Doctoral Dissertation

Development of Globularity and Torsion Based Analysis Methods For Protein Structures

구형성과 뒤틀림각에 기반한 단백질 구조 분석 방법론 개발

February 2013

Sunghoon Jung

Lab. of Computational Biology and Bioinformatics
Interdisciplinary Graduate Program in Bioinformatics
College of Natural Sciences
The Graduate School
Seoul National University

Development of Globularity and Torsion Based Analysis Methods for Protein Structures

February 2013

Sunghoon Jung

Lab. of Computational Biology and Bioinformatics
Interdisciplinary Graduate Program in Bioinformatics
College of Natural Sciences
The Graduate School
Seoul National University

Thesis submitted in fulfillment of the requirements for
the degree of Doctor of Philosophy to
Seoul National University

ABSTRACT

Development of Globularity and Torsion Based Analysis Methods for Protein Structures

Sunghoon Jung

Lab. of Computational Biology and Bioinformatics
Interdisciplinary Graduate Program in Bioinformatics
College of Natural Sciences
The Graduate School
Seoul National University

The structure of protein has intimate relationship with the function of protein. The structure of protein is experimentally determined through X-ray crystallography and NMR methods. However, X-ray crystallography is hard to obtain mobile protein structure and crystallization often causes practical problems. NMR structure is impossible in the observation of membranous or large proteins. Thus, theoretical methods for the determination of protein structures are highly concerned to circumvent practical problems. Homology, threading and *ab initio* modeling are the three typical approaches in protein structure modeling. *ab initio* modeling is often called as protein folding problem. The natural stable state of protein structure is believed to be the minimal energy state. The critical problem of protein folding research is the impossibility of the exhaustive search of possible conformations. Globularity of the protein structure was assessed in the pursuit of the universal structural constraint while approximated measurement name Gb-index was developed. Strong perfect globe-like character and the relationship between small size and the loss of globular structure was found among 7131 proteins which implies that living organisms have mechanisms to aid folding into the globular structure to reduce irreversible aggregation. This also implies the possible mechanisms of diseases caused by protein aggregation, including some

forms of trinucleotide repeat expansion-mediated diseases. Torsion angle constraint mimics natural process of conformational change of proteins which lacks significant movement along covalent bonds and change in bond angles. This torsion angle system was applied to structure alignment to prove the validity as a structural representation. It was more effective to accurately anticipate homology among 1891 pairs of proteins of 62 different proteases and among 1770 pairs of 60 proteins of kinases and proteases with the string of φ and ψ dihedral angle array than famous 3D structural alignment tool TM-align. Secondary structure database and structure alignment web server was constructed from PDB and SCOP entries based on the simple classification scheme according to the backbone torsion angles. The database introduced here offers functions of secondary database searching, secondary structure calculation, and pair-wise protein structure comparison. Visualization during the process of the protein folding simulation is quite interesting regarding the fast apprehension of the states while previous algorithms such as molecular dynamics offers very few options of interference. Computational application named ProtTorter which visualizes three-dimensional conformation, calculates the potential energy, and supplies the user interface for backbone torsion angle manipulation was developed. Using this application, simple folding algorithm was newly investigated. Cotranslational and torsional folding path was utilized in the context of Levinthal paradox. The validity of the folding method was investigated using the test sets of small peptides. Positive result for the possibility of this method was obtained as the stable negative energy minimal structures and fast convergence. Application of torsional system of which validity was proved in the structure alignment assays and globular constraints which might infer solvent interactions by minimizing solvent accessible surface area might be worth for further studies based on the folding algorithm using ProtTorter application.

Keywords : protein structure, protein folding, structural globularity, torsion angle system, Levinthal paradox, cotranslational folding.

Student ID : 2008-22789

DECLARATION

This thesis has been composed by myself and it has not been submitted in any previous application for a degree. The work reported within was conducted by me unless otherwise stated.

February 2013

Sunghoon Jung

For my parents and professor for their right and spiritual guidance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I.

Introduction

1.1 Background of Protein Research

Proteins are polymers of amino acid monomers. There are 20 types of amino acids which consist of proteins. These amino acids cover wide range of physicochemical properties including electric charge and conformation. All the amino acids exist in biological organisms are L type enantiomer. Side chain of amino acid determines specific characteristics of each type of amino acids. Some amino acids are polar with electric or partial charges, while others are non-polar or hydrophobic.

The synthesis of protein is one of the most crucial parts of metabolism of living organisms. The information of protein sequence flows from DNA in the nucleus to the cytosol or to the endoplasmic reticulum through messenger RNA. The DNA in the nucleus forms a structure called chromosome in combination with histone proteins. The transport of material into and out of the nucleus is tightly controlled through nuclear pore complex and proteins called importin and exportin. The messenger RNA is generated through the action of RNA polymerases. The DNA and RNA is composed of four types of nucleic acids; adenine, guanine, cytosine, and thymine or uracil. Combination of the three nucleic acids codes 20 amino acids from 64 possible ones by incorporating redundancy. These three nucleic acids are called as “codon.” The match of coding amino acid and the codon is called the genetic code. This genetic code is almost universal with small modifications in certain rare organisms indicating that all the present living organisms are from the common ancestor. There are region of coding and non-coding of protein amino acid in the sequence of nucleic acid of nuclear DNA. These are closely bounded by the start and stop codon. During the flow of information of

nuclear DNA to the outside of the nucleus, the region relevant to the protein coding sequence is copied by polymerase and synthesized into messenger RNA. This process is called as “transcription.” There are many steps of modification during the generation of mature messenger RNA from the nascent string of ribonucleic acid. There are also exceptions that messenger RNA does not indicate the information necessary to code a protein including the case of micro RNAs. However, the function of most RNA is to code the sequence of amino acid of the specific protein. Messenger RNA is sent out of the nucleus and transported either to cytosolic ribosome or to the ribosome attached to rough endoplasmic ribosome (rER). Ribosome employs transfer RNA to deliver necessary amino acid to the ribosome. Transfer RNA has anticodon region which signifies the matched codon sequence. Following the appropriate match between codon of mRNA and anticodon of tRNA, the ribosome correctly relates the information from the nucleus to the newly synthesized protein. tRNA also has amino acids attached by aminoacyl transferase. These attached amino acids are transferred to the carboxyl terminus of previously synthesized amino acid string. This process of transferring of nucleic acid information into protein amino acid sequence is called as “translation.”

The uniform flow of information from the information reservoir of nucleus into the final product of proteins with two steps of transcription and translation was proposed by Francis Crick and is known as the “central dogma.” Though there are exceptional cases including the cases of the information flow of retroviruses, majority of the living organisms strongly follows the central dogma.

1.1.1 The Function and Structure of Protein

The proteins synthesized following the central dogma is the most important component of living organism. The word protein itself indicates the primary importance of this material from the Greek etymology of *protos* which means “the first”. The major function of living organism is achieved by the functioning of proteins. Proteins have two major functions of structural element and chemical enzyme. Myriad number of phenomena of living organism is achieved by the interplay of these proteins with

utilization of other chemical resource.

The function of protein is strongly dependent to the three-dimensional conformation that the molecule can adopt. (Sheraga, 1957) Protein is believed to adopt a single native conformation unlike to many other polymers though it is also possible to contend that protein shows dynamic behavior of adoption of conformations. The native states are found under the similar circumstance of a living organism; i.e. aqueous solvent near neutral pH at 20-40ºC. When the environment where protein resides digresses from the native environment, the protein denatures (or unfolds) into strings without consolidations. The native structure of protein is often recovered or the protein renatures when the natural environment is tenderly recovers. The structure of a single protein molecule might be further divided into distinct region of separate function. These regions can usually fold independently into stable structure. Each of these regions are called “domain.”

1.1.2 Protein Secondary Structure

Three dimensional conformation of protein is very complicated and irregular. There are, however, certain typical motifs of local structure. These local structures are usually consolidated by hydrogen bond interactions. These structures are called as secondary structure being the structure of the next level of complicacy to the nascent string information of primary structure. Three-dimensional structure is similarly ascribed as tertiary structure and the combination of the tertiary structure is ascribed as quaternary structure. The simplest classification of the secondary structure categorizes into three types of helix, extended, and other structure.

Straight string of amino acid might wind up being supported by hydrogen bonds between backbone amide and carboxyl atoms of residues which are several positions apart. The most famous example of this helical structure is α-helix. There also exists another possibility of stabilized secondary structure. In this case, the amino acid residues do not wind up and extends like a string in a zigzag pattern. This “extended” structure is

stabilized by hydrogen bonds between the extended strings. Typical extended structure is β-strand which consist β-sheet structure. In β-sheet, the β-strand might run in parallel or anti-parallel direction. Any other structure except helix and extended structure is called as other structure. These irregular loop structures might connect helix and extended structures such as the β-turn motif between β-strands. (Wilmot et al., 1998)

1.1.3 Torsion Angle

Covalent bonds are intact during the conformational change of protein molecule. Thus, bond length and bond angle are not the subjects of consideration during the protein folding within the living organism. The change of the three-dimensional conformation is only due to the rotational variation of torsion angles of covalent bonds. Among many rotatable single bonds, torsion angles of backbone have the most profound impact on the conformation of the protein molecule. It is worth to note that rotation of a single torsion angle of the backbone causes perfect different location of the part resides in the rear to the position of change. Among the three backbone torsion angles of a residue, N-Cα bond torsion (φ) and Cα-C bond torsion (ψ) is freely rotatable. The peptide bond (amide bond) is not rotatable due to the planarity character originated from partial double bond. Most of the torsion angle of peptide bond is 180º establishing *trans* conformation while there are more frequent exception of *cis* conformation of 0º for proline.

Ramachandran plot is the plot of φ and ψ angle of each amino acid on a two-dimensional plane. There are frequently occupied region by amino acids on the ramachandran plot. (Ramachandran et al., 1963) The side chains also have a preferred conformation though there are many unusual or high energy structures. (Ponder and Richards, 1987) Further investigation revealed that the side-chain conformation is rather correlated with the structure with backbone; i.e. some backbone allows only particular type of side chain conformation (Summers et al., 1987; Dunbrack and Karplus, 1993)

1.1.4 Hydrophobic Effect

Globular water-soluble proteins typically have the conformation in which hydrophobic residues reside near the core area and hydrophilic residues reside on the surface. The factor that causes the packing of the hydrophobic core is "hydrophobic effect." This is a strong factor that sustains the stability of the protein conformation. This factor is believed to be originated from entropy though it is still contentious with some debates. This factor is explained as the limitation of possible microstates of water molecule around the non-polar surface of protein molecule. If the limitative surface area of the protein is larger, the entropy decreases and vice versa. However, it is worth noting that preliminary molecular dynamics simulations were unable to find any evidence for the enhancement of water structure at a hydrophobic protein interface (Kovacs et al., 1997; Leach, 2001)

Membrane proteins are also important functioning molecule in the living organism. This protein includes receptors and ion channels. The topology of this type of protein in the membrane spanning region is reversed from soluble protein; i.e. hydrophobic amino acids reside in the outside toward the membrane. Though the experimental determination of the three-dimensional structure of membrane protein is very difficult, many structures were revealed. Seven trans-membrane helices motif is a well-known structure among many and could be observed in the structure of bacteriorhodopsin and rhodopsin (Henderson et al., 1990; Habelka et al., 1995)

1.2 Experimental Structure Determination Methods

1.2.1 X-ray Crystallography

There are experimental methods that reveal the structure of proteins. The most popular and old method is X-ray crystallography. This method is the oldest method among the ones which determines the protein conformation. This method requires the crystal of proteins (1-2Å) prepared from slowly dried aqueous solution. The crystal has a lattice structure of consecutive protein molecules. When the strong beam of X-ray is traced onto the crystal, the beam of light scatters interacting with the protein molecule. By the summation of interference between waves of light, the particular pattern of diffracted points (~1000 spots) appears. From this information of constructive and destructive interference of light waves and the mathematical tool of Fourier function, the three-dimensional electron density map is modeled. Observing the electron density map, researchers detail the atomic information.

1.2.2 NMR Spectroscopy

NMR spectroscopy was developed in 1980s. This method employs nucleic resonance information which arises from near atom pairs instead of light beam interference patterns. From this resonance information, the distance and near pairs of atoms are calculated. The constraints of structural character are built from this information and the series of models (15-20) of structures are constructed from these constraints. Usually isotopes with radiation are used as the label for the lack of marked resonance property of usual atoms. Though the preparation of radio-isotope labeled protein is costly and cumbersome, this method supplies great advantage by enabling the possibility of observation of native, i.e. near physiological condition, structure in aqueous environment. Also this method enables the observation of dynamics of conformational change while X-ray crystallography only enables to observe a single conformation.

1.2.3 Limitations of Experimental Methods

There are, however, limitations in these methods. Though X-ray crystallography allows the determination of very large proteins and membrane proteins, this needs a delicate process of crystallization of proteins which are time-consuming and very hard to succeed. Obtaining a single structure might cost several year for a single researcher. X-ray crystallography is also unable to discern highly mobile proteins. NMR method is hard to obtain the structure of large proteins with the limit of 30kD. Also this method is unable to determine the structure of membrane proteins.

1.3 Protein Structure Prediction Methods

Though there are possible methods for the determining of the structure of proteins, the speed of accumulation of revealed protein sequence far exceeds the rate of protein structure determination. Due to the many genome projects of organisms, there is bountiful new information of gene sequence which codes for protein. However, the functional study of these genes is not always efficient for many possible limitations. Many laboratory molecular biology experiments need pain-taking illogical trials and observations. Elucidation of the function of a single unknown gene might take several years when the phenotype is masked by several epistatic elements. Furthermore, the demarcation of the object gene with unknown function is still difficult in spite of the aid from genomic sequencing. Exons from a single coding region might be recombined to produce proteins with different functions.

There are numerous species of organisms with different set of genome and proteome. The number of species might be tremendously increased if one considers unrevealed microorganisms which might have profound scientific and industrial implications. To elucidate the whole knowledge of the biology of these organisms by laboratory molecular biology alone with traditional genetics is strongly impossible to finish within a plausible time span. Thus, considering that three-dimensional structure is strongly informative to the function of the protein, the theoretical and computational method to reveal the three-dimensional structure from protein amino acid sequence alone is invaluable. This problem of anticipation of the structure from amino acid monomer sequence is known as the “protein folding” problem. This is one of the crucial sub-disciplines of “bioinformatics” which concerns with the collection, organization, and analysis of biological data. (Leach, 2001) The protein folding problem has 40 years of history and hundreds of papers of this field are published in each year. The general approach to solve the protein folding problem can be categorized into three subtypes of homology modeling, threading, and *ab initio* modeling.

1.3.1 Homology or Comparative Modeling Method

Homology modeling or comparative modeling method simply compares the sequence of protein with unknown structure with the sequence of proteins with predetermined structures. This is the most powerful and accurate method among the three, partly indicating that currently revealed protein structures covers wide range of possible structures. Though this method is strong for most cases, there are exceptions of failures when the appropriate known structure of similar sequence does not exist. Due to the main principle of comparison of sequences for the determination of structure of query protein, the quality of the prediction strongly depends on the degree of similarity between query protein sequence and sequence of known structure. Deviation of 0.3Å arises for each 10% reduction of sequence identity and protein of sequence similarity less than 30% to proteins in the structure library is considered impossible for prediction. (Baxevanis and Ouellette, 2005)

Homology modeling process has five major steps. First, the sequence alignment of the sequence of the query protein to the sequences of the library proteins is conducted. Referring the alignment with insertions and deletions, certain region of backbone of the reference structure is selected and replaced referring common backbone segment library. Side chain is also changed according to the difference in the alignment and the difference which arisen from previous step of backbone modifications. This built model is refined by energy minimization techniques which relieve collisions and steric strains. Visual and numerical validation is finally conducted by computational viewer or validation applications. Many of the steps among the five steps are supplementary modifications except the first sequence alignment step. To improve the quality of the modeling based on homology, most of the approach employs multiple structural homolog databases rather than single one.

Most of the homology modeling process needs extensive interference of human manipulation. Automated homology modeling applications have developed including Modeller (Sali, 1998), DeepView (http://us.expasy.org/spdbv), WHATIF (Krieger et al.,

2003). Homology modeling web server includes SWISS-MODEL server (Schwede et al., 2003; http://www.expasy.org/swissmod/SWISS_MODEL.html), the CPH Models server (http://www.cbs.dtu.dk/services/CPHmodels/), and the SDSC1 server (http://cl.sdsc.edu/hm.html). EVA (Evaluation of Structure Prediction Server; http://cubic.bioc.columbia.edu/eva/) ranked SWISS-MODDEL and SDSC1 the best with outperforming accuracy and speed among the participant servers.

1.3.2 Threading Method

Threading is another widely used method for the modeling of protein structure from protein sequence. This method is useful when there is no possible homolog of the query protein in the structure homolog database. This method shows limited accuracy when compared to the homology modeling for it is conducted from reference structure that lacks the sequence similarity. For the failure of finding the right reference structure by sequence alignment, threading method uses string of amino acid of the query protein as probe which flows through the pipe of backbone structure of temporary reference protein. This string is also called as a "snake." After the snake flowing through the pipe, the fitness of the probe with the reference backbone structure is quantified. Usually empirical energy function or some type of packing efficiency measurement is used for the fitness quantification.

According to several threading results, the similar amino acid sequence of protein does not necessarily imply the similar three-dimensional structure. Completely different sequence might fit well to the reference structure by threading. There are two types of threading method with different fitness evaluation method. First one is called as "three-dimensional threading." This is classified into distance-based method (DBM) because of the importance of three-dimensional distance. Three-dimensional threading was first developed by Novotony and colleagues (Novotony et al., 1984). It was reevaluated after the heuristic potential function became solid background (Jones et al., 1992; Sippl and Weitckus, 1992; Bryant and Lawrence, 1993). This method is typical one following the basic process described above. After the flowing of the probe through the backbone pipe,

the fitness is evaluated by distance-based (or profile-based) energy functions (Bowie et al., 1991). The energy of the conformation of protein can be assessed by this heuristic energy function that relies only on the distance and atom types of each residue. Thus, this method is possibly regarded as to be dependent only to the Cartesian coordinate information.

The other method of threading is called as "two-dimensional threading." This is classified into "prediction-based method (PBM)" and was first developed in mid 1980s (Sheridan et al., 1985). This method was widely used after the improvement of the accuracy of secondary structure prediction. (Fischer and Eisenberg, 1996; Rost et al., 1997). This method is primarily dependent on the predicted secondary structure of the query sequence. Solvent accessibility and sequence of amino acid residue information is optional during the method implementation. Secondary structure sequence rather than amino acid sequence itself is considered valuable for the more conserved information of the secondary structure. All information considered for the threading is integrated and converted into one dimensional string named "pseudo sequence." The threading process is able to be simply described as the alignment of two pseudo sequences. All sequence alignment algorithm could be applied during this process; e.g. dynamic programming (Needleman and Wunsch, 1970; Smith and Waterman, 1981) and BLAST variants (Altschul et al., 1990; Altschul et al., 1997). This method is called as two-dimensional threading for the dimensionality of the alignment space which has two axes of pseudo sequences. This two-dimensional threading is 10-100 times faster than three-dimensional threading for its reduced number of dimension. The two-dimensional method, however, shows comparable or better accuracy than three-dimensional method. (Baxevanis and Outllette, 2005)

Though new possibility originated from threading method promises more capability of protein structure modeling, the quality of the result of this method is quite limited with RMSD higher than 3Å. (Baxevanis and Outllette, 2005) The importance of this method, however, could be found in the fact that this method reveals approximate structure

model for proteins which have no homolog in the pre-built structure database. One of the successful examples of the academic achievement using threading method is the case of the protein named leptin which is the important obesity relevant factor and had no previously known structural homolog. (Madej et al., 1995) From this modeling, rather accurate activity mechanism of the protein was derived.

There are numerous web threading servers. Most of them are based on two-dimensional threading. EVA (Evaluation of Structure Prediction Server) lists and ranks threading web servers in addition to homology modeling servers. This list includes BLAST and PSI-BLAST servers though these are actually good sequence alignment server and are strictly not the threading servers. According to the result as of 2005, SAMt99 (http://www.cse.ucsc.edu/research/compbio/HMM-apps/T99-model-library-search.html), three-dimensional PSSM (http://www.bmm.icnet.uk/~3dpssm/), and FUGUE (Shi et al., 2001) were the best performers among the ranked servers. Metaservers which combine the results from multiple servers recently appeared and often resulted in better performance to original ones.

1.3.3 *ab initio* Method

The third method of protein structure anticipation other than homology modeling and threading is *ab initio* method. The word *ab initio* is latin language which literally means “from the start”; *ab* is a preposition which means “from” and *initio* is the ablative form (a form comes with prepositions) of *initium* which means the beginning. This method predicts the structure of a query protein from “nothing” as meant by the name *ab initio* (from the beginning). In other words, this method does not exploit previously determined protein structure of homolog. This method has not yet complete solution to result in very accurate structure model and still experimental. Usually, the models that are generated through widely known *ab initio* method are unreliable for practical or even academic uses. The problem of deriving perfectly reliable algorithm for *ab initio* modeling is regarded as the true “protein folding problem.” Though the accurate modeling for small proteins is considered possible, the large proteins are still

unable to be correctly predicted. In this sense, the protein folding problem is not solved.

1.3.3.1 Molecular Dynamics Simulation Method

Molecular Dynamics method could be easily interpreted as the method which simulates the natural dynamics of atoms in the protein molecule. The scheme of the simulation could be applied to any type of molecule in addition to proteins. The dynamics of classical Newton's mechanics is applied to the simulation. According to Newton's second law, the acceleration of an object depends on the ratio of force to mass exerted on the object as

$$\frac{d^2x_i}{dt^2} = \frac{F_{x_i}}{m_i}$$

, where x_i is the position of object i, t is the time, F_{x_i} is force exerted on i at position x_i, and m_i is the mass of the object i. The typical force field can be represented as follows (Höltje et al., 2008),

$$E_{tot} = E_{stretch} + E_{bend} + E_{torsion} + E_{vdW} + E_{electric}$$

, where $E_{stretch}$ is the potential energy term of covalent bond stretching, E_{bend} is the term of covalent bond angle bending, $E_{torsion}$ is the term of single bond torsion, E_{vdW} is the term of van der Waals interaction, and $E_{electric}$ is the term of Coulomb force interaction. However, the way of determining the force depends on the choice of simulation setter. Empirical force field is also possible to be used.

Using the force from employed force field, the acceleration could be calculated for any static moment of simulation. Using the information of acceleration and mathematical integration, the position of the next moment could be theoretically derived. Though this rationale is straightforward theoretically, there is a computational limitation of continuous calculation of derivatives and integrations. Thus, approximation is applied during the calculation of the positions along the time span. This usually uses Taylor series expansions.

Verlet algorithm (Verlet, 1967) is one of the most widely used integration method. This method can calculate the position of the next time step without calculating the velocity of the particle but only with the acceleration and the position of the particle in the previous step. The position could be derived as follows,

$$\mathbf{r}(t + \delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \delta t) + \delta t^2\mathbf{a}(t)$$

, which was obtained from the addition of the two equations below.

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t\mathbf{v}(t) + \frac{1}{2}\delta t^2\mathbf{a}(t) + \cdots$$

$$\mathbf{r}(t - \delta t) = \mathbf{r}(t) - \delta t\mathbf{v}(t) + \frac{1}{2}\delta t^2\mathbf{a}(t) + \cdots$$

The velocity of the particle at time t could also be derived as follows.

$$\mathbf{v}(t) = [\mathbf{r}(t + \delta t) - \mathbf{r}(t - \delta t)]/2\delta t$$

One of the problems of the simulation of molecular dynamics is that the conformational search space of the protein of typical size is too large to be thoroughly investigated. One might anticipate to reduce the size of the search space by appropriate constrains and restraints could be helpful. By this application, the simulation would more concentrate on the more realistic conformations. The distinction of the word of restraint and constraint might be worth to be noted. Restraint is able to be interpreted as the penalty which usually included into the force field. Thus, the conformation might violate the restraint condition. Constraint, however, is a condition that must be abided during the simulation. Thus, the conformation may not violate the conditions given by constraints. One example of the constraint dynamics might be the torsion angle molecular dynamics algorithm DYANA (Güntert et al., 1997). In this algorithm the degree of freedom is strongly reduced by only allowing the movement along the torsion angles. There is also a constraint dynamics using globularity criteria (Palù et al., 2004).

1.3.3.2 Levinthal Paradox

Though there are many constraints for the molecular dynamics and they are helpful for accurate folding of protein, the search space of the simulation is still too large for the whole scale exhaustive search. Simple estimation of the magnitude of the conformational space might be possible based on the scheme of torsion angle simulation. Typical protein with ~100 amino acid residues might have ~300 backbone torsion angles and several folds more number of side chain torsion angles. If the simulation grids each torsion angle into 360 steps, the total number of possible conformation is about $\sim 10^{700}$. Even if one assumes that each conformation is assessed within a millisecond (10^{-6}s), the whole search of the space takes $\sim 10^{687}$ year. This is perfectly not feasible to conclude the result of the exhaustive search. One might conclude that even in nature, the folding of a protein is not possible from the enumeration of all possible conformation. However, the protein amino acid string can fold into unique structure in just a few seconds or less. Thus, the supposition that protein sequence might fold following the preferred or predestined path rather than relying on the simple trial and error for all possibility. This supposition is called as Levinthal paradox (Levinthal, 1969).

1.3.3.3 Lattice Model

Lattice model rather than three-dimensional atomistic model is frequently employed. Though this model cannot give the information of specific interaction at the atomistic level, this can be used to investigate the fundamental questions about protein structure. Also, the exhaustive search of possible conformations is often possible with this method. From these extensive search results, the thermodynamic properties could be derived using the concept of statistical dynamics.

One of the simple lattice models is “hydrophilic-hydrophobic (HP)” model (Chan and Dill, 1993). In this model the amino acid residue of protein is considered either as hydrophobic (H) or hydrophilic (P) monomers. The typical force field of HP model is as follows.

$$E_{HH} = -1$$

$$E_{HP} = 0$$

$$E_{PP} = 0$$

, where E_{HH} is the potential energy of the hydrophilic atom pair, E_{HP} is the potential energy of the pair of hydrophobic atom and hydrophilic atom, and E_{PP} is the potential energy of the pair of hydrophobic atoms. The string of hydrophilic and hydrophobic monomer is sequentially grown on the two-dimensional lattice following the self-avoiding walks. The energy of the conformation is calculated from pairs of adjacent monomers referring the force field except the covalent bond pairs. Several features were obtained from this model. If hydrophobic-hydrophobic interaction energy term is weak, a large number of conformations are stable. If this interaction energy becomes strong, a sharp decrease in the number of stable conformation with hydrophobic core is observed. α-helix and β-sheet also arises naturally within this model. This finding of the formation of secondary structure might possibly suggest that hydrogen bonding pattern is not the only factor of the secondary structure formation. Simple packing might be one of the reasons for the local topological structure.

More sophisticated lattice models which are intended to generate fully detailed structure exist. Usually those fine-grained lattice model needs adjuvant algorithms including simulated annealing to search low energy conformation from the vast number of possible conformations. Simulated annealing helps to find more stable conformations by broader mutations originated from the up-lifted temperature. Repetitive process of rising and falling of temperature leads to the overcome of the limitations from the trapping of potential well during the search of minimal conformations. Skolnick developed lattice models which are used in a three-stage procedure (Godzik et al., 1993; Skolnick et al., 1997). In the first stage of the procedure, a coarse lattice model is used to generate a set of preliminary candidate conformations. Monte Carlo simulated annealing is used for

the generation of conformations. Monte Carlo method is purely randomized method which is different from molecular dynamics. The low-energy structures from the first stage are refined in the finer second lattice model which uses more accurate representation of side chains. This second model is believed to be more similar to the actual structure of the protein. In the final stage the structure from the second stage is converted into a full atomic model using typical simulations with standard force fields.

Lattice model could be ascribed to be one of the constraint simulations for the restricted possible movements. Though the attempts of reducing the degree of freedom has not always been successful many genuine approximations have been suggested. One of them is Gō model (Ueda et al., 1975). Gō model uses a single pseudo-atom to represent the amino acid. The position of the pseudo-atom is regarded to be on the Cα atom. The concatenated Cα string is treated to be connected through harmonic potential with experimentally determined minima at 3.8Å. Purely repulsive or Lennard-Jones type interaction is usually applied. The performance of these approximated methods is questionable if the intention of the simulation is to reproduce the exact protein three-dimensional structure. However, if one considers the impossibility of exhaustive search of the candidate space and Levinthal paradox, the constraint or approximated approach could be regarded as the possible leader to the feasible solution for the protein folding problem.

1.3.3.4 Monte Carlo Method

Monte Carlo simulation method generates conformations of proteins by utilizing random changes of positions of atoms of the molecule. This method is an easy method to implement for its simple manipulation of atom positions; i.e. only simple random movements are necessary. Though the method is simple to apply, it is still difficult to simulate flexible molecule like protein unless the size of the simulation system is small either by the small size of the molecule or by specific constraint that reduces the degree of freedom. The utility of reducing the degree of freedom is clear especially in the case of Monte Carlo method. If one allows the movement of atoms

along any direction in the Cartesian coordinate system there might be very frequent deviations of typical bond length among bonded atom pairs. The movement away from the equilibrium bond length by any random change of atomic position might induce strong increase in the potential energy of the molecule. In the scope of finding the minimal energy conformation among the whole possible cases, these high energy conformations resulted from the bond length deviation are strictly useless. Thus, appropriate freezing of some degrees of freedom is necessary.

Recently, there was a trial to simulate the folding of protein using Monte Carlo method with degrees of freedom of torsion angles. (DeBartolo et al., 2009) This method uses library of move set which signifies amino acid sequence dependent backbone torsional angle preference. It utilized statistical potential derived from backbone hydrogen bond requirements, chemical property, and packing preference of 20 amino acids. It was simulated in the coarse grained model with amino acids without side chains which only allowed C_β atom representation. Monte Carlo simulated annealing was applied with increasingly restrictive constraints. The simulation process iterates the process of conformational change and restriction of freedom of movement referring the state of determined secondary structure. Every round of iteration conducts the folding simulation from stretched string. The only difference of subsequent rounds from previous ones is the restricted move sets of φ and ψ angle from torsional angle preference library. Once the secondary structure is determined, the torsional propensity of φ and ψ angle is inclined to the formation of the determined secondary structure. The authors contended that this approach mimics the real process of protein folding. This example of restrictive Monte Carlo method demonstrates the importance of appropriate limitations of the internal freedom of movement. With robust restrictions and constraints, proper simulation of Monte Carlo would be performed in conjunction with adjuvant algorithms for accurate conformational search. Possible helping algorithm could include energy minimization methods which try to follow the gradient of the potential local minima well.

1.3.4 Competition of Protein Structure Prediction Methods: CASP

There is an academic competition named CASP (Critical Assessment of Structure Prediction) which is organized in 1994-1995 and held every two years for protein folding researchers. The coordinates of newly revealed protein structure is donated by crystallographer and NMR spectroscopists to the organizers of CASP. Using these unrevealed structures as reference, CASP holds blind tests for the protein model calculating methods of participants. Registered predictors submit their predicted structural models for the test protein sequence within 6 or 8 weeks from the question announcement. After the close of the session, the submitted structures are evaluated using structure comparison tools. The robust performers publish and explain their methods.

In the first CASP challenge, competitors predicted three-dimensional structure from seven protein amino acid sequences. After the choice of the protein sequence to be predicted, participants had to construct sequence alignment. Final prediction of the protein structure was made from this alignment information. RMSD of just 0.6Å was achieved for the best anticipation. The accuracy of the built model was strongly dependent on the degree of the sequence similarity; i.e. the sequence identity and the quantity of insertions and deletions. According to the results of the CASP assessments, there have been profound improvements among the protein structure prediction methods. ROSETTA (Bonneau et al., 2001) is one of the famous applications among numerous protein folding methods which showed high rank of performance in the CASP competitions.

1.4 Studies and Concerns of the Protein Folding Research

The natural stable state of protein structure is believed to be the minimal energy state among every possible state. The most critical problem of protein folding research, however, is the impossibility of the exhaustive search of possible conformations. Within the limited scale of search space, simulation would bind up to result in local minimal energy conformation. This is specifically true for the results of energy minimization methods which would wander around minimal energy structures which is still different from global energy minima state. The simulation of energy minimization method would be trapped within a local minimum by the hindrance from the height of the minimal energy potential well. Many methods to escape the trapping of the local potential energy minimum, including replica-exchange method and simulated annealing, were developed. Replica exchange method exchanges multiple copies of sample conformation derived from molecular dynamics simulation performed in different temperatures to help to cover more range of conformational space. However, there has not been a perfectly suitable solution for the problem of feasible finding of the native state.

If the simulation which covers all possibility is practically unreliable, the importance of Levinthal paradox becomes evident. If one can reveal the native path of folding of protein amino acid string, regeneration of the accurate conformation by the simulation becomes tenable. Thus, it might be an interesting trial to postulate the path along the conformational space that reflects the way of protein folding and apply this folding path to *ab initio* modeling. Among many of the possible paths, the cotranslational protein folding has been suggested and recently reviewed. (Fedorov and Baldwin, 1997; Basharov, 2000; Basharov, 2003; Kolb, 2001; Giglione et al., 2009; Kadokura and Beckwith, 2009; Ellis et al., 2010; Saunders et al., 2011; Srivastava et al., 2011) The postulation that the native structure of protein is strongly dependent on the folding of the amino acid sequence during the translation process is quite plausible considering the short time needed to fold the protein compared to the long time of amino acid residue

addition and peptide extension. If the postulation of the cotranslational folding is applicable, the degree of freedom of the simulation system could be drastically reduced for it is necessary to consider only the movement of newly added amino acid. Also, there is a possibility of further modification after the synthesis of the protein. For proper modeling of proteins, additional path following Levinthal paradox would be necessary. In chapter 6, this author tried to validate the robustness of the concept of cotranslational protein folding and further Levinthal path optimization in the modeling of peptide structure. The computational application with graphical user interface which enables such validation was referred in chapter 5.

There are many possible methods to implement the initial structure generation using cotranslational folding. Molecular dynamics and Monte Carlo methods are ones of the examples. These automated simulation methods, however, are somewhat cumbersome for interactive human interventions of detecting information and manipulation. The human perception of the intermediate results during the simulation process might provide valuable genuine interpretations and findings. This scheme of modeling could be considered to be originated from traditional physical modeling studies which were conducted when computational approaches were not tenable. The application which was introduced in chapter 5, utilized very interactive user interface for the modeling. Using this interactive application, the research of the trial of a possible Levinthal path of the folding of small peptides was conducted in chapter 6.

The construction of conformational system is also necessary. Simple three-dimensional Cartesian system and other constrained systems including torsion angle system are possible to be used. This author focused on the use of torsion angle system for it is more intimate to the scheme of the conformational change in the natural environment; i.e. torsion angles are the only members which are practically free to be changed. Before the application of torsion angle model to the research of chapter 5 and chapter 6, the effectiveness of the system was scrutinized by applying it to the structure alignment studies in chapter 3. In this chapter, the validity of structure alignment using torsion

angle string was investigated. Also, secondary structure and protein structure pair-wise comparison web server application was constructed and referred in chapter 4.

Cotranslational protein folding is quite interesting and might be the promising solution of Levinthal path for the protein folding problem. However, though this method strongly decreases the degree of freedom, more robust and straightforward additional constraint that fundamentally frames the broad shape of the protein is quite necessary. Finding the universal structural characteristics of native structure of proteins, thus, might provide very helpful insights to the formulation of the proper simulation constraints. In chapter 2, the globular structure was found to be almost universal structural characteristic and the degree of geometrical globularity of diverse types of proteins was quantified.

Even if that folding during translation strongly influences the conformation of the protein as an initial structure, the modification of the conformation after the release from the ribosome is possible to significantly modify the preliminary structure. In chapter 6, the additional path to the initial cotranslational folding followed the order of the possible magnitude of the torsional disturbances relayed to the backbone torsional bonds from the side chains of amino acids. Further study of the appropriate simulation method after the cotranslational folding might be an important subject of future research. Whatever that method would be, the concept of Levinthal paradox might be still valuable regarding the scale of the space of possible conformations.

We tried to investigate possible constraint, validity of torsion angle model, and the validity of protein folding following the concept of Levinthal paradox. Based on the findings of the chapters from 2 to 6, genuine approach for the reliable folding simulation could be possibly made. The globular character of protein might be utilized as the folding criteria and also for the representation of the solvent effect while torsional system could strongly reduce the folding space. Concept of Levinthal paradox and torsional paths of folding might help for better apprehensions of the nature of the protein folding and structural analysis. This author hopes these studies to be of help for the

development of the correct solution for the protein folding problem which has been the primary concern for structural bioinformaticians for several decades.

CHAPTER II.

Analysis of the Globular Nature of Protein

2.1 Introduction

Protein structure is considered to be the most important primary information in molecular biology, especially in pharmaceutical studies. The difficulty, however, of deriving structural information from proteins using protein crystals or protein solutions leads to the development of protein structure prediction methods based on amino acid sequences and other already revealed structures. The most critical problem of the protein structure anticipation method is the colossal magnitude of possible conformational states. Numerous restraints and simplifications have been developed to be appropriately applied to reduce the search space while minimizing false structures. Thus, if one could obtain the universal structural characteristics of protein structures, very helpful constraints and restraints could be derived for the simulation of protein folding.

Recently, Palù et al. (2004) incorporated globularity as their protein folding simulation criterion using Constraint Logic Programming. Globularity, expressed by the radius of gyration, was used to improve the packing and accuracy of NMR structures in previous research (Kuszewski et al., 1999), and the validity of the globular restraint for NMR protein structure determination was examined by Huang and Powers (2001). Globularity was also successfully used to assess the quality of models submitted to the Critical Assessment of Techniques for Protein Structure Prediction center (CASP; Constantini et al., 2007). Although protein globularity is assumed to be a valid criterion in many studies, to our knowledge, an analysis of the globularity of proteins investigating a whole database of protein structures had not been performed.

Previous studies used only about a score of proteins to validate their globular suppositions. Here, we investigated if globularity is a general character of most proteins and whether it can be applied as a valid constraint in protein structure simulations using virtually all the protein structures in the Research Collaboratory for Structural Bioinformatics' Protein Data Bank (RCSB's PDB) database. We removed redundant entries and divided the proteins into subcategories to enable a more detailed analysis.

Chaperones are known to protect the aggregation of misfolded proteins by binding and aiding the recycling of the folding process, especially in the endoplasmic reticulum (ER) during protein synthesis. The delineation of correct- and misfolded states by chaperones suggests a conundrum because many more proteins exist than chaperones and related molecules. Complete recognition of a correctly folded structure by a structural protein–protein binding site interaction is almost impossible because one protein might possess numerous structural characters. Accordingly, a possibility exists that the globular character might be the checkpoint of the correct folding in biological organisms. This assumption is supported by the aggregation of misfolded proteins because non-globular proteins tend to bind more tightly with one another due to the larger surface area provided by the loss of globularity.

The correct postulation for the underlying mechanism of chaperones in protein synthesis is quite important and has tremendous biological implications. It is quite probable that structural globularity would be the checkpoint of correct folding for the possible universality among other candidate structural characteristics. This structural globularity might have other biological functions in the biological organisms than the possible checking criteria of correct folding. One of such possibilities includes the inhibition of irreversible aggregations. Aggregational diseases including prion diseases, Altzheimer disease, and some forms of trinucleotide repeat expansion-mediated disease are strong concern of present biological studies. Plaques of aggregates of proteins are thought to be one of the major causes of the diseases. Globular proteins provide less binding area than linear proteins for the convex characters of the surface. Thus, it is probable that globular

proteins prohibit pathogenic aggregations. Proteins might have sequence information to make the structure into the globe to protect irreversible aggregations which might occur among nonglobular rod-like proteins. There is a strong native tendency of molecules to reduce the surface area within the water environment to form globular structure. On the other hand, there still exist rather linear protein molecules. Whichever of the sequence-driven and solvent-driven methods the underlying mechanisms of the formation of the globular structure might be, scrutinizing if the globular structure is the general tendency of the real proteins with previously revealed structures might be valuable and has been carried out in this study.

Unexpectedly, most of the proteins showed strong structural globularity (i.e., mode of approximately 76% similarity to the perfect globe) with only a small proportion of the proteins being outliers. This strong perfect globe-like character implies partial validity to the postulation that living organisms have mechanisms to aid folding into globular structures to reduce irreversible aggregation. It also implies the possible mechanisms of protein aggregation diseases including some forms of trinucleotide repeat expansion-mediated diseases.

2.2 Materials and Methods

2.2.1 Data Sets

PDB files were collected based on the Structural Classification of Proteins (SCOP; Murzin et al., 1990). We used all-alpha, all-beta, alpha/beta, alpha+beta, multidomain proteins, and other minor proteins including peptides, small proteins, and coiled-coil proteins. We excluded membrane proteins and peptides because of their topological difference and lipid membrane surrounding environment, which has different characteristics from that of soluble proteins. We also excluded fragmented and nucleic acid-containing structures, but included ligand-bound proteins. We removed structures that have 90% or more sequence identity to others to reduce redundancy.

Redundancy also arose from proteins with multiple chains belonging to different sources of organisms or different SCOP classes. PDB entries with redundant source organisms were removed but structures with two or more different SCOP classes were not excluded to allow the investigation of as many protein structures as possible.

In total, 7131 PDB structures were analyzed with 1365 all-alpha chain, 1503 all-beta chain, 2690 alpha/beta chain, 2067 alpha+beta chain, and 182 multidomain chain containing proteins and 547 other proteins. Programs to sort proteins according to their structural classification, source of organism, oligomeric states, and to filter out redundant and fragmented structures were all written in JAVA language.

2.2.2 Globularity Measurement

We defined new simple geometric quantities to represent globularity other than the radius of gyration because the radius of gyration might misinterpret internal cavities. Our globularity index (Gb-index) was defined as the ratio of the length of the longest displacement of any two atoms of the protein to the average of the longest lengths of two displacements that are orthogonal to each other and to the longest displacement. This approximated measure was chosen because cubic proteins are assumed to be extremely rare in real cases. The orthogonal criterion was surveyed within a 2˚ span from a perfect orthogonal angle. A range of 2˚ was successful for all the cases tested. We calculated the mean, standard deviation (S.D.), median, and the minimum and the maximum values of these indices of globularity. All the necessary programs were written in JAVA.

2.3 Results and Discussion

The distribution of the degree of globularity was analyzed according to the source of organisms (Figure 2-1a) and SCOP classification (Figure 2-1b). Except coiled-coil proteins and peptides, all kinds of proteins of SCOP classifications including all-alpha, all-beta, alpha/beta, alpha+beta, multidomain, and small proteins showed mean Gb-indices from 0.69 to 0.73 and modes from 0.76 to 0.84. Their median values ranged from 0.71 to 0.74. The mean and mode of all proteins was 0.71 (S.D. 0.14) and 0.76, respectively, with a median of 0.72. The mean Gb-index of peptides was 0.59 (S.D. 0.16) and the median was 0.58. Coiled-coil proteins showed the lowest Gb-index with a mean of 0.42 (S.D. 0.22) and a median of 0.40. Modes of the Gb-indices of peptides and coiled-coil proteins were 0.6 and 0.2, respectively. Details of the values of each type of protein are shown in Table 2-1.

No significant difference of globularity was observed between proteins from different organisms. Gb-indices showed similar average (0.70–0.71) and median (0.72-0.73) values among proteins from different organisms, except viral proteins, which showed a slightly lower average Gb-index of 0.67 (S.D. 0.18) and median of 0.70. The modes of the Gb-indices of the proteins from different organisms ranged from 0.72 to 0.76 with the mode of 0.76 for the whole protein, 0.72 for archaeal and eukaryotic proteins, and 0.76 for bacterial and viral proteins.

Viral coiled-coil proteins gave a minimum Gb-index of 0.08 (PDB ID: 1 pjf). The mean Gb-index of viral proteins was slightly lower (mean of 0.67) compared to proteins from other sources (mean of 0.71). Though this deviation may have been due to the small sample size (234 entries) compared to other proteins (total of 7131), the lower globular character might have originated from the abundance of structural capsid proteins.

Ninety-five percent of the proteins from any biological source had Gb-indices higher than 0.453, and 97% of the proteins had Gb-indices higher than 0.413. This result strongly indicated that almost every protein is globular, partly validating previous

attempts that used the globularity criterion in anticipating protein structures (Palù et al., 2004). However, non-globular, linear proteins were observed, as represented by low Gb-indices, which implies that the folding criterion based on globularity might not be suitable for all cases. Rather than the uniform mathematical formula for the radius of gyration according to the length of the polypeptide chain (Skolnick et al., 1997), sequence- and other character-based homology search for expected globularity might be more suitable for proteins with varying degrees of globularity.

A few percent of non-globular proteins existed although most of the proteins were globular. We investigated the possible relationship with the size of the protein and the tendency to lose the globular structure. We drew a graph of the mean, minimum, and maximum Gb-indices of proteins along with the number of atoms in the proteins (Figure 2-2). In all cases, the means were always approximately 0.7 and the maximum Gb-indices were always just below 1.0. However, the minimum globularity of proteins showed logarithmic growth with the square correlation coefficient (R^2) of 0.62 to the regression line, indicating that smaller-sized proteins were more likely to deviate from globular structures.

We also analyzed the relationship of the numbers of proteins with globularity lower than 0.453, i.e., the lowest 5% of non-globular proteins with the size of the proteins (Figure 2-2). The findings showed that the smaller the protein, the more non-globular in structure, with the square correlation coefficient (R^2) of 0.79 for the observed data and the power regression line. This and the logarithmic increase in the minimum Gb-index with increasing protein size strongly implied that non-globular characteristics might be more acceptable for smaller proteins than larger ones.

Calnexin (Bergeron et al., 1994) and calreticulin (Michalak et al., 1999) are chaperones that are known to retain inappropriately folded proteins in the ER. The delineation of correct and incorrect folded proteins is known to be the function of another ER enzyme, glucosyl transferase (Ellgaard et al., 2001). The interplay of these three key enzymes

retains incompletely folded proteins in the ER. A significant portion of proteins in the ER are misfolded and translocated back into the cytosol and degraded. (Plemper and Wolf, 1999). How the myriad of numbers of misfolded proteins is recognized individually, and why misfolded proteins are likely to aggregate with each other to make plaques or crystals inside the ER, remains unclear.

565 eucaryotic secretory proteins, which originates from the rough ER and proceeds to Golgi apparatus, showed mean Gb-index of 0.70 (s.d. 0.14) and minimum and maximum Gb-indices of 0.12 and 0.96 each, partly indicating that globularity might be useful in preventing the irreversible aggregation. 31 proteins (i.e. 5.49% of the 565 proteins) showed Gb-index lower than 0.453, which is the threshold value of the least 5% of non-globular proteins. The degree of globularity of these proteins was as strong as, or might have been even stronger than non-secretory proteins considering the smaller sizes with mean atom number of 2686 (s.d. 3285) than the size of the total proteins investigated with mean atom number of 4808 (s.d. 5834) and the correlation of the loss of the globularity with protein's small size.

The globular structure of proteins might help prevent irreversible and pathological aggregation because it has the minimum surface area of a specified volume. The size of the interacting surface area is widely known to strongly correlate with the binding strength. Two perfect globes will have virtually no contacting area because of the convex shapes of both. However, two rodlike proteins would line up side by side with the strong interaction through the long contactable area. The relationship between low protein size and the propensity to lose globular structures might be explained by the smaller contactable surface area of smaller proteins than larger proteins with the same globularity.

It is generally believed that the folding of nascent chain of amino acids is mainly motivated by hydrophobic effect which is through the action of the water molecules. The globular structure is thus a natural conformation considering the minimal surface area

that interacts with the water. There might be contentions about the existence of amino acid sequence level influence that encodes the globular structures. Discussion of the domain structural entity might be helpful for these contentions. Domain is a substructural entity that exists within the structure of the whole protein which can be clearly delineated from other structures as independent. These domains are not a whole protein but might form rather globular conformations. If the globular character is proved to be general in these subprotein domains, it might suggest partial effect of sequence information in the formation of globular structure since the domain does not have obligations to be independently form globes with the interaction with the water molecules. However, the information from sequence might not be a complicated one but be simple seed information of folding start site.

It might be impossible for each domain in the proteins to be independently perfectly globular while coalesce to form the globular structure of the whole protein. The domains should be pressed and be fit with each other by further modifications to form protein-wise overall globular structure. For exceptionally strongly globular domains, the explanations might be provided based on the seeds of folding within the domain. The globular structure formed during the folding which has started from the seed might not have been able to be resolved for its strong stability. For the exact assessment of the globular nature of individual domains, analysis with the SCOP domain conformations provided by ASTRAL database might be appropriate. It is important to only consider the SCOP domains which are the subunit of a polypeptide chain rather than independent chains to scrutinize the possible general globular nature of domains within a whole protein which might prove the assumptive aid of sequence information in the formation of globular structure. The detailed postulations of the roles of sequence information might be possibly made after the certification of the general globularity of subprotein domains.

2.4 Conclusion

We investigated the structural globularity of proteins with an approximate measurement. The results strongly indicated that virtually every protein (95%) was significantly globe-shaped with Gb-indices larger than 0.453. Some oddities were found mainly among small proteins. The small size of the protein and the tendency to have significantly non-globular structure showed a rather high correlation (R^2 = 0.79). The minimum Gb-index showed a logarithmic increase along the increase of protein size (R^2 = 0.62).

The suggestion that globularity might function to prevent aggregation is somewhat intriguing considering the interest in protein aggregation associated with neurodegenerative diseases. Pathogenic aggregations of proteins may have been caused by the loss of globularity of normal proteins or by the overproduction of the proteins with low globularity. Figure 2-3 displays the least globular structure (model 7) of the amyloid-beta peptide (PDB ID: 1BA4), which is known to aggregate and make plaques in neurons in Alzheimer's disease. The Gb-index of this structure was 0.2094, which is in the smallest 0.24% of the 7131 proteins examined. Although confirming that the less globular structure has a primary effect on pathological aggregation is not sufficient, one can still aid in the irreversible aggregation.

This supposition might also be applied to some form of trinucleotide repeat expansion diseases, which also show pathological plaques; i.e., inserted poly-amino acids might induce the loss of globularity in normal proteins. As in the case of Huntingtin in the Huntington's disease, stretch of multiple residues of a single type of amino acid is expanded within the pathogenic protein to far exceed normal quantity. More than 40 glutamine repeat within the allele is required for the full penetrance of the disease. (Walker, 2007) This rather long stretch of a single type of amino acid might possibly cause the disruption of the integrity of the protein structure leading to the loss of the predetermined globularity.

The major coded amino acid in the trinucleotide repeat expansion mediated disease is glutamine as in the case of Huntington's disease, spinal and bulbar muscular atrophy, dentatorubral-pallidoluysian atrophy, spinocerebellar ataxia type 1, and Machado-Joseph disease. Other trinucleotides that are repeated do not resides in the coding region as in the case of fragile X syndrome, fragile XE mental retardation, myotonic dystrophy, and Friedreich's ataxia. The exact effect of the local structure of a poly glutamine stretch to the structure of the whole protein is hard to be clearly anticipated. However, most of the free homogenous amino acid stretch would form helices (Scott and Sheraga, 1966; Ooi et al., 1967). These helical structures would collapse into a lumped structure if the helices are long enough. After the shrink of the lengthy dimension into a lump, further modifications of the local structure would induce extended sheets and other necessary loops formation. It is possible that this newly formed local structure would provide a planar binding surface while pushing other native domains apart.

Though most normal amino acids including glutamine would induce helical or extended sheets, proline repeated stretch would induce more turns and random coil structures as partly supported from its frequent appearance as the secondary structure disruptor and turn structure former. It is possible that plenty of non-regular turns caused by proline would inhibit stable planar β-sheet surface and deter the irreversible aggregations. Lengthy alpha helix and possible helix-turn-helix motif formation might also be deterred by the effect of prolines. Glycine with less steric hinderances from its simple side chain might permit more random configuration which would result into a lumped local domain causing less strong β-sheet or α-helix conformations leading to less strong aggregations.

However, the contention that globularity is the major reason for the pathogenicity of polyglutamine expansion mediated disease should be more cautious because the aggregation is not the only major cause of the neurodegenerative diseases including Huntington's disease. For instance, the toxicity has strongly reduced for the case of the repeat of glutamines from the nucleotide with CAACAG where CAA is also the codon that codes glutamine. (Bonini, 2008) This indicates that the pathogenic mechanism of

the disease does not solely depend on the existence of the repeated single type amino acids.

Our Gb-index has shortcomings in the delineation of some polygonal structures from globular proteins. Although most of the polygons would be implausible for real protein structures, cubic structures and other structures with strongly planar or concave surfaces are still possible. The analysis of the curvature of the surface of a protein would result in a more accurate inference of the degree of globularity for these exceptional cases. Also, the search span of 2° for searching the orthogonal displacement to the longest displacement among all possible atom pairs may be too small and might cause the Gb-index to decrease. Our finding regarding the universality and the distribution of globularity of known protein structures can be used to suggest proper constraints or restraints for protein folding algorithms. The supposition that the deviation of small proteins from general structural globularity might be due to their more tolerance to the aggregation from smaller binding surface might aid in the research of aggregational diseases which are strongly concerned in recent biological and medical research. This analysis of the simple physical character of protein structures might be helpful in anticipating protein structures and in deciphering the underlying mechanisms of protein sysnthesis and many of the aggregational diseases.

Table 2-1. Gb-index of Different Types of Proteins

Type	Mean(s.d.)	*Median*	Mode	Min.	Max.	Number
SCOP classes						
all-α	0.70(0.14)	*0.72*	0.84	0.19	0.99	1365
all-β	0.70(0.14)	*0.71*	0.80	0.14	0.97	1503
α+β	0.73(0.14)	*0.72*	0.76	0.12	0.99	2690
α/β	0.71(0.13)	*0.74*	0.76	0.15	0.99	2067
Multidomain	0.70(0.13)	*0.72*	0.76	0.28	0.96	182
coiled-coil	0.42(0.22)	*0.40*	0.20	0.08	0.86	49
Peptides	0.59(0.16)	*0.58*	0.60	0.24	0.93	129
small proteins	0.69(0.13)	*0.71*	0.76	0.21	0.97	369
Source of Organisms						
Archaea	0.71(0.13)	*0.73*	0.72	0.28	0.98	568
Eukarya	0.70(0.14)	*0.72*	0.72	0.12	0.99	3672
Bacteria	0.71(0.14)	*0.73*	0.76	0.15	0.99	2657
Virus	0.67(0.18)	*0.70*	0.76	0.08	0.95	234
Total	0.71(0.14)	*0.72*	0.76	0.08	0.99	7131

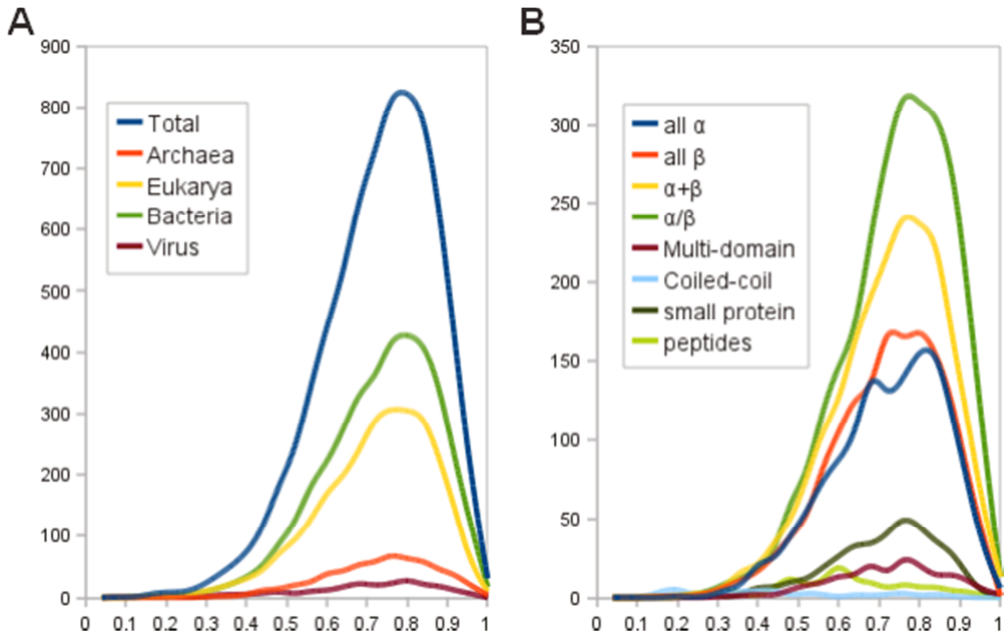

Figure 2-1. Distribution of Gb-indices among Proteins of Different Types and Different Sources of Organisms. (a) Distribution of Gb-indices of proteins from four different types of organisms (archaeal, bacterial, eukaryotic, and viral proteins). All four proteins from different organisms showed similar distributions. Details of the distribution of the values are listed in Table 2-1. (b) The distribution of Gb-indices of eight different types of proteins according to the SCOP classification. Peaks were between 0.7 and 0.8, except the peak of peptides (0.6) and coiled-coil (0.43) proteins.

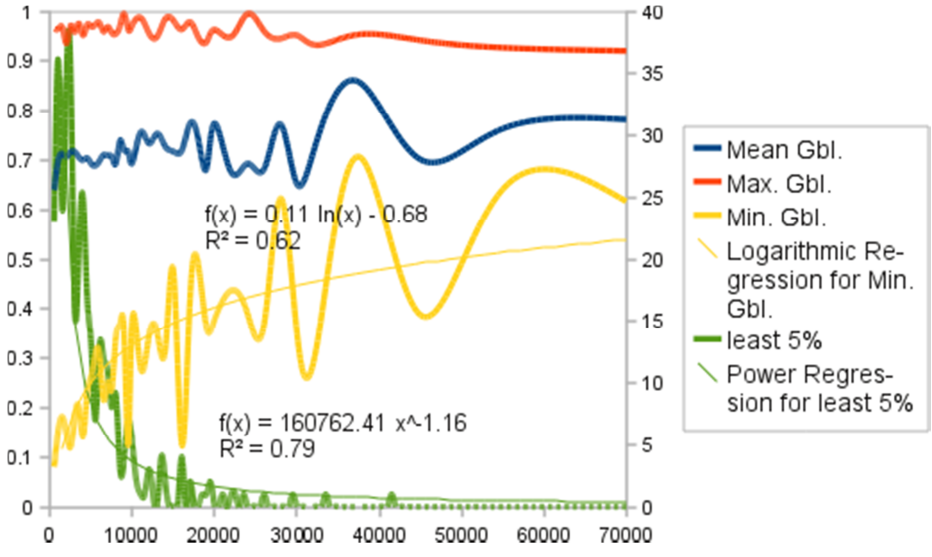

Figure 2-2. Change in Protein Globularity with Protein Size. The mean, maximum, and minimum Gb-indices are plotted against the atom numbers of proteins. As the atom number increased, the minimum globularity measure showed a logarithmic increase, indicating a correlation coefficient of 0.79 with the regression lines. The maximum and mean values, however, stayed rather constant along the whole range of protein size. The relationship between protein size and the minimal globularity index indicates that the non-globular structure might be more permissible in smaller proteins. The number of proteins with a globularity index lower than 0.453 (the lower 5% of non-globular proteins) was also plotted against protein size. A correlation coefficient of 0.62 was shown with the decreasing power regression line, possibly indicating again that the small size permits less globular structures of proteins.

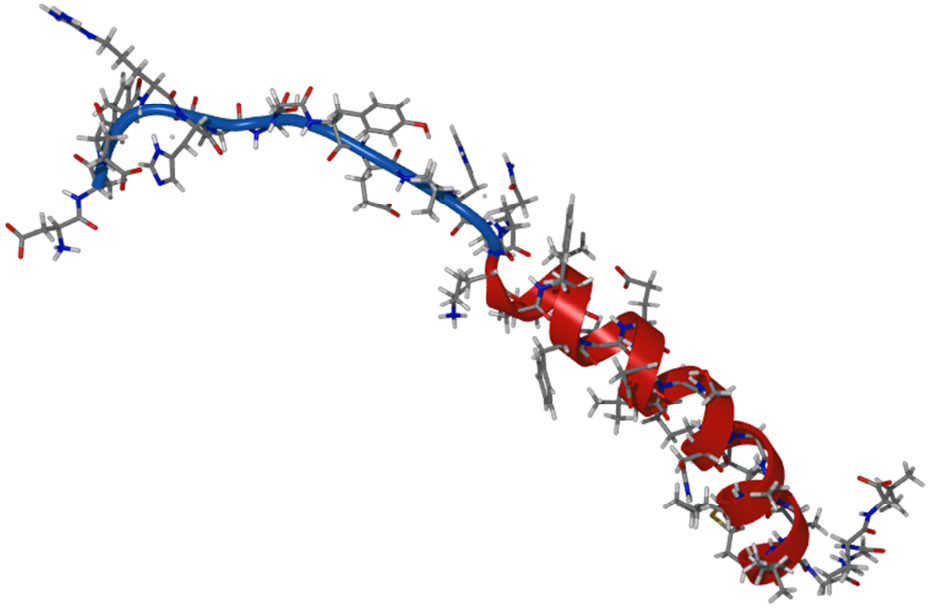

Figure 2-3. Structure of the Amyloid-beta Peptide. The fibrillar structure of the amyloid-beta peptide is shown by a ball-and-stick model and helix ribbons. The Gb-index was 0.2094, which belongs to the lowest 0.24% of non-globular proteins. This non-globular structure of the molecule might aid in irreversible aggregation. Image was prepared with Sirius visualization system.

CHAPTER III.

Validity of Protein Structure Alignment Based on Backbone Torsion Angles

3.1 Introduction

Protein structure has always been a significant concern among molecular biologists because it provides intimate information regarding the function and mechanism of the given protein. This knowledge regarding proteins, which are key molecules in the biology of living organisms, can be used in a variety of ways, ranging from protein structure modeling (Moult et al., 2003; Zhang and Skolnick, 2005) to structural genomics (Skolnick et al., 2000; Baker and Sali, 2001; Marsden et al., 2007). The number of published protein structures has increased to approximately 70 000; this increase represents the interest and perpetuating importance of the knowledge of protein structure for biological and pharmaceutical studies.

Atoms of the molecule would not change their positions along the covalent bond length or to change the angle of covalent bond for the high energy barrier. They instead rotate around the axis of the covalent single bond. Thus, torsion angle system is supposed to regenerate the correct movement of the atoms of molecules. However, it would be still worth to validate the robustness of this space representation system which regenerates the real movement of atoms. As a partial validation method of this system of space description, the torsion angle system was applied to structure alignment in this study.

Numerous structural alignment algorithms have been published. Five of these, namely TM-align (Zhang and Skolnick, 2005), FATCAT (Ye and Godzik, 2003), CE (Shindyalov and Bourne, 1998), MAMMOTH (Ortiz et al., 2002), and TOPMATCH

(Sippl and Wiederstein, 2008; Sippl, 2008), were employed by RCSB as structure alignment service tools (www.rcsb.org). All of these algorithms are similar in their using three-dimensional (3D) coordinates of atoms. Structural alignments that mainly use 3D coordinates take much more time than do sequence alignments, which align 1D sequence strings. Due to their complexity of three-dimensional property, the apprehension of each state during the process of alignment is strictly limited. The operator of the alignment could not firmly aware of the point of improvement of the alignment or any other possible way of improvements with three dimensional numerals. The case is also similar for the computational operations. The computational approach has similar limitations due to the three-dimensional complexity which strongly deters fast perception and modification of states. In typical algorithms of structure alignments, iterative rotations for gradual fitting of two global structures are generally conducted. More complicated modifications to the alignments including local alignment using seed match and extensions like BLAST (Altschul et al., 1990) are, however, strictly impossible with typical 3D methods. Specifically, the 3D character would not permit fast conversion of comparison frames; i.e. every comparison should be made after the massive rotations which are derived from iterative calculations for the selection of matching residues. This necessity for the calculation for every possible comparison frame arises from the nonsequential character of the position information of each residue in the 3D space, where sequentially far residues can adopt very close positions. These limitations strictly prohibit sophisticated manners of structural analysis including comparison.

The limitations which arise from innate complexity of 3D system also influence the speed of operations. Whole genomes of human and mouse can be aligned in approximately 38 days with 100 machines using a well-known sequence alignment tool, BLAST (Kim et al., 2006). If 3D structure coordinates can be transformed into a 1D vector, whole proteomes of human and mouse could be aligned within 1 day with a rapidity similar to that of BLAST analysis with a single computing machine because the

proteome is many times smaller than the genome. Structural genomics and other high throughput analyses including drug target discovery might be possible to become more amenable with this achieved fast speed. These improvements might help for the advances of the structural studies in deciphering functions and interactions among proteins of biological organisms.

Karpen and colleagues (Karpen et al., 1989) and Miao and colleagues (Miao et al., 2008) noticed that a 3D backbone structure can be mathematically represented with a 1D φ and ψ dihedral angle. In addition, it is widely accepted that backbone structural information can be used for structural alignment validation with fair credibility. For example, the widely accepted algorithm TM-align uses only alpha carbon atom coordinates (Zhang and Skolnick, 2005). The notion of Karpen et al. (Karpen et al., 1989) and Miao et al. (Miao et al., 2008) may thus be plausible to be implemented to compare structural similarity between proteins with reliable credibility using fast 1D alignment algorithms.

The utilization of a reduced dimensional quantity for structural alignment using dynamic programming algorithms was previously attempted by Rose and Eisenmenger (Rose and Eisenmenger, 1991). Although Rose and Eisenmenger remarked that torsion angles might be useful for structural alignment based on the Needleman-Wunsch dynamic programming algorithm, they used differential geometry (Rackovsky and Scheraga, 1980; Louie and Somorjai, 1982; Louie et al., 1983; Rackovsky and Goldstein, 1988) of protein chains instead. This differential geometry is more complicated to derive from 3D coordinates than φ and ψ angle values, and its superiority of accuracy and performance is doubtful. Sklener et al. (Sklenar et al., 1989) also attempted to represent the helical status of the backbone structure using atom coordinates of protein backbones, but they didn't use the φ and ψ dihedral information to represent the backbone structure. Recently, YAKUSA (Carpentier et al., 2005) used 1D α angle arrays to reduce the dimension of the comparing information for fast structural alignment with BLAST-like algorithm. SHEBA (Jung and Lee, 2000) uses 1D "environmental profiles" containing information about sequence homology and residue-dependent information such as solvent

accessibility, hydrogen bonds, and side-chain packing as initial alignment, which is then refined for three-dimensional geometry by dynamic programming (Stivala et al., 2010).

Karpen and colleagues (Karpen et al., 1989) showed RMSD of φ and ψ dihedral angles (Δt) between pairs of substructure fragments of two proteins correlates with the RMSD of 3D coordinates (Δr) of the backbone atoms from the alignment using the method of Kabsch (Kabsch, 1976; Kabsch, 1978). Recently, Miao and colleagues (Miao et al., 2008) also showed the higher coverage of local structure alignment based on backbone dihedral angles (φ and ψ angles) with Smith-Waterman dynamic programming algorithm than SSM (Krissinel and Henrick, 2004), DALI (Holm and Sander, 1993), and CE (Shindyalov and Bourne, 1998) with reliable validity proven by the alignment of several of the most challenging pairs of proteins among the 68 pairs presented by Fischer and colleagues (Fisher et al., 1996) and phylogenetic analysis of class II aminoacyl-tRNA synthetases.

These two researches, however, didn't support enough size of test materials for the quantifiable evaluation of the effectiveness of backbone torsion angle alignment algorithm. Karpen and colleagues proved the reliability of their method from the case study of two proteins (i.e. ribonuclease A and the first 124 residues of actinidin) (Karpen et al., 1989). TALI of Miao and colleagues only used four pairs of proteins (i.e. 1cewI-1molA, 1cewI-1r4cA, 1hngB-1a64A, and 1nj8D-1b76A) (Miao et al., 2008). It would be, therefore, a worth attempt to evaluate the accuracy and effectiveness of the structural alignment based on the backbone dihedral angles with large enough test sets, considering the utility of their 1D representation of structural information.

The present study attempted to evaluate the accuracy of the structural alignment with strings of backbone torsion angles using a 1D comparison algorithm by observing the correctness of the classification of homology among 1891 pairs of proteins from three kinds of 62 proteases. Phylogenetic clusterings of 62 proteases were also analyzed for the validation of this approach. Simple gapless global alignment was conducted to

evaluate the appropriateness of backbone dihedral angle method. We used simple geometrical and statistical similarity measurements applying simple arithmetic operations to the angle difference to determine the degree of structural identity.

3.2 Materials and Methods

Phylogenetic and homologic analyses were conducted to test the validity of backbone dihedral angle method. Sequential and structural information of 62 proteases with intermingled homologous groups were used. Detailed descriptions of these proteases are in the following section. Sequence alignment, TM-align, and two backbone dihedral angle difference measurement methods were used to build phylogenetic trees, which might reflect different levels of accuracy by different clustering patterns.

The accuracies of homology delineation of dihedral angle method and that of TM-align were measured and compared after setting optimal thresholds. The performance was measured by ROC values, accuracy (ACC), balanced error rate (BER), the Matthews correlations coefficient (MCC), and other quantities, while the sensitivity and specificity were displayed with TP vs. FP and TN vs. FN plots and an ROC plot. The details of the experimental settings and preparation of materials follow.

3.2.1 Definition of φ and ψ angles

The φ dihedral angle of the i^{th} amino acid is defined as the torsion angle of C_{i-1}-N_i-$C\alpha_i$-C_i, and the ψ dihedral angle of the i^{th} amino acid is defined as the torsion angle of N_i-$C\alpha_i$-C_i-N_{i+1}. Similarly, we can define angle ω as the torsion angle of $C\alpha_i$-C_i-N_{i+1}-$C\alpha_{i+1}$. We assumed ω to be 180˚ because it is usually close to 180˚ with a minor exception of 0˚ due to the partial double bond character. Relative 3D backbone atom coordinates can be accurately determined by simple mathematics using these three angles. The program used to calculate dihedral angles from PDB files was written in JAVA.

3.2.2 Ramachandran plot RMSD (RamRMSD)

We used the Ramachandran plot RMSD (RamRMSD) as the quantity that represents structural similarity based on φ and ψ angles. Similar measurement was used by Karpen and colleagues as Δ*t* (Karpen et al., 1989). It is natural to use the RMSD of points on the Ramachandran plot as a parameter indicating structural similarity because we used φ and ψ angle information for comparison. We calculated RMSD of the Euclidean distance of every two points of matched residues on each of the two Ramachandran plots. The Euclidean distance can be defined as follows:

$$D = (\Delta\varphi^2 + \Delta\psi^2)^{1/2}$$

,where *D* is the distance and

$$\Delta\varphi^2 = (\varphi_1-\varphi_2)^2, \quad \text{if } (\varphi_1-\varphi_2)^2 \leq 180^2$$

$$(360-|\varphi_1-\varphi_2|)^2, \quad \text{if } (\varphi_1-\varphi_2)^2 > 180^2$$

$$\Delta\psi^2 = (\psi_1-\psi_2)^2, \quad \text{if } (\psi_1-\psi_2)^2 \leq 180^2$$

$$(360-|\psi_1-\psi_2|)^2, \quad \text{if } (\psi_1-\psi_2)^2 > 180^2$$

where φ_1 and φ_2 are φ angles from each residue, and ψ_1 and ψ_2 are ψ angles from each residue. Conditional terms are added to find the smallest distance between any two angles with our -180˚ to +180˚ notation; i.e., not to consider the distance of two angles, +180˚ and -180˚, as 360˚ apart rather than 0˚ apart, for example. The RamRMSD would be as follows:

$$Ram\ RM\ SD = \sqrt{\frac{\sum_{k=1}^{n} D_k^2}{n}}$$

where n is the total number of residues to be compared, and D_k is the distance of points of k_{th} residues of each protein on each Ramachandran plot as defined above.

3.2.3 Statistical Similarity Measurement with Weight Imposition

Although RMSD is a common measure of structural similarity, it is weak to small number of local deviations (Zhang and Skolnick, 2005). To circumvent the problems of RMSD, TM-score was used with the Levitt-Gerstein weight factor (Levitt and Gerstein, 1998), which weighs close residue pairs more than distant residues. Here, we defined logPr, which weighs smaller differences, to suggest a possible substitution for RamRMSD, which is vulnerable to local deviations. We defined the probability value (Pr-value) as the probability of finding closer angular similarity than observed similarity in a random environment for each torsion angle pair of compared polypeptide chains, and used logPr (base 10) as our additional informing quantity to RamRMSD; we used Pr rather than P to avoid confusion with the hydrophobicity descriptor **log***P* (Mannhold and Waterbeemd, 2001) or with the P-value for evaluating statistical significance of homology from null hypothesis distribution (Ortiz et al., 2002; Altschul, 1990).

If the difference of the φ and ψ angles is defined as a vector Ω ($\omega_{\varphi 1}$, $\omega_{\psi 1}$, $\omega_{\varphi 2}$, $\omega_{\psi 2}$, … , $\omega_{\varphi n}$, $\omega_{\psi n}$), where $\omega_{\varphi k}$ is the difference of 2 φ angles of the k_{th} amino acid of each n-residue-long string and $\omega_{\psi k}$ is the difference of 2 ψ angles of the k_{th} amino acid of each n-residue-long string, the constant probability density function ρ(ω) and the Pr-value in a random environment can be mathematically written as follows:

$$\rho(\omega) = \frac{1}{180^\circ}$$

where ω is the angular difference, and

$$\mathrm{Pr} = \prod_{k=1}^{n} \left[\left(\frac{1}{180^\circ} \right) \omega_{\varphi k} \left(\frac{1}{180^\circ} \right) \omega_{\psi k} \right]$$

where n is the number of total residues being compared and every angular difference is presumed to be statistically independent. The uniform p.d.f. could be heuristically adjusted using observations from non-homologous alignment data of large enough sizes in further studies. Naturally, if the Pr-value is small, the structural similarity between

two proteins is higher. Because multiplied values range from 0 to 1, the Pr-value is more strongly dependent for small values than for large values. A 180˚ difference has no effect on the Pr-value because the multiplied value is 1, but a 0˚ difference has a critical influence on the Pr-value because it immediately changes it to 0. We heuristically assumed that the absolute 0˚ difference was 1.0×10^{-8} for practical reasons; this was the highest accuracy possible based on the format of our dihedral angle data file.

Although the Pr-value is the original descriptor of the significance of similarity, we used the logPr-value to circumvent a computational overflow problem. We used log base 10 for easy comprehension of the order of magnitude of the probability, Pr.

$$\log \Pr = \sum_{k=1}^{n} \log \left[\left(\frac{1}{180^{\circ}} \right) \omega_{\varphi k} \left(\frac{1}{180^{\circ}} \right) \omega_{\psi k} \right]$$

If the logPr-value is smaller, then they are more similar. The logPr-value of a single residue ranges from -16 to 0. For global alignment, we should normalize the difference in compared amino acid residue lengths. We divided the logPr-value with residue number n and calculated the average:

$$\log \Pr_N = \frac{1}{n} \sum_{k=1}^{n} \log \left[\left(\frac{1}{180^{\circ}} \right) \omega_{\varphi k} \left(\frac{1}{180^{\circ}} \right) \omega_{\psi k} \right]$$

where N denotes a normalized value. A normalized logPr signifies the average logged probability of finding closer alignment between all residue-pairs compared in a random environment.

3.2.4 Alignment Algorithm

We employed a simple alignment algorithm for single-chain proteins. Using the shorter chain as a probe on the template of the longer chain, we moved the probe chain by one residue for each calculation. The probe chain's N-terminus began probing from the template chain's N-terminus. When the C-terminal region of the probe passed through the C-terminus of the template, the probe's protruding C-terminal region was compared to the N-terminal area of the template chain according to the boundary

conditions. That is, where n_1 and n_2 are the lengths of the polypeptide chains S_1 and S_2, respectively, and $n_1 < n_2$, where $S_k(0)$, $S_k(1)$, …, $S_k(n_k-1)$ denote from the first to the last amino acid residues of $S_k(k=1, 2)$, the calculation of values (logPr and RamRMSD) should be as follows:

List 1. Alignment Algorithm

```
for(int i=0; i<n2; i++) {
   for(int j=0; j<n1; j++) {
      if(i+j≥n2) CalculateValue(S1(j), S2(i+j-n2))
      else if(i+j<n2) CalculateValue(S1(j), S2(i+j))
}}
```

During the probing, the calculated Pr-value and RamRMSD were recorded and the alignment frame that yielded the best value was selected. The best alignment frame between the logPr- and RamRMSD-based methods may differ. The alignment program was written in JAVA.

3.2.5 Parameter Settings for Alignments and Clustering

Global alignment with a gap open penalty of 13, extension penalty of 3, and free end gap penalty was conducted for sequences of 62 proteases. A UPGMA algorithm with bootstrapping of 100 replicates was used for tree construction from sequence of proteases. CLC bioinformatics workbench was used for alignment and tree calculation and Geneious workbench was used for graphical representation. (8+logPr), RamRMSD, and (1−TM-score) were used for distance, and a Fitch-Margoliash algorithm was employed for building trees from protein structures. TM-score was normalized by the size of the target protein of the comparison pair. An appropriate integer (8) was added to logPr to make distances positive. Trees were generated from a distance matrix using the

FITCH program of the PHYLIP package. Geneious workbench was used for graphical representation of trees.

3.2.6 Performance-evaluation Quantities

Quantities used to evaluate the performance of the four methods (logPr and RamRMSD of backbone dihedral angle method and RMSD and TM-score measurements of TM-align) were defined as follows(Wei et al., 2010): we considered clustering between the same type of proteases as true, and that between different types of proteases as false. There were 656 true pairs and 1235 false pairs. After setting an appropriate threshold for delineation of positive and negative classes, we defined true positive (TP), true negative (TN), false positive (FP), and false negative (FN). From these, we calculated the true positive rate (TPR), or sensitivity, and the true negative rate (TNR), or specificity; these were defined as:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P_{exp}}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N_{exp}}} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}$$

where P_{exp} and N_{exp} were the numbers of true and false pairs, respectively. The positive predictive value (PPV) and negative predictive value (NPV) were defined as follows:

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{P_{pred}}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{N_{pred}}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$$

where P_{pred} and N_{pred} were the number of positive and negative pairs. ACC and BER were also calculated and were defined as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P_{exp}} + \mathrm{N_{exp}}}$$

$$\mathrm{BER} = \frac{1}{2}(\mathrm{FPR} + \mathrm{FNR}) = \frac{1}{2}(1 - \mathrm{TPR})(1 - \mathrm{TNR})$$

where 1-TPR was the false positive rate (FPR) and 1-TNR was the false negative rate

(FNR), which were defined as:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N_{exp}}}; \mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{P_{exp}}}$$

The MCC (Matthews, 1975) was also calculated and was defined as follows:,

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{\mathrm{P_{exp}}\,\mathrm{N_{exp}}\,\mathrm{P_{pred}}\,\mathrm{N_{pred}}}}$$

After choosing a threshold for positive and negative class delineation referring to the above quantities from various thresholds, we calculated $\mathrm{ROC_{100}}$, $\mathrm{ROC_{200}}$, $\mathrm{ROC_{300}}$, and $\mathrm{ROC_{350}}$ values to assess the ranking quality of each method. ROC values were defined as follows (Lee et al., 2008):

$$\mathrm{ROC_t} = \frac{1}{\mathrm{P_{exp}}} \frac{\sum_{i=0}^{t} \mathrm{T_i}}{\mathrm{t}}$$

where $\mathrm{T_i}$ was the number of true positives ranked ahead of the $\mathrm{i_{th}}$ false positive. ROC curves were drawn and AUROC for each of the four methods were calculated from specificity and sensitivity values of various thresholds. The calculation of the AUROC was conducted numerically. Various grid widths for 20 predictions for specificity were used with the application of midpoint rule.

3.2.7 Test Set Preparation

Although SCOP (Murzin et al., 1990) and CATH (Orengo et al., 1997) classifications are often used as references for the evaluation of alignment quality, some argue that these classifications are so discrete that detailed alignment quality might not be properly assessed (Zhang and Skolnick, 2005). In addition, databases such as CATH use other structure alignment tools for classification (Zhang and Skolnick, 2005), and significant structural similarity has been shown to exist in proteins belonging to different classes (Zhang and Skolnick, 2005; Yang and Honig, 2000; Kihara and Skolnick, 2003). Thus, we used functional classification of proteins as our classification reference, focusing more on the practical utility of the backbone torsion angle based structure alignment algorithm to correctly annotate functions of unknown proteins.

We used PDB files of 62 peptidase proteins with 20 serine-type peptidases (GO ID: 8236), 30 metallopeptidases (GO ID: 8238), 7 cysteine-type peptidases (GO ID: 8234), and 5 aspartic-type peptidases (GO ID: 70001). We chose the peptidase family mainly for its amenable size and the number of subgroups. We selected single-chain proteins without any missing residues. We neglected structures only with alpha carbon coordinates and with modified amino acids whose order of backbone atom coordinates were inverted. Fragmented structures, which compose only a partial portion of the whole protein, were also omitted while selecting proteases. The search tools of the RCSB webpage and JAVA codes were used for searching and selecting PDB files for test set preparation.

3.3 Results and Discussion

3.3.1 Sequence and Structure Trees of Different Groups of Proteases

The structure of a protein is known to have more intimate relationship to its function than does its sequence. If φ and ψ angle alignment is reliable, the pair-wise alignment results should be accurate and the tree built from these alignment distances should be appropriate. We derived phylogenetic trees from proteins with intermingled members of various functional homologies using global alignment of backbone dihedral angles (φ and ψ angle). A total of 62 protein structures of different peptidases as described in the Materials and Methods section were used to construct phylogenetic trees (Figure 3-1). Distances of structure alignments were measured as described in Materials and Methods. Overall correctness could be partly assumed by the strength of clustering of proteins of the same groups without any heterologous interruptions, although strict evaluation of the pair-wise distances may have differed from the aggregation of leaf nodes depending on the branching patterns.

The clustering of structure alignment-based trees using backbone dihedral angle methods of RamRMSD (Figure 3-2b) and logPr-value (Figure 3-2c) showed clustering

with accuracy comparable to that of TM-align (Figure 3-2d) and better than that of the sequence alignment tree (Figure 3-2a). Structure-based trees showed overall concrete distributions of the same homologous group members, while sequence-based trees showed stronger dispersion of serine-type peptidases and metallopeptidases.

A maximum of 14 metallopeptidases were posed next to each other without any interruption of other peptidases in our logPr tree, and a maximum of 9 metallopeptidases were posed next to each other in the RamRMSD tree. TM-align also showed a maximum of 14 metallopeptidases right next to each other without any heterologous interruption. Sequence-based clustering showed a stronger dispersion of metallopeptidases; a maximum of only six metallopeptidases were clustered together. A maximum of nine serine-type peptidases were posed next to each other in TM-align without any interruption, and six were posed next to each other in both of logPr and RamRMSD methods. Sequence-based clustering showed only five serine-type peptidases posed next to each other.

All four methods showed similar clustering among aspartic-type peptidases and cysteine-type peptidases. Omega-amino acid-pyruvate aminotransferase (3a8u) and protein disulfide-isomerase A3 (2alb) of cysteine-type peptidases and hydrogenase 3 maturation protease (2i8l) of aspartic-type peptidases were diverged from the others. Five lactoferins of serine-type peptidases (1b1x, 1ce2, 1i6q, 1lcf, and 1lct) were clustered very closely to each other by all four methods.

Internodes of trees from structural alignments, especially the two trees from backbone dihedral angle methods, showed relatively closer positions to the root compared to the length from leaf nodes to internodes. A comparatively shorter length from internode to root indicated that the structural information was rather discrete compared to the sequence information. This made the difference between different groups of proteins comparatively smaller than the difference between any two proteins. The small difference between the distances from leaf to root and from leaf to internode implies that

a delicate setting of cutoff values would be required for accurate delineation of different homologous group members using structure alignments. This also signifies that structural information that can be employed as characters for clustering is only a small fraction of the total structural information. It is probable that concentrating on the more representative characters, thus discarding the background difference, would yield better results.

3.3.2 Comparison of Backbone Torsion Angle-based Method and TM-align

The trees (Figure 3-1) of 62 proteases were drawn based on the alignment distances of 1891 pairs. Trees drawn with backbone dihedral alignment methods showed reliable results as explained above (Figure 3-1b, 3-1c). However, quantification of the accuracy of dihedral angle method and comparison of this accuracy with other methods is still necessary. Based on our analysis, φ and ψ dihedral angle method showed reliable and even better performance. Among the 1891 pairs of proteins from 62 proteases, protein pairs with the same type of proteases were regarded as true pairs, and pairs with different types of proteases were regarded as false pairs.

The thresholds of each of the four methods to delimit true pairs and false pairs varied from the values that approximately yielded the maximum sensitivity (1.00) and minimum specificity (0.00) to the values that approximately yielded the minimum sensitivity (0.00) and maximum specificity (1.00). An increase in sensitivity generally induced a decrease in specificity during the change of the threshold value. For a proper comparison between methods, we selected the optimum threshold value as that which showed both sensitivity (TPR) and specificity (TNR) of more than 0.5 for TM-align and 0.6 for φ and ψ dihedral angle method with the highest MCC. MCC was used instead of ACC because this test set is imbalanced, having approximately twice as many false pairs as true pairs (Baldi et al., 2000; Murakami and Mizuguchi, 2010). We applied different criteria for the threshold because TM-align could not show both TPR and TNR of more than 0.6 at the same time. log(1/45) for logPr, π/1.9375 for RamRMSD, 5.5 Å for RMSD of TM-align, and 0.285 for TM-score were chosen as optimal thresholds.

φ and ψ dihedral angle methods showed performances comparable to those of TM-align based on the results of these selected thresholds (Table 3-1). The sensitivity (TPR) and specificity (TNR) of φ and ψ dihedral angle methods were above 0.6 as selection criteria. Sensitivities of the methods ranged from 0.62 of logPr and 0.64 of RamRMSD to 0.50 of RMSD of TM-align and 0.52 of TM-score. Specificity was the highest at 0.68 in TM-align RMSD and the lowest at 0.53 in TM-score, while logPr showed a specificity of 0.66 and RamRMSD showed a specificity of 0.63. Of the four methods, logPr showed the highest PPV (0.49), the highest NPV (0.77), the highest ACC (0.65), the lowest BER (0.36), and the highest MCC (0.27), while TM-score showed the lowest PPV (0.37), the lowest NPV (0.67), the lowest ACC (0.52), the highest BER (0.48), and the lowest MCC (0.04). RamRMSD showed similar values to those of logPr for PPV (0.48), NPV (0.77), ACC (0.63), BER (0.36), and MCC (0.26). TM-align RMSD showed similar performance to that of logPr and RamRMSD with a PPV of 0.45, NPV of 0.72, ACC of 0.62, BER of 0.41, and MCC of 0.18.

The overall performance of backbone dihedral angle approach was quite valid compared to that of TM-align, both with logPr and RamRMSD measurements, regarding the above statistics. We further investigated the quality of prediction using ROC_{100}, ROC_{200}, ROC_{300}, and ROC_{350} values, where a higher ROC value signifies better quality. The values are displayed in Table 3-2. TM-align RMSD showed the highest ROC_{100} (0.204), the second highest ROC_{200} (0.246), and the third highest ROC_{300} (0.290) and ROC_{350} (0.313). This signifies that TM-align RMSD was the most accurate in the range of 1st to 100th false positives, but failed to be the best in broader ranges. ROC_{100} (0.153, 0.149) of logPr and RamRMSD were both less than the ROC_{100} of TM-align RMSD (0.204) and TM-score (0.193). However, ROC_{200} (0.251) of logPr and ROC_{300} (0.324, 0.304) and ROC_{350} (0.354, 0.336) of logPr and RamRMSD were higher than the best values of the TM-align methods.

To further evaluate the sensitivity and the quality of the prediction represented with ROC values, we drew a classical chart of TP versus FP (Söding, 2005) (Figure 3-2a). As

can be seen in Figure 3-2a, TM-score and RMSD of TM-align showed better performances in the region from the 1st to approximately the 100th false positive. However, RamRMSD and logPr performed better in the region of the 100th false positive or more. The worse performances of backbone dihedral angle method in the top 100 positive guesses indicates that backbone torsion angle-based anticipations are less robust than TM-align in clearer cases.

We also analyzed the accuracy of negative anticipation. Figure 3-2b shows the number of true negatives along with the increase in the number of false negatives. Backbone dihedral angle method, using both logPr and RamRMSD measurements, showed more valid performances than TM-align methods in all ranges. logPr and RamRMSD showed similar performances with a slightly better performance of logPr. To further analyze performance, we graphed the ROC curves of the four methods using specificity and sensitivity values observed at various thresholds (Figure 3-3). The performances of our two methods (with areas under the ROC curve [AUROCs] of 0.6743 [logPr] and 0.6694 [RamRMSD]) were comparable to those of TM-align RMSD and TM-score (with AUROCs of 0.5965 and 0.5494, respectively).

Backbone dihedral angle methods showed comparable performances, and in some cases outperformed, when delineating the functional homology of the 62 proteases, as shown by the high ACC, BER, MCC, and ROC values. The chart of TP vs. FP and TN vs. FN (Figure 3-2) also demonstrate the comparable performances of this approach. The ROC curve (Figure 3-3) and high AUROC values also support the validity of our new method.

Weighted dihedral angle method (logPr) showed improvement over RamRMSD. However, in this set of 1891 pairs of 62 proteases, the Levitt-Gerstein weight factor (Levitt and Gerstein, 1998)-exploited TM-score performed worse than did non-weighted TM-align RMSD, especially in the obscure cases of delineation pairs, which is shown in Figures 3-2a, 3-2b, and 3-3. TM-align aligns two proteins with TM-score-based heuristic iterations and uses RMSD only as an optional quantity; i.e., the different performance

only depends on the application of the weight factor to the distances of the aligned residues. This implies that weighting of closer similarity based on 3D coordinates might mislead the delineation of homology in difficult pairs, indicating that local deviations might be important information in less significant cases. Weighting on closer backbone torsion angle similarity, however, did not distort the appropriate alignment, as can be seen by the high performance measurements in Tables 3-1 and 3-2 and in the sensitivity (Figure 3-2a), specificity (Figure 3-2b), and ROC curve (Figure 3-3) graphs, signifying that distance based on backbone torsion angle information is more robust for comparison than that based on 3D information.

Backbone dihedral angle approach showed reliable accuracy compared to sequence alignment, as shown in Figure 3-1, and with TM-align, as shown in Figures 3-1~3-3 and Tables 3-1 and 3-2. In addition, the Spearman rank correlation coefficient and Pearson's correlation coefficient of the pair-wise comparison from the four methods were calculated (Table 3-3) for further validation of backbone dihedral angle method. The correlation between our two methods of logPr and RamRMSD and TM-align RMSD ($r = 0.53$ and 0.55; $r_s = 0.45$ and 0.47) was stronger than the correlation of each with TM-score ($r = 0.41$ and 0.44; $r_s = 0.13$ and 0.16). The rather solid correlation of TM-align RMSD and TM-score with backbone dihedral angle methods partly indicates the validity of our new approach. Backbone torsion angle method showed very high correlation between the two measurements (logPr and RamRMSD) based on both the Pearson's (0.95) and Spearman's rank correlation coefficients (0.92), higher than those between TM-align RMSD and TM-score ($r = 0.56$ and $r_s = 0.33$).

3.3.3 Clustering Trees and Accuracy Analysis with Delineation Set of 30 Kinases and 30 Proteases

The robustness of the backbone torsion angle alignment method was partly validated by the clustering analysis of 4 types of proteases as shown in the section above. Proteins of more distantly homologous groups were used for the further validation. A mixed set of

30 kinases and 30 proteases were employed for the clustering (Fig. 3-4) and accuracy analysis of delineations. Trees of clustering from logPr (Fig. 3-4a), RamRMSD (Fig. 3-4b) showed apparent delineation of kinases from proteases. The distance between these two groups are very long that one can easily recognize the separation of each from the other. The distances within each group of kinases and porteases were rather similar in the case of the tree from logPr measurements while the tree from RamRMSD showed much shorter distances between proteases than the distances between kinases. The difference between the distances of two groups might have originated from the weighting of smaller distance in the case of logPr, which might have caused the distances between kinases to shrink to be relatively more similar to the distances between proteases. As being observed in the case of 62 proteases set, logPr and RamRMSD tree showed smaller distance between internodes than the distance between terminal leaves and the last internode indicating that structural information is quite homogenous within the group of identical functional homology. The clear discrimination of two homology group members indicates that structural classification is strongly robust in the case of lucid functional difference especially with our new method. The rather obscure delineation of previous 62 proteases set might have originated from the very similar functional homology among 4 clustered subtypes of proteases and the informational homogeneity among the members of a group of functional homology.

Clustering tree from TM-score measurement did not show clear delineations between kinases and proteases while it showed some aggregation of kinases as subcluster. Though 22 kinases were posed close to each other, members of this major cluster of kinases were in very proximity to the members of proteases group, making them hard to clearly separate referring the pairwise distances. Furthermore, 8 kinases were posed within the clusters of proteases. The overall distribution of pairwise distances of 60 proteins was rather even. This incomplete separation markedly shows the better ability of our backbone torsion angle method than the TM-align method. The inferior performance of TM-align might be due to the inconsideration of connectional

information among matched Cα atoms. Neglect of this information makes an algorithm crucially vulnerable to the similar ostensible shape with different topology of the backbones. According to our previous finding in chapter 2, the proteins generally adopt strong globular structure and the most significant difference among proteins is the topology of the backbone structure. Thus, there is a possibility that TM-align might have been misled by general globular configuration of the positions of Cα atoms. This indicates the robustness of backbone torsion angle alignments and the possible problems of 3D methods which does not considers the information of the connections of each matched reference point.

The numerical measurement of accuracy of both of our new backbone torsion angle based method and the typical 3D method of TM-align was conducted with ROC curve analysis (Fig. 3-4d). Surprisingly, our new method with both logPr and RamRMSD measurements showed perfect accuracy with AUROC of 1.0 which means that every pairwise distance might be correctly classified as true and false pairs. This clear discrimination is displayed in the trees of figure 3-4a and 3-4b. TM-align method with RMSD and TM-score measurements, however, showed no marked improvements in this more clearly distinctive set than previous set of protease subtypes possibly indicating the less robustness of 3D methods. This is shown in the similar AUROCs of RMSD (0.6846) and TM-score (0.6319) to the previous ones.

The statistical analysis performed with ROC curve graphs are rather seems inappropriate regarding the results of the perfect accuracy of the delineations of our method. In fact, the compared test set should be clearly discriminated for the strongly different topology of the backbones which also reflects the absence of homology of sequence information. However, TM-align failed to correctly reflect the starkly different topology of the backbone structure by erroneously considering discrete Cα positions. This approach might be appreciable if proteins are of diverse morphology. However, structures of proteins are mostly spherical as shown in the previous chapter. Thus, missing the connectional information of the polypeptide backbone structure might mislead the

classification. The rather erroneous result of TM-align might be a valuable illustration of the importance of the backbone information from this typical and clear discrimination accuracy analysis.

3.3.4 Computational Time and Complexity

The computational complexity of alignments could be reduced to O(nm) with pre-calculated dihedral angle arrays from $O(m^2n^2)$ of typical 3D coordinate-based alignments, where m and n is the length of the compared proteins. Computation time of backbone dihedral angle methods was calculated and drawn (Figure 3-4) from the results of 1891 pairs of 62 proteases. Both the logPr and RamRMSD methods showed linear relationships with R^2 of 0.83 (logPr) and 0.69 (RamRMSD), with the search space calculated by multiplying the lengths of each peptide chain of the pair-wise comparison. The logPr method took slightly more time than RamRMSD. The mean and median CPU times of 94.42 and 90 ms each for logPr and 79.14 and 80 ms each for RamRMSD were needed to calculate a pair of proteins among the 1891 pair-wise comparisons with 3.0 GHz AMD phenome processor on an openSUSE 11.2 platform. TM-align took an average CPU time of 754.30 ms to calculate one pair of comparisons in the same environment.

Although backbone dihedral angle method was approximately 8-fold (logPr) or 10-fold (RamRMSD) faster on average, TM-align tended to be much slower when the size of the compared protein pair increased. For example, the pair with the largest search space of 809568 ($res.^2$), 1Q2L (939 res.), and 2GTQ (867 res.) consumed only 220.0 ms (logPr) and 130.0 ms (RamRMSD) using backbone dihedral angle methods, but took 9160 ms with TM-align, which is approximately 40 times slower than logPr and approximately 70 times slower than RamRMSD. Considering that our JAVA program needed an interpreter (JVM) to perform the calculation, the rapidity of backbone dihedral angle algorithm might be more than proved here. Applying more sophisticated sequence alignment algorithm, however, would consume more computational resource than this

simple performance evaluating algorithm. The average and median values of the search space were 2.98×10^5 and 1.38×10^5 (residue^2).

3.4 Conclusion

Backbone dihedral angle approach is possible to be considered as being robust based on the results from 1891 pairs of proteins as presented herein. BLAST and other methods can be applied with minor modifications as shown by the case of YAKUSA (Carpentier et al., 2005) with comparable rapidity as sequence alignment by changing the 3D backbone structure to 1D torsion angle strings. Though the rapidity and validity of the backbone dihedral angles approach is comparable and even better for more obscure comparisons than famous 3D alignment TM-align as shown here with 62 proteases and 60 proteins of kinases and proteases, this approach's robust performance is currently not very much appreciated. The result of better accuracy of obscure cases of alignments might be due to the general globular structures of usual proteins; i.e. the comparison of topological characteristics of backbone which consist the ostensible sphere structure might provide more information than the comparison of broad 3D globular shapes especially in the case of marked functional difference.

Three-dimensional representation of structures strongly limits human apprehensions and sophisticated computational analysis. In this study, 3D protein backbone structure was converted into 1D torsion angle strings to allow more amenability and rapidity to the structural analysis. The outstanding rapidity of about 10 folds and comparable or better accuracy of 1D backbone torsion angle method was validated here. In native environment, change of covalent bond length and angles are very rare while torsional movement along the axis of covalent single bond is general. The regeneration of this native movement by implementing torsional space system is possible to be regarded as valid from the application of the system to the structure comparison in this study with fair credibility. The fast speed and better accuracy for functionally different proteins and

obscure cases of the same function proteins of the method developed in this study might contribute to the more massive and complicated analyses for large scale structural genomics.

This method could also be further enhanced by, for example, cumulating φ, ψ, and ω angles for exact backbone structure matches to improve accuracy. Future studies might consider investigating the use of numerous possible weighting schemes. Regarding the validity of backbone dihedral angle alignment in structure comparison proven here and its simplicity which can be further exploited, we are hopeful that this approach could be used as a reliable basis in structure related protein researches.

Table 3-1. Performance of the Four Methods

Methods	TPR	TNR	PPV	NPV	ACC	BER	MCC
logPr	0.62	0.66	0.49	0.77	0.65	0.36	0.27
RamRMSD	0.64	0.63	0.48	0.77	0.63	0.36	0.26
TM-RMSD	0.50	0.68	0.45	0.72	0.62	0.41	0.18
TM-score	0.52	0.53	0.37	0.67	0.52	0.48	0.04

Table 3-2. ROC values of the Four Methods

Methods	ROC_{100}	ROC_{200}	ROC_{300}	ROC_{350}
logPr	0.153	0.251	0.324	0.354
RamRMSD	0.149	0.229	0.304	0.336
TM-RMSD	0.204	0.246	0.290	0.313
TM-score	0.193	0.241	0.277	0.293

Our logPr and RamRMSD showed worse performance for the clearer cases (protein pairs before 100^{th} false positives) but showed comparable accuracy for more difficult cases (protein pairs after 100^{th} false positives) as can be seen by high ROC_{300} and ROC_{350} values.

Table 3-3. Pearson and Spearman Correlation Coefficients

	logPr		RamRMSD		TM-RMSD		TM-score†	
	r	r_s	r	r_s	r	r_s	r	r_s
logPr	1	1	0.95	0.92	0.53	0.45	0.41	0.13
RamRMSD	0.95	0.92	1	1	0.55	0.47	0.44	0.16
TM-RMSD	0.53	0.45	0.55	0.47	1	1	0.56	0.33
TM-score	0.41	0.13	0.44	0.16	0.56	0.33	1	1

†We inverted the sign of TM-score values because TM-score scores closer distance with higher TM-score making the correlation with others negative.

r: Pearson's correlation coefficient

r_s: Spearman's rank correlation coefficient

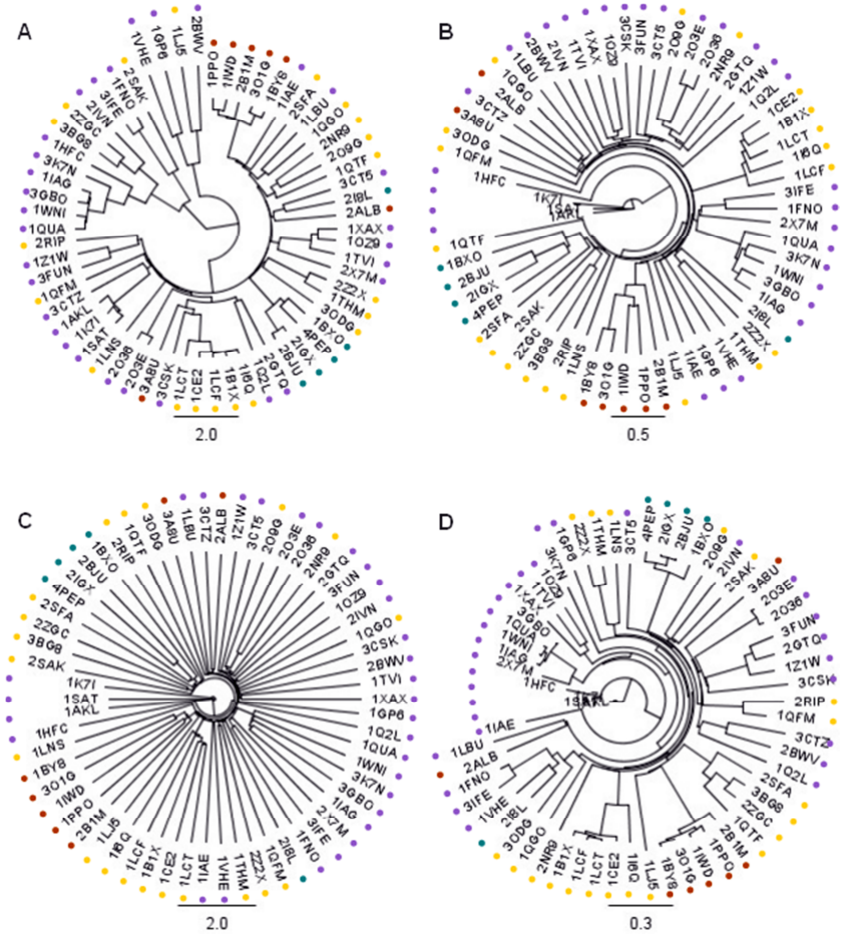

Figure 3-1. Phylogenetic Trees of Different Types of Proteases. Phylogenetic trees of different proteases were built from sequence analysis(a) and structure analyses including backbone dihedral angle structure alignment method(b,c) and TM-align (d). Sequence alignment generated rather obscure clustering between serine-type proteases (yellow dots) and metalloproteases (purple dots). Aspartic-type proteases and cysteine-type proteases were dotted with cyan and red color each.

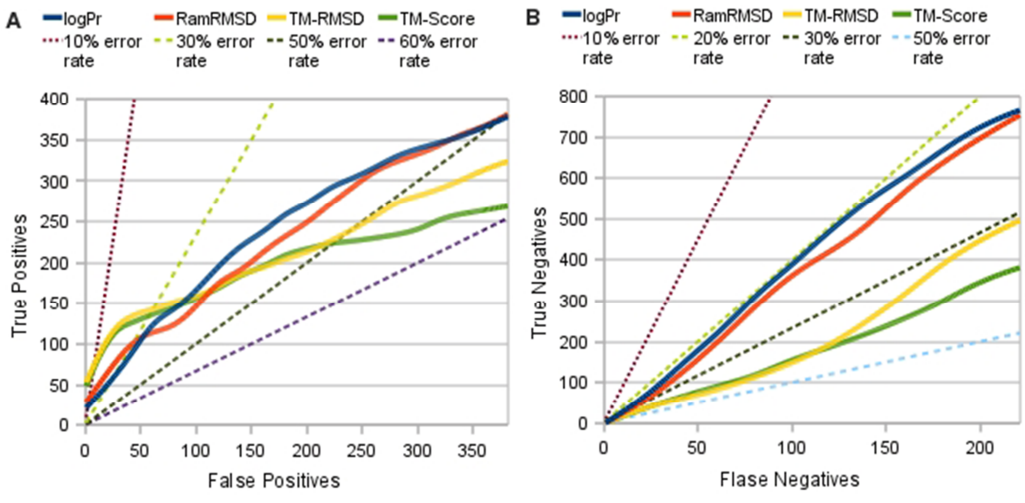

Figure 3-2. Performance Displayed by TP vs. FP and TN vs. FN plot. Curves tilted to upper left indicates better accuracy. In TP vs. FP plot(a), backbone dihedral angle methods (logPr and RamRMSD) showed comparable performance to TM-align methods, performing worse for clearer cases but better for more obscure cases. In TN vs. FN plot(b), our methods showed better performance than TM-align methods for all the cases. Dashed lines signifies error rates.

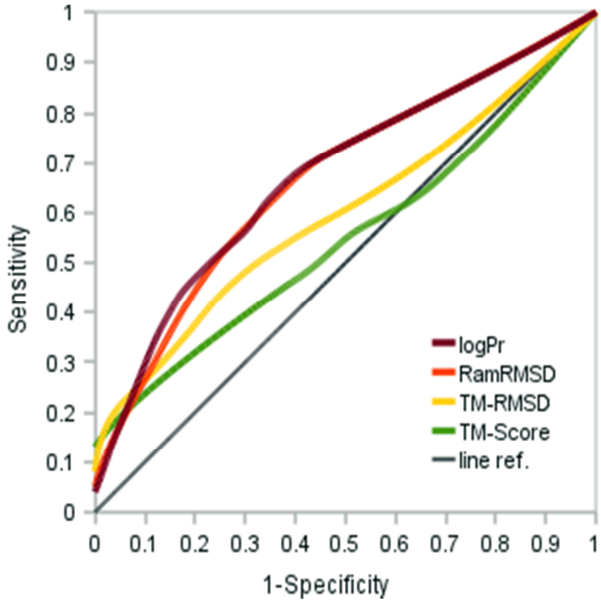

Figure 3-3. ROC curves of Different Methods. logPr and RamRMSD showed similar performance with AUROC of 0.6743 and 0.6694 respectively. This was comparable with the performance of TM-align which showed AUROC of 0.5965 for TM-align RMSD and 0.5494 for TM-score. AUROC was calcaulted with the grid with of 0.02 following rectangle, or midpoint, rule.

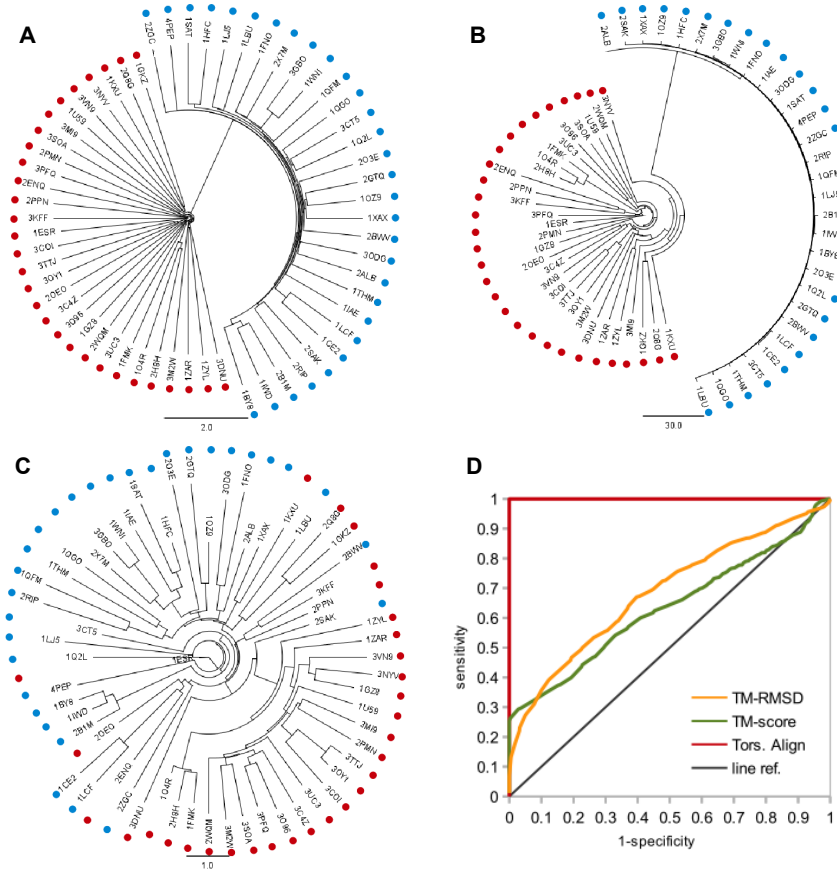

Figure 3-4. Clustering trees from 30 kinases and 30 proteases and the accuracy of the four methods. The clustering trees of 30 kinases (red dots) and 30 proteases (blue dots) built with logPr (A), RamRMSD (B), and RMSD of TM-Align (C) are displayed. Backbone torsion angle methods showed perfect delineation of kinases and proteases while TM-align showed promiscuous clusterings with partially correct cluster of kinases; i.e. cluster of 22 kinases is not clearly separated from proteases. ROC curve (D) shows the perfect performance of backbone torsion angle alignment with AUROC of 1.0 while TM-align method showed similar performance than that of the 62 protease set with AUROC of 0.6846 (RMSD) and 0.6319 (TM-score).

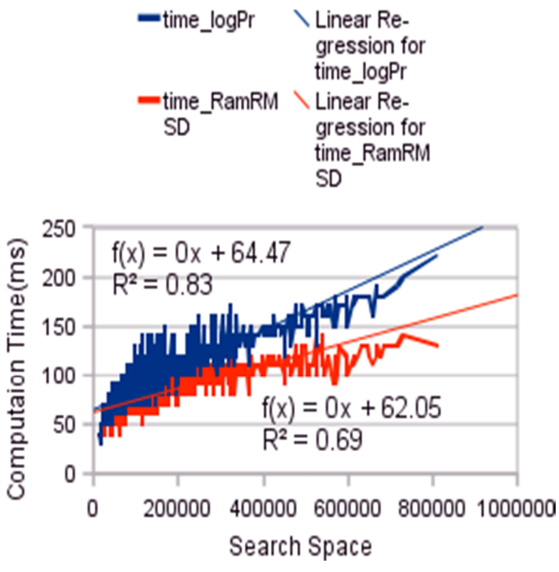

Figure 3-5. Computation Time Along the Search Space. The CPU time of logPr was slightly longer than that of the RamRMSD. The relationship between the CPU time and the space of surveillance was linear with high correlation coefficients (0.83 for logPr and 0.69 for RamRMSD).

CHAPTER IV.

Secondary Structure Information Repository from Backbone Torsion Angle

4.1 Introduction

Protein is the primary component of the biological system of living organisms. Protein structure is hierarchically classified as primary, secondary, tertiary and quaternary structures. Primary sequence refers the sequence of amino acid residues and tertiary structure refers the three-dimensional structure. Quaternary structure originates from the combination of the three dimensional structure moieties. Tertiary structure can be regarded as the topological organization of smaller local structures. About 500 common folds were found from about 20000 proteins and about 1000 common folds were estimated to exist in all possible proteins. (Baxevanis and Ouellette, 2005) The secondary structure of a protein might be considered as the local structure with repetitive hydrogen bonds which consists the tertiary structure when combined.

Secondary structure is determined by numerous factors. These factors include inter-residue interaction, backbone-backbone interaction, solvent interaction (Pauling et al., 1951), hydrophobic or hydrophilic interaction, and global or local interaction. (Baxevanis and Ouellette, 2005) Secondary structure is usually classified into three categories of helix, strand, and other. Helix has backbone of cork-screw like spirals with projecting side chains. Typical example of helix secondary structure is alpha-helix structure. This structure is the most common secondary structure. There are 3.6 residues in a turn and the helix is sustained by the hydrogen bond between carbonyl (CO) group and amino (NH) group of distinct amino acids.

Strand has extended backbone structures created by zig-zag pattern. This zig-zag pattern is different from helices in the alternating direction of backbone torsions. Typical example of strand secondary structure is beta-strand. Two or more stretches interact through hydrogen bonds to create beta-sheet structure. Other structure is the local structures that does not fall either into helix or strands. Sometimes more fine classification is used for some classification applications. DSSP uses 7 categories of α helices (H), 3/10 helices (G), π helices (I), β bridges (B), extended β strands (E), hydrogen bonded turns (T), and bends (S) for the classification of secondary structures. (Kabsh and Sander, 1983) Classification scheme of five classes is also usually used. In this scheme, β-turns (with sharp chain reversals), omega loops (with loops which resemble Greek letter omega), and 3/10 helix are additionally used to α helices and extended β strands. (Baxevanis and Ouellette, 2005) Though most of the amino acids could be classified into these categories, there are some amino acids that don't belong to any category. These unclassifiable structures are called as random coil or unstructured region.

Many prediction algorithms and applications were developed to predict the secondary structure from amino acid sequences. Early studies focused on the different propensity of individual amino acids emphasizing the importance of local environment. (Baxevanis and Ouellette, 2005) Global interaction, however, might play important roles in secondary structure formation. Hundreds of new ideas are now using various biochemical insights and computational and mathematical algorithms to anticipate secondary structure.

PHDsec (Rost et al., 1994; Rost et al., 1996) and PROFsec (Rost et al., 2003) of PredictProtein (Rost et al., 2003) service use feed-forward artificial neural network algorithm for the prediction. These two are basically the same, while PROFsec is an elaborated version. Feed-forward neural network is the simplest artificial neural network with no cycles but only with flow-through processes. When the query sequence is input, sequence homologues are searched. These homologs are aligned by MaxHom algorithm.

The evolutionary history of conservation and substitution of amino acid residues is obtained from this alignment. The features obtained from this alignment are input to the artificial neural network. The alignment step is the most important and sensitive step in PHDsec and PROFsec for the determination of accuracy. PSIPRED (McGuffin et al., 2000) is similar to the PHDsec. After the query sequence is input, PSI-BLAST is used to search homologues instead of MaxHom. The features obtained from sequence alignment are input to the artificial neural network. Newer version explored support vector machine instead of artificial neural network. (Ward et al., 2003) SVM (support vector machine) stand alone application showed similar accuracy to the neural network system and combining of the two yielded significantly better accuracy. (Ward et al., 2003) SAM-T99 (Karplus et al., 1998; Karplus, 2003; Karplus, 2005) is similar with the previous two applications. When a query sequence is input, the application searches and aligns homologues using hidden Markov model approach. Features derived from this alignment are input to the machine learning algorithm. (Baxevanis and Ouellette, 2005) Hidden Markov model endows more sensitivity to detect remote homologues than other methods including blast derivatives. (Lee et al., 2008, Baxevanis and Ouellette, 2005)

Currently, there exist protein fold classification databases which partly exploit automatic structure analysis tools. Secondary structure is one of the most important criteria that are used for the categorization. CATH (Class, Architecture, Topology, Homology; Pearl et al. 2000) and SCOP (Structural Classification Of Proteins; Murzin et al. 1995; Hubbard et al., 1999; Lo Conte et al., 2000; Andreeva et al., 2004; Lo Conte et al., 2002) are one of the most well-known hierarchical classifications of common protein folds. These databases revealed unexpected relationship among distant proteins previously from sequence analysis including convergent evolutions. (Baxevanis and Ouellette, 2005)

Classification criteria of CATH are similarity of structure and sequence and secondary structure content. (Pearl et al. 2000) It is classified hierarchically and built from high resolution (<3.0Ǻ) proteins and domains. (Pearl et al. 2000) The highest level of hierarchy is Class level which is automatically determined by secondary structure

content. (Pearl et al. 2000) Three categories of mainly-alpha, mainly-beta, and alpha/beta are in the Class level. (Pearl et al. 2000) Architecture level is manually classified by overall domain shape and orientation of the secondary structure. (Pearl et al. 2000) Topology level further classifies the protein structures by dividing according to the secondary structure connectivity and general shape using automatic SSAP algorithm. (Pearl et al. 2000) At the bottom of the whole hierarchy, the categories are clustered according to sequence identity (>35%) and length of the sequence match (>60%). (Pearl et al. 2000)

SCOP is primarily manually classified with the aid of computational tools. (Murzin et al. 1995) It has four-partite hierarchy of Class, Fold, Superfamily, and Family. (Murzin et al. 1995) Class level is consisted with 11 classes of all alpha proteins, all beta proteins, alpha and beta proteins (a/b), alpha and beta proteins (a+b), multi-domain proteins (alpha and beta), membrane and cell surface proteins and peptides, small proteins, coiled coil proteins, low resolution protein structures, peptides, designed proteins. These classes are classified based on secondary structure content and protein size. (Murzin et al. 1995) Class named as alpha and beta proteins (a/b) is consisted with mainly parallel beta sheets (i.e. beta-alpha-beta units). Class named as alpha and beta proteins (a+b) is consisted with mainly anti-parallel beta sheets (i.e. segregated alpha and beta regions). Multi-domain proteins contain folds which have two or more domains belonging to different classes. Membrane and cell surface proteins and peptides do not include proteins in the immune system. Class named as small proteins is usually dominated by metal ligand, heme, and/or disulfide bridges. Designed proteins contain experimental structures of proteins with essentially non-natural sequences.

Protein folds are classified into Fold if entries of the same Class have the same major secondary structures in the same arrangement and with the same topological connection. (Murzin et al. 1995) Same Fold does not signify the same origins, but does signify the similar topology. (Murzin et al. 1995) Next hierarchy of classification to the Fold is Superfamily. This is classified based on the sequence or structure similarity. (Murzin et

al. 1995) The elements of the same Superfamily are suspected to come from the common evolutionary origin. (Murzin et al. 1995) The lowest level of the hierarchy of classification of SCOP database is Family. The entries of this lowest classification level are suspected to have clear evolutionary relationship among them. (Murzin et al. 1995) These are classified based on the matter of having common structure, common function, or sequence identity of more than 30%. (Murzin et al. 1995) These fold classification databases are strongly influenced by computational structure comparison tools. Structure alignment based on backbone torsion angle is introduced in the previous chapter. We would like to introduce database of secondary structure in this chapter. The most well known secondary structure database would be the database of DSSP (Dictionary of Secondary Structure of Proteins). The DSSP (Define Secondary Structure of Proteins) algorithm of this database is first described in 1983 and is one of the standard tools for evaluation and annotation of protein structure. (Baxevanis and Ouellette, 2005) DSSP produces compact summary of local protein structural features along the amino acid sequence. DSSP uses very stringent methods to identify hydrogen bonds and bonding patterns. DSSP is *de facto* reference for PDB database and other tools. (Baxevanis and Ouellette, 2005)

Here, we constructed secondary structure database from 92998 PDB chains and 64799 SCOP entries based on the simple classification scheme according to the backbone torsion angles. SCOP entry 3D structure file was obtained from ASTRAL PDB style database (Brenner et al. 2000; Chandonia et al., 2002; Chandonia et al., 2004). The database introduced here offers functions of secondary structure database searching, secondary structure calculation, and pair-wise protein structure comparison. Secondary structure query can specify the helix, extended, and other structure content. Sequence length is also could be specified. Secondary structure of a specified ID of PDB and SCOP is also possible to be drawn. One can upload protein structure file to calculate its secondary structure. Protein structure comparison tool supports pair-wise comparison of proteins based on the backbone torsion angles.

4.2 Materials and Methods

The database uses web-server based architecture. User interface of the database is accessible through internet browsers. The server was built on workstation with quad core AMD phenom II CPU (3.0GHz) and operating system of openSUSE 11.2. This server exploited Glassfish 3.0 server application for web infrastructure and MySQL 5.1 for database construction and data retrieval. JAVA programming language and JSP language was used for the main component of the web application which serves the secondary structure database through internet. MySQL JDBC 5.1 was used for the connection between java language and MySQL.

4.3 Results

4.3.1 User Interface and Architecture

The first page of the web application is index.jsp page (Figure 4-1). This page contains welcome statement and necessary explanations of the whole application. The index.jsp page displays three functionalities of secondary structure database search, secondary structure calculation, protein pair-wise comparison using backbone torsion angle. This page also displays the limitations of database. Proteins containing nucleic acids, containing no backbone nitrogen atoms, proteins with missing residues, with abnormal backbone atom sequence, with alternative atom locations, and of one or two amino acids long are omitted. User of the web application can jump to the specific utility page by clicking the hyper-linked text of index page. If user clicks secondary structure database search hyper-link, then the web application leads user to “sss.html” page. When user clicks secondary structure calculation hyper-link, “ssc.html” page appears. “tasa.html” appears for protein pair-wise comparison.

“sss.html” page is titled as “Search Interface of Secondary Structure” (Figure 4-2). This page explains how one can perform searching by specifying query criteria. One can specify the database from which the search would be conducted. ASTRAL PDB-style SCOP database and isolated chains of PDB entries database are supported. For the

support of PDB-style data of ASTRAL, it was very convenient to implement secondary structures of SCOP entry. Selection of the database is enabled through selection list implemented into the “sss.html” file. The selection of either value of SCOP or PDB is stored into variable named DB and sent to “sss.jsp” file after the submit button is clicked.

User can also delimit secondary structure content and amino acid sequence length by selecting check box of each argument and input the span of the percentage of structure content and amino acid length. The “input” tag of “text” and “checkbox” type of html was used for these delimitations. Secondary structure is categorized into helix, extended, or other structure. User can directly search the secondary structure of PDB and SCOP entry by specifying PDB and SCOP ID. User also should signify the chain name after the PDB ID. User can order the results according to the alphabetical order of ID, helix, extended, other content, and sequence length. This ordering method can be set by the selection list above the “submit” and “reset” buttons. The default ordering method is to ignore ordering. “submit” button sends variables in which values are stored to “sss.jsp”. “submit” and “reset” button is enabled by “submit” and “reset” tag of html file.

After the selection of query criteria and clicking “submit” button, the result page generated from “sss.jsp” file appears (Figure 4-3). If there is any error in the query, the page is redirected to the search interface page. It first summarizes the query that user has sent to the application. The selected database, helix content, extended structure content, other structure content, sequence length span is displayed. This page also displays the MySQL query made from query criteria, specified ID if there is any, and the ordering method. Result of the query is displayed in the result table. The table has seven headings of ID, amino acid sequence, secondary structure, helix (%), extended (%), other (%), and amino acid length. These headings are colored in blue. The content of the result is colored in orange. The sequence of amino acid and secondary structure is displayed by 20 characters for each line.

Users can upload their own protein structure file and analyze secondary structure

through secondary structure file upload interface of “ssc.html” file (Figure 4-4). Secondary structure file upload interface explains the limitations and shows the file upload dialog. The application is capable only for single chain structure. Thus, PDB file with multiple chains should be edited manually before uploaded. The relevant explanations are written in “sss.html” page. Alpha carbon only file, nucleic acid containing file, alternative atom position containing file, missing residue containing file, and backbone atom disordered file are also not supported by the application. File upload interface is enabled by “file” type “input” tag of html.

After the file is uploaded, the result page of “ssc.jsp” appears. This page explains the content of the result and shows the result of checking of errors of the file. Erroneous cases include multi-chain containing error, alternate atom position containing error, nucleic acid containing error, backbone atom order error, alpha carbon only error, and missing residue containing error. Error check table shows the result of the error test for each of these cases. If the test result is valid, word “OK!” is printed next to the error case label. Else if the test result is not valid, word “Failed!” is printed. The result table contains headings of amino acid sequence, secondary structure, percentage of helix, extended, and other structure, and amino acid sequence length. Relevant result of calculation is printed under each heading. Each line contains 20 characters at maximum.

User can compare two protein structures through protein structure comparison interface of “tasa.html” file (Figure 4-5). This interface page describes the function of interface and query type. Pre-deposited structures are possible to be compared by specifying IDs of either PDB or SCOP database. User should also type chain identifier when input PDB ID for this database has isolated each chain from structures with multiple chains. Text type input tag of html was used for information input. File uploading for comparison is also optional. File type input tag of html was used for information upload. Uploaded file is temporarily stored with the file name of uploading time formatted in millisecond. These temporarily saved files are not to be abused for malicious intentions including treachery on novel structural findings.

Input data is sent to “tasa_DB.jsp” page and “tasa_file.jsp” for ID specification case and file upload case each. If there is no deposited structure with the same name as input ID, “tasa_DB.jsp” page shows error remark indicating possible error in the input name of IDs. “tasa_file.jsp” page shows error remark if the files are not properly uploaded and file contains multiple chains, nucleic acids, missing residues, disordered backbone atom sequence, alternative atom position, and only alpha carbon. The error remark recommends editing of inappropriate files with text editors.

Result of backbone torsion angle based structure alignment is displayed in either “tasa_DB.jsp” page of “tasa_file.jsp” page (Figure 4-6). The alignment algorithm was named as TASA (Torsion Angle based Structure Alignment). In this algorithm, the structural similarity is presented as RamRMSD and logPr value. RamRMSD is the RMSD (Root Mean Square of Deviation) of distance of points in two Ramachandran plots of each protein compared. (Jung et al., 2011) logPr signifies the average probability of finding more similar torsion angle difference assuming the distribution of difference is uniform along the span of π radians. (Jung et al., 2011) logPr value ranges from -16 to 0 where -16 is a heuristic integer which substitutes mathematical infinity. Result table has blue headings of logPr, RamRMSD, calculation CPU time of logPr and RamRMSD each, amino acid sequence length of the first protein and second protein each, and two-dimensional search space. Content of the relevant result is printed below to the headings in orange color.

4.3.2 Computational Mechanisms

The application is mainly consisted with two major parts of user interface page of html file and query searcher and result viewer page of jsp file. All information which is input into html page is transported to relevant jsp files through the implementation of “action” attribute of “form” tag of html with “post” method. “post” method was used instead of “get” method for possible security reason.

“sss.html” secondary structure search interface file receives query information and send

it to “sss.jsp” file. The specific MySQL query command is generated by parsing and combining this information through operations with methods which are relevant to String class of JAVA language. The “sss.jsp” file checks which criteria among H (%), E (%), O (%), and sequence length were selected. Specified range is used for the generation of database query command. This file also lists the specified structure IDs of SCOP and PDB and uses this string of list for the construction of database search query. The method of result ordering is also fetched from html file and input to the jsp page.

The database server of MySQL is located in the local server computer host. The port number of the database service is 3306, which is the default one of MySQL application. Database name of phipsi was used for the deposit and retrieval of secondary structure information. Using DriverManager, Connection, and Statement classes of java.sql package, the query command was executed. The result of the query was managed by ResultSet clsss of java.sql package. Connection class object was constructed from DriverManager by calling getConnection() method. Statement class object was constructed from Connection class by calling createStatement() method. ResultSet class object was constructed from Statement class by calling executeQuery() method. Each result of concerned content is fetched by getString() method of ResultSet class using variable name as arguments. If SQLException occurs, try and catch statement handles this exception. Most of the error occurs when there is no element in the database which fits the query since the query command syntax is always correct for our right parsing and combining of information. Other possible case of error might include systematic error of database server application. Because this case would be quite rare, the error message which indicates that there is no such element is printed in the case of SQLException.

“ssc.html” secondary structure calculation interface file uploads query file and send the file information to “sss.jsp” file. The uploaded file is parsed from the request object of “ssc.jsp” page. ServeletFileUpload class of org.apache.commons.fileupload package, DiskFileItemFactory class of org.apache.commons.fileupload.disk package, and

ServletFileUpload class of org.apache.commons.fileupload.servlet package is used to fetch file upload. List and Iterator classes of java.util package was used to deal with parsing of uploaded files. File class of java.io package was used to contain uploaded file before writing to local hard disk. FileItem class object created from parseRequest() method of ServeletFileUpload class was used to write the file into hard disk of server computer. If there is an error during the file uploading and saving, the error is recorded and all further process is stopped.

After the file uploading and saving to local disk is performed, "ssc.jsp" page checks the uploaded file to see if there is any error. This error checking is performed using classes of FileErrorFind package of this web application which parse necessary information from PDF format file. Further process is stopped if any error occurs. Sequence of the protein is fetched from SeqWriter class of SecStrFind package of this application. TorsAngCalc class of TASA package utilizes AtomCoordRead class and AngleCalc class of the same package to generate phi/psi backbone torsion angle file. AtomCoordRead class parsed the coordinate, atom type and residue type information. This torsion angle file has angles in radians and has file extension of ".fs". The calculation of torsion angle of A-B-C-D exploited inner product of the normal vectors of the planes of the A-B-C and B-C-D. Secondary structure determination for each residue was conducted referring the backbone torsion angle data file. Secondary structure of amino acid residue is classified into helix if the backbone torsion angles belong to the range of (φ, ψ) = (-155°~-47°, -62°~-52°), (-104°~-47°, -52°~-37°), and (-117°~-104°, -52°~-37°). A residue is classified into extended secondary structure if the backbone torsion angles of phi and psi belongs to the range of (φ, ψ) = (-155°~-138°, 90°~155°), (-140°~-64°, 90°~180°), (-64°~-53°, 90°~100° and 110°~168°). Residues of which backbone torsion angles belong to other range were considered to belong to other secondary structure. Percentage of each secondary structure is calculated by calc() method of PercHEO class using the calculated secondary structures. The result string is displayed on the jsp page.

"tasa.html" protein pairwise comparison interface file either receives information of

PDB and SCOP database ID specification or uploads a pair of query files and send the information to “tasa_DB.jsp” file and “tasa_file.jsp” file each. The DB IDs are sent to the “tasa_DB.jsp” file. “tasa_DB.jsp” file uses getParameter() method of request jsp page object with variable name as argument to parse transported information from “tasa.html” file page. Names of necessary files are derived from parsed string and are used for fetching backbone torsion angle data. Using TASA class of TASA package of the application, the pairwise comparison is conducted. TASA class reports spent CPU time by exploiting ManagementFactory and ThreadMXBean class of java.lang.management package. TASA class implements doTASA class as calculation module. doTASA class calculates the logPr and RamRMSD value by shifting frames with non-gapped alignment. Among the frames that were shifted, the frame which yields the minimal logPr and RamRMSD values is selected. The result of TASA class is a string which contains the information of logPr, RamRMSD, CPU time (logPr), CPU time (RamRMSD), sequence length of the first protein, sequence length of the second protein, the search space size, and the minimal sequence length between both proteins.

“tasa_file.jsp” page is invoked when the query is a pair of protein structure file. The uploaded protein files are stored into the local disk by utilizing FileItem class of org.apache.commons.fileupload package, DiskFileItemFactory class of org.apache.commons.fileupload.disk packge, and ServletFileUpload class of org.apache.commons.filupload.servlet package. The file name of the saved file is the time of the save in millisecond format. The error is checked from this stored file by utilizing classes of FileErrorFind package of the application. Amino acid sequence is fetched using SeqWriter class of SecStrFind package. If no error is found, the pair-wise backbone torsion angle comparison is conducted using TASA class of TASA package. TASA class utilizes doTASA class as mentioned above. The detailed process of pair-wise alignment and similarity calculation is the same as for the “tasa_DB.jsp” page.

4.4 Discussion

In this chapter, the application of secondary structure information server based on protein backbone torsion angle was introduced. Secondary structure could be regarded as the building blocks of 3D tertiary structure. Among the secondary structures, local hydrogen bonded helix and extended structure is frequent. This web application provides search interface for deposited secondary structure of PDB and SCOP entries, secondary calculation utility for user's own structures, and pair-wise protein structure comparison utility.

The web server employed jsp infrastructure for easy accommodation of previously created JAVA based classes and packages to the developed web application. Simple query information forms were built with HTML format files. While jsp pages mainly conduct database searching, implemented JAVA packages conducts checking of file content error, calculating torsion angle, determines the secondary structures of each residues, and aligns and compares protein structures. The application tried to explain the details as much as possible for every change along the computational processes and any occurrence of exceptions.

Secondary structure search through query string of three types of structures using sequence alignment algorithms and through query string of backbone torsion angles using torsion angle based protein structure alignment algorithm might be useful improvements. Future protein structure researches based on the secondary structure might be more aided through the addition of homology search utility.

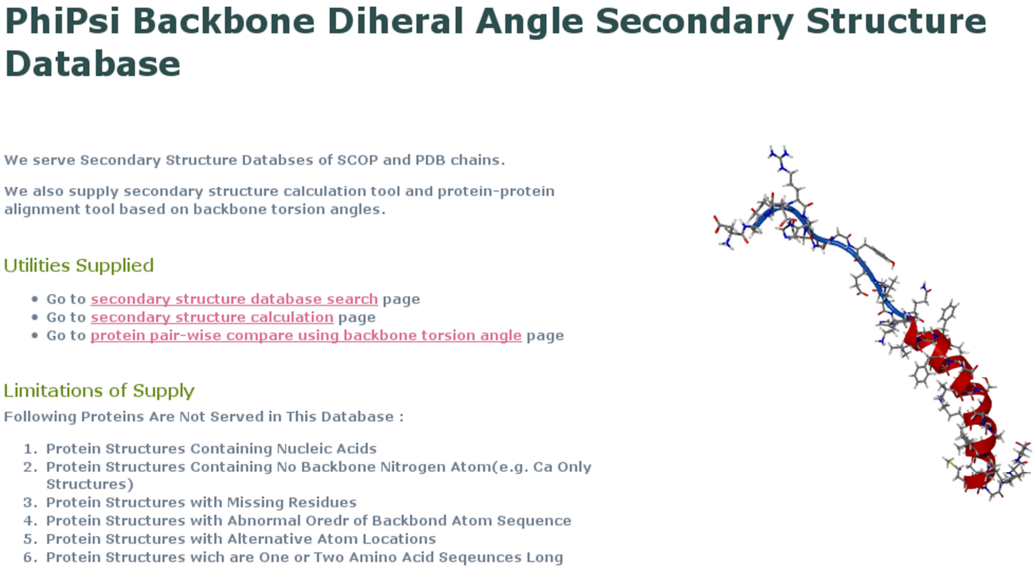

Figure 4-1. First Page of the Secondary Structure Database Web Application. Web application of secondary structure information repository serves three functions of secondary structure database search, secondary structure calculation, and protein pair-wise structure alignment. Limitations of the supplication of information are also clearly described. Currently, PDB and ASTRAL SCOP database information is possible to be retrieved.

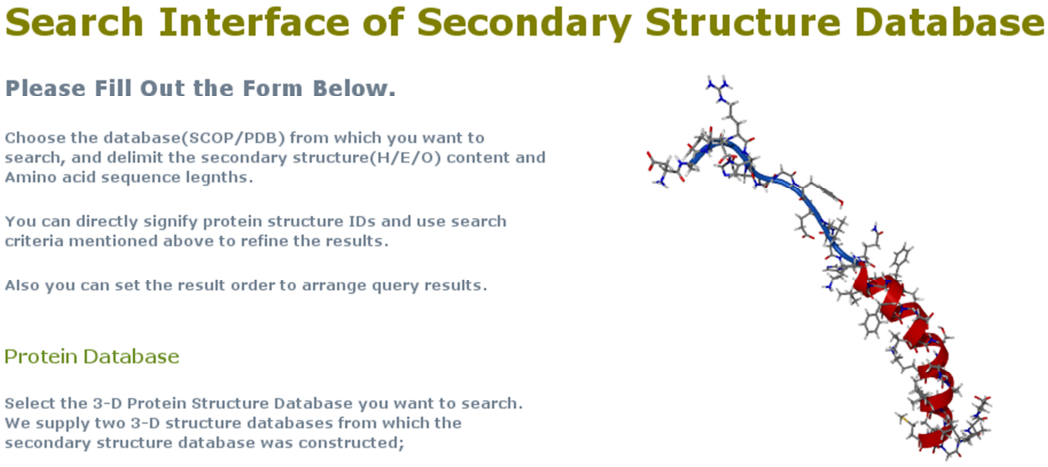

Figure 4-2. Secondary Structure Database Search Interface. The headear region of the secondary structure database search interface is presented. The directions urge users to choose searched database, secondary structure (H/E/O) content and peptide length. It also informs that one can delimit the searched space within the specified list of IDs of database entries. This also informs that user can set the oredering scheme of the results.

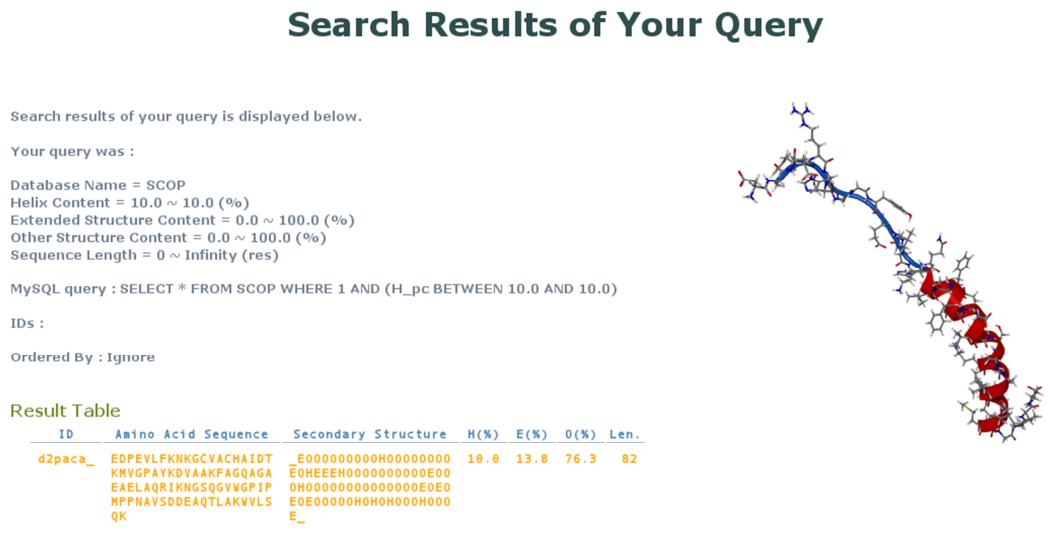

Figure 4-3. Secondary Structure Query Result. Sample query results are presented. The query was for the entries of SCOP database with helix content of 10.0%. MySQL query statement is presented below to the query summary information. The results in the result table section are asendingly ordered by the helix content.

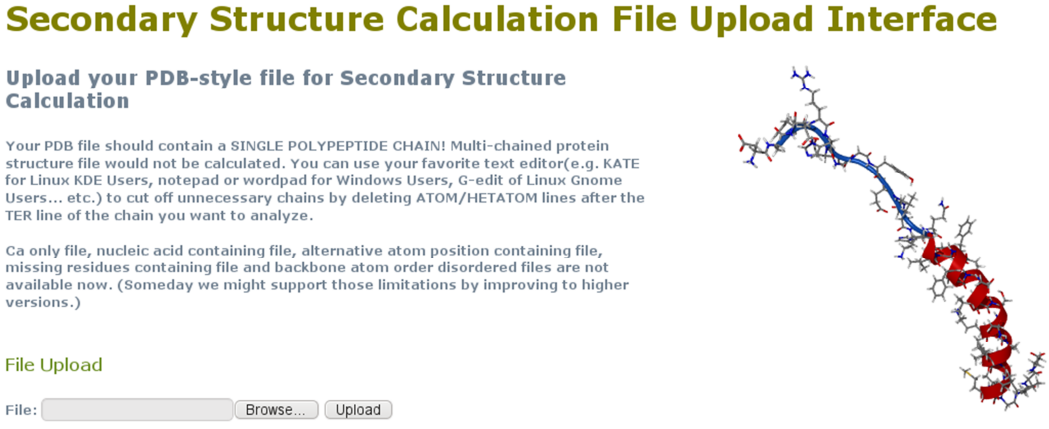

Figure 4-4. Secondary Structure Calculation Query Interface. Header and file upload dialog of secondary structure interface is presented. Calculation of only single chained PDB format file is available. Further limitations of availability is described. This is enabled by file input component of HTML format.

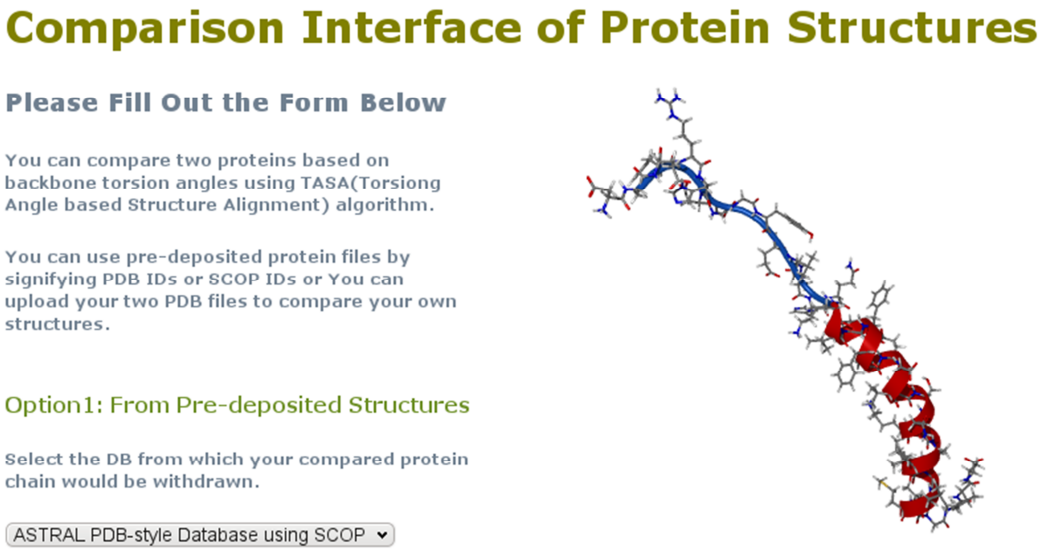

Figure 4-5. Pair-wise Protein Structure Comparison Interface. Header of protein pair-wise comparison interface is presented. The alignment algorithm named TASA (Torsion Angle based Structure Alignment) was used for the structural alignment and similarity measurement. Pre-deposited structures and uploaded structures can be compared.

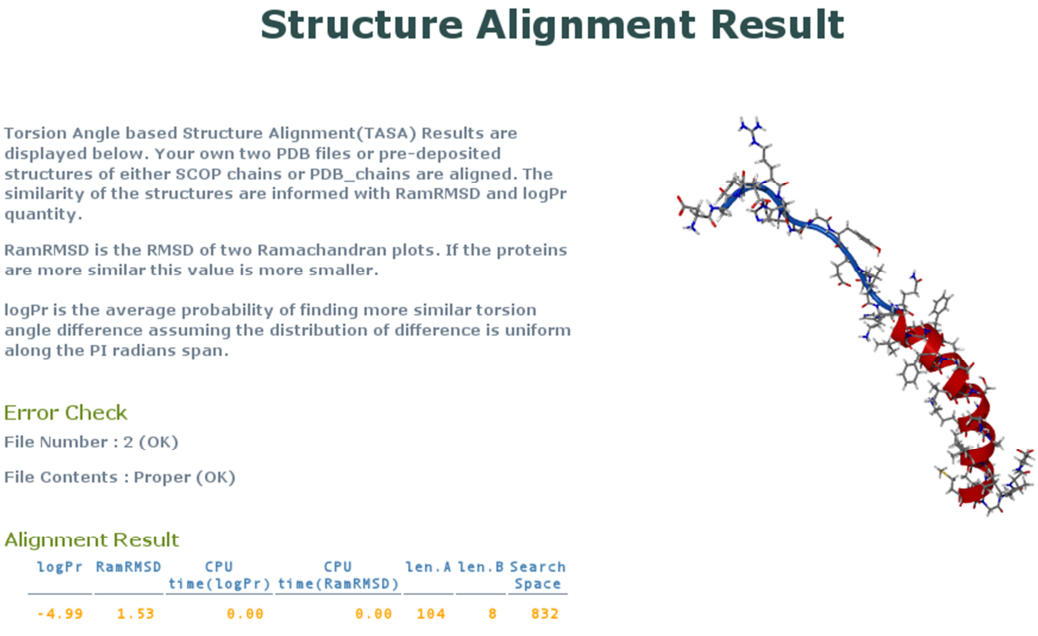

Figure 4-6. Protein Pair-wise Structure Comparison Result. The sample result page of protein pair-wise structure smiliarity comparison analysis is presented. The similiarity is informed through RamRMSD and logPr. RamRMSD ranges from 0 to 3.14 while logPr ranges from -16.0 to 0. -16.0 was used instead of negative infinity for the limitations of accuracy of dealt numerals. CPU time for the calculation of logPr and RamRMSD each, amino acid sequence length of each proteins and the magnitude of search space are also displayed in the result table.

CHAPTER V.

ProtTorter : A Protein Structure Modeler with Torsion Angle System

5.1 Introduction

Computational studies of protein structures have been carried out since the first attempt of deciphering the three dimensional structure of sperm-whale myoglobin from X-ray diffraction patterns in 1950s. (Kendrew et al., 1958) The utilization of computers was inevitable for the complicity of possible locations of atoms from the relational information. Graphical representation of three-dimensional protein structure is quite valuable for various structural analyses considering the complexity of three-dimensional information. Previously, especially when the computational representation of molecular structure was quite limited, structural representations usually exploited metal, wooden, or plastic balls and wires. This approach has, however, several limitations including the long time for the construction and conformational modification, limited methods for display, and difficulty of apprehension for its fixed physical dimensions.

Three-dimensional representation using computational application enabled fast and easy modeling of protein structure. One can easily color, shadow, tint, texture, and illuminate subject portion and zoom, rotate, shrunk, and expand the structure. (Baxevanis and Ouellette, 2005) Computational viewer applications supply better utility of apprehension by exploiting improved computational resource. Well known structural viewer applications include RasMol (Raster Molecule; Sayle and Milner-White, 1995), Chime (Chemical MIME; http://www.mdlchime.com), WebMol (Walther, 1997), Cn3D (Hogue, 1997), SwissPDB-Viewer (Kaplan and Littlejohn, 2001), Sirius (http://sirius.sdsc.edu) and PyMOL (http://www.pymol.org). RasMol is written in C language and very fast for

its innovative code design and fast ray-tracing algorithm. (Baxevanis and Ouellette, 2005) It is compatible with PostScript (MolScript; Kraulis, 1991) format. The command-line language of this application is regarded as a typical type for structure viewer applications.

Chime is written in C++ and numerous packages including ProteinExplorer (Martz, 2002; http://proteinexplorer.org), Sting Millenium (Hilga et al., 2004), and FirstGlance (http://bioinformatics.org/firstglance/fgij) were based from this application. (Baxevanis and Ouellette, 2005) WebMol is a java implementation of RasMol. (Baxevanis and Ouellette, 2005) It has no OS dependence for using JAVA language and supports broad options. Cn3D is a freeware and also a helper application for structure viewer in MMDB of NCBI's Entrez service. (Baxevanis and Ouellette, 2005) It is written in C++ with OpenGL (open graphic library). SwissPDB-Viewer (or DeepView) is a closed-source program which utilized OpenGL and supports very much options and it is even inappropriate for the first users. (Baxevanis and Ouellette, 2005) This also supports POV-Ray (Persistence of Vision-Ray Tracing) function. Sirius is a structure analysis and molecular modeling application which supports advanced users whose need exceeds simple displaying of structure of molecules. (http://en.wikipedia.org/wiki/Sirius_ visualization_software) It supports high quality 3D graphics and visualization of molecular dynamics trajectories from CHARMM and AMBER simulation output files. (http://en.wikipedia.org/ wiki/Sirius_visualization_software) PyMOL is an open-source, non-free binaries program which utilizes Python programming language. (http://en.wikipedia.org/wiki/PyMOL) It can produce high-quality 3D images of biological molecules and the authors argue that a third of publicized 3D protein structure images are from PyMOL. (http://en.wikipedia.org/wiki/PyMOL)

There are more than two dozens of free viewer applications which are stable, functional, and easy to use. (Baxevanis and Ouellette, 2005) The ultimate object of the viewer application is to convey scientific information in a lucid visual representation. Graphical representation of the three-dimensional structure of proteins is quite valuable

considering the complexity which is hard to clearly understand with three-dimensional numerals. The visual representation of the structure information during the folding simulation is further valuable regarding the fast apprehension of the structural and energetic information of each state. This is partly applied to the molecular dynamics study as the visualization of the simulation trajectories. However, typical previous algorithms including molecular dynamics and Monte Carlo method are not much eligible for the application of the retrieved visual information for the modifications during the simulation. This is partly due to the complexity of the system that represents the conformational space and to the lack of strict constraints that eliminates unnecessary possibilities. Torsion angle system with appropriate limiting path following the Levinthal's postulation is, on the contrary, quite adaptable for the structural modifications with the reference of graphical information.

Molecular conformation changes mainly through the torsional movements of atoms along the axis of covalent single bond while the movement that changes the length and angle of the covalent bond is quite rare. Thus, the torsional system that was introduced in chapter 3 could be regarded as to regenerate the natural movement of conformational changes. This system of the representation of the conformational space was partly validated by the successful application to the pair-wise structural alignment. This torsional system is succinct in the description of the structural information and is also very interpretative for human. This amenability of interpretations and fast operations enables concurrent monitoring and modifications of structures during the structural modeling processes. A modeler named ProtTorter (**Prot**ein **Torter**) was constructed using this representation system for the primary step for the applicability to protein engineering and massive structural genomics. This newly developed application is a protein three-dimensional structure viewer and modeler which is based on the backbone torsions. It is possible that many genuine simulation algorithms to be applied with this new modeler.

Protein folding might be processed sequentially from N-terminus considering the

sequential nature of amino acid addition to the growing peptide and the short time of folding compared to the synthesis of protein. There are positive results of the direction-dependent folding postulations including the one with modified ROSETTA algorithms. (Ellis et al., 2010) Though the structure generated from cotranslational folding might have to undergo further modification in the cellular environment to become more stable structure, the direction-dependent or cotranslational protein folding algorithm is quite interesting in the points of which offers very broad option of interference of manipulation compared to the previous algorithms such as molecular dynamics. The fitting of each torsion angle of backbone peptide bond could be separated from any other torsion which resides in the rear. The torsion angles in the front also might be conserved during the calculation of the specific torsion angle. Thus, conformations are determined by considering and modifying very finite number of torsions while folding through the chain of backbone torsions. This character enables bountiful human interaction during the process of folding with visual representation. In molecular dynamics, however, there are very few things that could be done by human for its self-propagating property of simulation and indivisibility of the movement of atoms from others.

In addition to finding real native structure from structural simulations, one would also like to interactively modify the structure of a protein model of a given amino acid sequence for certain purpose while observing potential energy and rotating atoms along torsion angles. This might be especially important in the case of protein engineering including refinements after local structural perturbations. Thus, ProtTorter was made to visualize three-dimensional conformation, calculate the potential energy, and supply the user interface for backbone torsion angle manipulation. This new application has two main functions of protein folding and structure viewing. One can select each torsion angle of each peptide bond and manually rotate it around the torsional axis. Space-fill and wire-frame model can be applied for viewing. Finding global minimum and local minima from calculated energy landscape of φ and ψ torsion angles enables quick jump among local minimal conformations. It is possible to model a protein structure by

sequentially adding new amino acid residues using this computational application. It is also easy to modify built model by rotating part of the conformation. One can also observe the change of potential energy while varying the conformations. Plausible candidate conformations could be derived fast and easily by utilizing our new application. In the following sections, we would describe the details of this structure viewer and modeler.

5.2 Methods and Material

5.2.1 Computational Framework

Computatioanl implementations were mainly performed through JAVA programming language. JAVA SDK (Standard Development Kit) was used for the calculations and graphical displays. Graphical implementations were mainly performed through JAVA's javax.swing packages. The GUI was composed with five components of structure viewer, result viewer, sequence viewer, option panel, and menu bar. Local computational algorithms coded with JAVA language performs PDB file import and export, structure zooming, rotating, and moving actions, wire-frame or space fill representations, and concurrent energy calculation during the structural modifications. Sub-algorithms of JAVA codes enables search and jumps between local minima. Graphical representation of the three dimensional structure, backbone torsional modification, potential energy calculation functions are supplied by codes with JAVA SDK packages.

5.2.2 Model Energy Calculation

Every conformation built on the modeling panel has its own potential energy. This energy can be calculated based on various force fields. ProtTorter uses modified AMBER02 force field (Cieplak et al, 2001). All the information of the AMBER02 force field is stored in the AtomicInfo class and PotentialEnergy class in a static context.

As all the bonds and angles of atoms are fixed in our modeling procedures, part of the

force field (bond stretching and angle bending term) was omitted. Since all the fixed position of our atoms are from the native positions, the contribution of bond stretch potential and bond angle bending potential would be minimal. The use of those bond stretching and bond angle bending term was originally for the three-dimensional molecular dynamics where atoms move freely along any directions within the coordinate space. In this application, however, the atoms do not move along bond axis and the bond angles are fixed either. Thus, the necessity of referring omitted portion of potential is quite low. Under this supposition, the folding stability could be assessed only by van der Waals, electrostatic and torsional potentials. The potential energy thus can be described as follows.

$$E_{total} = \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\varphi - \gamma)] + \sum_{i<j}^{van\ der\ Waals} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}\right] + \sum_{i<j}^{electrostatic} \frac{q_i q_j}{\varepsilon R_{ij}}$$

AtomicInfo class contains charge, AMBER force field atom index, and AMBER force field atom type and PotentialEnergy class contains electric constant and parameters of van der Waals potential, torsional parameters, and improper torsional parameters. Improper torsional and torsional parameters are used in the same formula while improper torsional parameter is for more accuracy by representing planarity criteria. PotentialEnergy class contains unused bond stretch and angle bending parameters also for possible modification of utilization of the force field.

All the parameters are indexed based on the 40 types of AMBER atoms. Lennard-Jones potential was used for van der Waals potential. Parameters of van der Waals potential include the equilibrium internuclear separation (van der Waals radius) and potential well depth for self-interactions among 40 types of atoms. The specific parameters of van der Waals radius and well depth between two different types of atoms are calculated as follows

$$r_{eqm,XX} = \frac{1}{2}(r_{eqm,XX} + r_{eqm,YY})$$

$$\varepsilon_{XY} = \sqrt{\varepsilon_{XX}\varepsilon_{YY}}$$

, where r_{eqm} is van der Waals radius and ε_{XY} is well depth. By these calculated well depth and radius the potential of van der Waals interaction is derived as follows.

$$V(r) = \frac{C_{12}}{r^{12}} - \frac{C_6}{r^6}$$

, where

$$C_{12} = \varepsilon_{XY} r_{eqm,XY}^{12}$$

$$C_6 = 2\varepsilon_{XY} r_{eqm,XY}^{6}$$

Whenever the potential is to be calculated, the information of AMBER atom type and coordinate is transported to the calcPotential() method of PotentialEnergy class. Parameters of van der Waals potential between different types of atoms are calculated referring arrays of 40 self-interaction parameters following the formulae above. The potential is calculated from these parameters and the distance from coordinate pairs as shown above. Calculated potential is cumulated to account all interactions.

Electric potential is calculated referring charge and coordinate information transported to calcPotential() method of PotentialEnergy class. Charge is referred according to the 40 AMBER atom types. Electric constant of 332.0522173 ($kcal^{-1} \cdot Å^{-1} \cdot mol$) was used with dielectric constant of 1.0. Torsional parameter and improper torsional parameters include reference energy, reference torsion angle, and circular period determinant. Four atom types of a torsion angle are used as indices of the parameter arrays in torsional potential. Improper torsional array indices are also types of atoms but in this case the sequence of atoms is intended to describe plane.

For possible future modifications, bond stretch and angle bending potential parameters were also contained in the PotentialEnergy class. Bond stretch parameter includes energy parameter and reference length of bond. Angle bending parameter includes energy parameter and reference angle. The elements of bonds, angles, torsions and improper torsions of a created peptide chain are added to the specific array list in the

CoordinateBuild class by static method findBondageInfo() of BondsAndParameters class. Atom indices which belong to each element are stored into each array. This information stored in array lists is used to graphically represent bonds in the WorkPanel class object and also for the calculation of potential energy after every change of backbone torsion angles.

5.3 Results

5.3.1 User Interface

ProtTorter exploits graphic user interface (GUI) which offers easy apprehension of each status of option settings and structure manipulations (Figure 5-1). The modeler has three main functions of graphical representation of three dimensional structure of proteins, conformational modification through backbone torsion adjustments, and potential energy calculation. Energy calculation enables concurrent calculation of the potential energy during the conformational modifications and search and selection of local energy minima. The whole user interface which supports these functions is constructed within a single frame. Several structure viewer system including VMD (Visual Molecular Dynamics; Humphrey et al., 1996) employs split multi-frame user environment. Single one was chosen, however, for the succinct and lucid representation of the application system. We chose “windows” user interface manager instead of java default “metal” user interface manager for aesthetic reasons. Frame environment was constructed utilizing JFrame class of javax.swing package.

The frame contains five main subcategories of menubar, option panel, structure viewer, computational result viewer, and sequence viewer. Main possible operations of the application includes PDB file import and export, zooming, rotating, and moving of the structure, wire-frame and space-fill model representation of the structure, and energy calculations. Menubar used JMenuBar class of javax.swing package and is located at the top of the frame. Menubar contains three menus; file menu, edit menu, and model menu.

Each menu was built from JMenu class of javax.swing package. File menu contains functions of file import, PDB file export, new modeling panel creation, and application termination. Each function is tenable through “Open PDB” menu item, “Save PDB” menu item, “New Chain” menu item, and “Exit” menu item. Each menu item can be selected by short key of “Ctrl+O”, “Ctrl+S”, “Ctrl+N”, and “Ctrl+X” each.

Edit menu contains options of structure viewer. One can set background color of structure viewer panel by “Set Background Color” menu item. This menu invokes JColorChooser class of javax.swing package. One can zoom and center the structure automatically to fit the size of the viewer panel by “Fit To The Panel” menu item. “View Model” menu item enables the selection of structure representation model between space-fill and wire-frame (Figure 5-2). “Rename Tab” menu item enables users to change the title of the tab of working panel. Model menu contains 20 menu items which add new amino acid among 20 possible ones for the selected new peptide chain. All menu items mentioned above was implemented with JMenuItem class of javax.swing package and relevant process of functions were called by implementing ActionListener and ActionEvent classes of java.awt.event package.

Option panel is located right below to the menubar. It contains buttons, sliders, spinners, labels and text areas that are either packed or not packed into the toolbar. Option panel contains functional elements for modulations of torsion angles of each peptide bond, calculation of energy landscape to find minimal energy conformations, and displaying current status of energy and torsion angles. Modifying conformation is only allowed for newly constructed models for the current lack of treatment for irregularities among various PDB files.

Peptide torsion angle modulation components are packed into toolbar. Toolbar was built from JToolbar class of javax.swing package. These components include numeral spinner for the selection of peptide bond between specific residues and three sliders for varying the angles of torsions along with labels signifying pertinence. Peptide number spinner

was constructed from JSpinner class of javax.swing package. Three sliders for φ, ψ, and peptide bond angle modulations were constructed from JSlider class of javax.swing package. Relevant functional processes were called by implementing ChangeListener and ChangeEvent classes of javax.swing.event package.

Energy landscape of φ and ψ torsion angles of specified peptide bond could be calculated by clicking “phi pot.” and “psi pot.” buttons. Buttons were built from JButton class of javax.swing package and relevant functional process was called by using ActionListener and ActionEvent classes of java.awt.event package. After the calculation of energy landscape the local minima are displayed in the relevant combo boxes with the order from the global minimum. User can set the structure into local energy conformation by selecting the items of combo boxes and selecting set buttons next to each box. Global minimal energy conformation can be set by selecting the local minima of the lowest rank that appears in each box and click “set” button. Combo boxes were built from JComboBox class of javax.swing package. Each button was built from JButton class of javax.swing package and relevant functional processes were called through ActionListener and ActionEvent classes of java.awt.event package.

Every change of the status of φ, ψ angles and the potential energy is displayed on the relevant text field components signified by labels. User can input desired value for φ and ψ angle into these text fields to change the values. Text fields were constructed using JTextField class of javax.swing package and labels were made using JLabel classes of javax.swing package. Angle value changing functionality was implemented with java.awt.event.ActionEvent class and java.awt.event.ActionListener class.

Structure viewer, result viewer, and sequence viewer region was built as tabbed panes. Structure viewer panel was built with JTabbedPane class and JPanel class of javax.swing package. When a new structure is opened or a new model space is created, a new tab is opened with the title of the structure file name. Every new panel on which the protein structure is shown is added to the tabbed pane. The status displaying components and

local minima displaying combo boxes of option panel is subsequently updated when the selection of the tab is changed. User can close the specific tab by right clicking on the tab label. Computational result viewer is built with JTabbedPane class and JTextArea class of javax.swing package. After every energy landscape calculation of specified peptide bond of modeling chain, the result text is updated with new information and displayed in the text area of tabbed pane. Sequence viewer is also built with JTabbedPane class and JTextArea class of javax.swing package and is updated after the addition of new amino acids.

The overall user interface of ProtTorter is not complicated and easy to apprehend. Though this interface does not offer rich functions for almost every possible occasion, it contains all the elementary and necessary functions for structure viewing and modeling, especially the one through direction-dependent or cotranslational backbone torsion angle folding. ProtTorter also fully utilizes the benefits of graphical user interface to the text user interface.

5.3.2 Protein Structure File Import

Structure viewer of ProtTorter supports PDB format. The import process is invoked when user selects “Open PDB” menu item on the File menu. JFileChooser class of javax.swing package was exploited to browse and specify the objective protein structure file. Dialog window of file browser and opener is displayed when this menu item is selected. Dialog window is shown by invoking showOpenDialog of JFileChooser class. Java file class object is generated by getSelectedFile() method of JFileChooser class. This fetched File class object is sent to the Coordinate class of the package of ProtTorter and used for the parsing of information.

5.3.3 Protein Structure File Export

The modeled peptide structure should be exported to discrete file for further possible modifications. One might use created structure for the generation of professional graphical representations or bioinformatical calculations. User can choose

"Save PDB" button to save created structure into file. This button invokes showSaveDialog method of JFileChooser class and fetches File class object of saving file. Atom serial number, atom name, residue number, residue name, Cartesian coordinates, occupancy, and thermal R-factor, atom type information was recorded into new saved file with PDB format of 80 columns (http://wwpdb.org). Occupancy and thermal R-factor value of 1.00 and 0.00 each was used. Thermal factor was set to 0 for the perfect clarity of the position of atoms of the model. The saved file also contains a simple header line which indicates the name of the structure.

5.3.4 Parsing and Initialization of Structure File

ProtTorter first parses the atom number information. It accounts only the first model of NMR ensemble file. Every occurrence of atoms of different name and residue to the previously counted ones is counted. The counted atom number was used for the generation of arrays for parsing of structural information. We parsed coordinates (x, y, and z), atom name, atom type, residue name, and residue number. We excluded the alternate atom location information for display and used only the first atom among the possible alternatives. After parsing the coordinates and other necessary information, the application reposes the origin of coordinates into the center of the atom coordinates and changes the atom coordinates relatively to this origin. Centering the origin is for the convenience of rotatory transformations. After the parsing and repose of the coordinate, the application determines the priority of the painting following the rule which paints the closest object last to overlap on the top of more distant ones. The priority of every atom was determined according to the distance along the z-axis by comparing the z coordinates; smaller z coordinate was regarded as closer. All the parsed information is saved in the object of Coordinate class. This information containing object is sent to the ProtTorter's WorkPanel class for three-dimensional structural representation. WorkPanel class object draws the atoms on the screen while rotating the structure according to the mouse events using various methods of Coordinate class for the calculation of coordinates.

5.3.5 Structural Representation

Object of WorkPanel is created after the file of structure is imported and parsed by Coordinate class object. Initialization of the WorkPanel object includes fetching structural information from object of Coordinate class. Bondages among atoms are determined using this information and information previously stored in BondsAndParameters class. Bond information of each amino acid residue and inter residues is designated in the BondsAndParameters class. Appropriate bonds are assigned and each pair of bonds is input into the ArrayList class object. ArrayList class object can be declared without prior knowledge of the array size and was used for the variability of bond number by residue sequence.

Drawing process is different between space-fill model and wire-frame model. Default drawing model is wire-frame. The drawing priority of each bond is determined by the priority of the first element of a pair in the wire-frame model. Each half of the line of bonds was painted with colors specifying each atom following the drawing priority. Black was used for carbon atoms, blue was used for nitrogen, red was used for oxygen, yellow was used for sulfur, and green was used for hydrogen. Antialising method was applied to all renderings of graphics using setRenderingHint() method of graphics2D class.

Drawing space-fill model was quite simple. Each atom was represented with colored spheres with varying radius. Overlap pattern was determined according to the drawing priority calculated by Coordinate class object. The radius of each sphere was determined by covalent radius measured by Slater. (Slater, 1964) Covalent radius of hydrogen, oxygen, nitrogen, carbon and sulfur was 0.25pm, 0.60pm, 0.65pm, 0.70pm, and 1.00pm, respectively. Factor of 3.0 was multiplied to these radii to give distinguishable dimensional difference. While determining the coordinate of painting on the WorkPanel in both model, appropriate zoom rate, and transformation constants of both X- and Y-axis were included for calculation. For the fast speed of structural representation with space-fill model, we used previously drawn colored sphere pictures with radial gradient

of brightness instead of drawing and calculating the gradients during the viewer panel representations and modifications. Five spheres were used each for carbon, oxygen, sulfur, nitrogen and hydrogen atoms. Each drawn sphere was saved into .png format and was fetched using getResource() method of ClassLoader class, getClassLoader() method of Class class, and getClass() method of Object class. Fetched image file was imported into object of Image class of java.awt package.

5.3.6 Modifying Graphical Representation of Structure

User can rotate, zoom and move the structure representation by invoking various mouse events. If user rotate mouse wheel upward, zoom rate is increased by heuristic factor of 0.5 for every rotation. The number of rotation is counted by using getWheelRotation() method of MouseWheelEvent class of java.awt.event package. MouseWheelListener class and MouseWheelEvent class of java.awt.event package were used to listen to the changes of mouse wheel status.

User can move the structure either horizontally or vertically by dragging the mouse with right button. When the right button of the mouse is first pressed, the 2D coordinate of the point of press is recorded and when the button is released after the dragging, the final 2D coordinate of point of release is recorded. The amount of planar displacement is calculated relative to the mouse point displacement. MouseAdapter class of java.awt.event package and MouseEvent class of java.awt.event package were used to listen to relevant mouse events. Method isMetaDown() of MouseEvent class was used to discern the right button event. Methods of getX() and getY() of MouseEvent class were used to fetch the mouse point coordinate information. User also can rotate the structure either horizontally or vertically by dragging the mouse with left button. When the left button of the mouse is first pressed, the 2D coordinate of the point of press is recorded and when the button is released after the dragging, the final 2D coordinate of point of release is recorded. The amount of rotatory displacement is calculated relative to the mouse point displacement. New rotated coordinate of the structure is calculated by calling methods of Coordinate class of the application with the arguments from mouse

dragging information. Simple rotational transformation matrix was used for the calculation of new coordinates. MouseAdapter class and MouseEvent class of java.awt.event package were used to listen to relevant mouse events. Method isMetaDown() and isAltDown() of MouseEvent class was used to discern the left button event. Methods of getX() and getY() of MouseEvent class were used to fetch the mouse point coordinate information.

One can set various options to the structure viewer panels including selection of background color and model of representation, and centering and fitting of the structure to the panel. These options can be achieved by menu items from Edit menu. “Set Background Color” menu item invokes static object of JColorChooser class of javax.swing package when selected. Color selection window appears by invoking showDialog() method and color chosen in the dialog window is imported into Color class object of java.awt package. This imported color is used to set the background of currently selected structure viewer tab by invoking setBackground() method of JPanel class of javax.swing package. “Fit To Panel” menu item reposes and zooms the structure to center and fit the structure of the currently selected tab to the panel. This sets the instance field of WorkPanel class (Boolean instances named FIT_OR_NOT and CENTERING) which signifies centering and fitting option and repaints the structure when invoked by selection. “View Mode” menu of Edit menu contains two menu items signifying space-fill model and wire-frame model each. Each menu item of “Space-fill” and “Wire-frame” sets the model of WorkPanel object of currently selected structure viewer tab using setDisplayMethod() method of WorkPanel class.

Overall, ProtTorter offers viewing of PDB format structure with wire-frame and space-fill models. Zooming, rotating, and displacing both horizontally and vertically is possible. Automatic fitting and centering to the screen and changing of background color are also possible. This module of ProtTorter performs all the necessary functions for viewing a PDB protein structure.

5.3.7 Protein Model Building

Users can build and fold their own peptide chain using ProtTorter. New WorkPanel for model building is opened and added to the structure viewer tab when the "New Chain" menu item of File menu is selected. CoordinateBuild class is invoked and an object is generated after the action is performed by "New Chain" menu item selection. This generated object is transported to the WorkPanel and used for the generation of a new work field. This CoordinateBuild class contains the same information with Coordinate class (i.e. coordinates, atom name, atom type, residue name, residue number, and priority for drawing) with additional manipulation methods which are used for modeling.

The underlying mechanism is quite different from simple parsing of structure data file though the containing information is the same. All the coordinates of newly added amino acid is mathematically transformed from template coordinates to fit into correct orientation. The orientation is determined as the orientation of tailing two atoms of previous amino acids. All the template coordinates of 20 amino acids is stored in AminoAcids class and fetched in static context. AminoAcids class has atom name and atom type information additionally to the coordinates. Template coordinates were obtained from NMR structures which generally contains hydrogen atom coordinates.

User can add new amino acid by selecting menu items of "Add Amino Acid" menu of Model menu. addAminoAcid() method of CoordinateBuild class is called by the action performed to these menu items. The method addAminoAcid() inserts atom coordinate, atom name, atom type, residue name, and residue number information of specified type of amino acid into the relevant array list by copying them from static instance arrays of AminoAcids class. Information other than the coordinates of the tailing two atoms of nitrogen and carbon is not inserted into the array list. Only coordinate information of the tailing atoms are inserted. Each template coordinates of 20 amino acids has two tailing coordinates of nitrogen and alpha-carbon which were from those of the atoms of the following amino acids in a real peptide chain. The coordinates of new amino acid is

rotationally transformed to fit the orientation of the N-Ca vector of newly added residue to the vector of the tailing two atoms. After the orientation of the newly added amino acid is determined, the amino acid is linearly transformed to be connected with the previous amino acid. Coordinates of the atoms of amino acids including the coordinates of the tailing two atoms are inserted into specific array list instance of CoordinateBuild class object.

Concept of spherical coordinate is applied when orienting new amino acid. Inclination angle and azimuth angle of vectors of both tailing N-Ca and new N-Ca is calculated with nitrogen atoms as origins. Then, the new amino acid is rotationally transformed with nitrogen atom coordinate as origin to make its inclination and azimuth angle to be zero. The zenith direction was set to the direction of the x-axis of Cartesian coordinate. After this transformation the new amino acid is rotationally transformed into the angles of the previous tailing vectors. During the addition of amino acids, all the backbone torsion angles are set to 180º by setPhi(), setPsi(), and setPeptBondTorsion() method. This torsion angle setting follows similar transformations as the addition of amino acids. For convenient rotational transform along the axis of zenith direction (x-axis), the inclination and azimuth angles of N-Ca, Ca-C, and C-N vectors for each of the φ, ψ and peptide bond torsion angle are calculated. After the inclination and azimuth angle is set to zero by rotational transform, torsion angle could be adjusted by simple rotational transform along the axis. All atomic coordinates which are subject to the variation of the torsion angles are rotated during the transformation. Hydrogen atom of backbone nitrogen was set to be planar with other peptide bond atoms with similar transformations mentioned above by the method setPeptTorsNH().

After the addition of amino acid template coordinates and atomic information, modification of amino-terminal atoms and carboxy-terminal atoms are made. Amino-terminal modification is done by calling setAminoTerminal() method of CoordinateBuild class object. This method sets the amino-terminal information of specified CoordinateBuild class object. This method first finds the atom coordinate of

hydrogen atom of backbone nitrogen of the first amino acid residue. Coordinate of delta-carbon atom is used in the case of proline. This coordinate is rotated twice while copying those two coordinates into two additional N-terminal hydrogens. Spherical coordinate was exploited during the rotations with calculated inclination and azimuth angles. Newly generated coordinate information is stored into the relevant array list while atom name, atom type, residue name, and residue number information are also stored in each specific array list.

Carboxy-terminal modification includes both the removal of one of the backbone oxygen atom of previous terminal and addition of oxygen atom to the newly added terminal amino acid. ProtTorter calculates the terminal oxygen coordinate referring the coordinate of the tailing nitrogen atom. The terminal atomic and residual information except the coordinates are deleted from relevant array lists during the removal of the terminal oxygen. The remaining coordinate of the tailing atoms are used for the guidance for the orientation of the addition of new terminal residue. After the addition of the terminal amino acid, the terminal backbone oxygen atom information is designated by inserting the atomic information with the index of the tailing nitrogen atom. This insertion is simply conducted by adding information of oxygen (i.e. atom name, atom type, residue name, and residue number) to the relevant array lists with the index of the tailing nitrogen in the coordinate array list. The other tailing atom, alpha-carbon, is not displayed in the structure viewer.

5.3.8 Model Modification

User can modify each backbone torsion angles with sliders below to the menu bar. Each label which is left to the sliders signifies which sliders are for the modification of each of the torsion angles of φ, ψ and peptide bond torsion angles. The major tick spacing of slider for the modification of peptide bond torsion angle was set to 180º for the planarity from partial double bond character. We allowed 0º torsion for some possible cases of peptide bond of proline. Each of the three sliders has values from -180º to +180º. User can select the number of peptide bond by changing the value of the

spinner. Every time the state of spinner is changed, the event listener added to the spinner is automatically called to do processes including changing the instance field of peptide bond number information (integer instance field of peptBondNum) of MainFrame class and resets the values of torsion angle sliders with newly calculated torsion angles from selected peptide bond.

When the slider handles are moved, the change of the state is observed by the event listener which was added to each slider object. This listener invokes series of methods including rotational transform, energy calculation, and repainting of the drawing panel. Rotational transform is conducted by invoking relevant methods (e.g. setPsi(), setPhi(), and setPeptBondTorsion()). Each method calculates current angle value and subtracts the argument angle value from this calculated value. Then, the method rotates relevant atoms by the value of this subtraction after the preliminary transformation according to spherical coordinates. After these torsion angle settings, the newly calculated potential energy is displayed on the text field on the right of the second row of the option panel. One can freely construct any kind of protein structure from peptide chains by setting backbone torsion angles and checking the potential energy of each conformation. Every change of current selection of model is reflected on status displaying components. Torsion angle is newly calculated and reflected to the sliders and text filed when value of peptide bond spinner and selection of tab is changed. Sequence viewer tab which displays the amino acid sequence of the peptide chain of currently selected tab also changes when different modeling tab is selected.

5.3.9 Local Energy Minima Calculation and Cotranslational Folding

The folding of a protein might be sequential from N-terminus regarding the nature of protein synthesis where amino acids are sequentially added from the N-terminus. This is supported by the results of several direction-dependent or cotranslational folding researches (Ellis et al., 2010, Srivastava, 2011, and Saunders, 2011). As proteins are synthesized in the ribosome, new amino acid is selected through the codon-anticodon interaction between tRNA and mRNA and is attached to the

terminal carboxyl group of preexisting polypeptide. Each newly added amino acid adopts a conformational change following the changes along the path in the torsional space. During the torsional conformational changes, the torsion angle would settle into the angle of the one of local energy minima. Among possible local minima, the global minima would be the most plausible one for the set conformation. For the case of Endoplasmic Reticular (ER) proteins, the folding would occur following the intrusion of the polypeptide into the ER lumen through the membrane channel. In the case of the actions of chaperone proteins, the sequential folding would follow the order of the release of chaperone molecules which allows free movement of each residues and finding of the energy minima. ProtTorter offers a function of calculation of potential energy and finding local minima for each of φ and ψ angle while sequentially adding amino acids from N-terminus with the force field described above.

User can select a specific peptide bond by changing the value of the spinner of upper left of the option panel. After the selection of peptide bond, user can perform local minimum search of φ and ψ angle by clicking “phi pot.” and “psi pot.” button each. When the listener of this button is invoked, potential of φ or ψ angle of the selected peptide bond is calculated for every 1º along the whole rotation of 360º using for iteration loop. The calculation of φ and ψ angle is separated and the torsion angle is grained into 360 elements representing the range from -180º to 180º during the calculation. Usually φ angle energy calculation follows the calculation of ψ angle energy landscape of the same peptide bond following the concept of cotranslational or directional protein folding. After the potential energy calculation of 360 cases of torsion angle grains, local minimum energy is scrutinized. The width of 2º is used for the searching of local minima well on the potential graph. The local minima are searched by finding point where the energy of flanking points within the well-width is larger than the point. Angle and energy values of local minima of φ and ψ angle are stored into each array list. These array lists are prepared and conserved for every residue of a peptide, and for every peptide of a run application. Energy calculation of a torsion angle depends on the

circumstances determined by torsions nearer to the N-terminus. The indices of global minima among local minima for each residue of a peptide are stored into array list. This array list are prepared and conserved for every peptide which is modeled in WorkPanel tabs.

After the calculation of potential energy and the finding of local minima and global minimum, the combo boxes of local minima angles are set to the ones of global minimum. The structure also automatically changes to the global minimum energy conformation after the calculation and finding of energy minima. One can change the structure to the local minimum energy conformation by modifying combo box values. After the selection of wanted local minimum torsion angles, one can click “set” button to change the conformation. Also, user can always go back to the global minimum energy conformation by selecting the lowest minimum indicated by the ordering numeral which resides before the angle value.

Whenever the conformation is changed, the φ and ψ torsion angles of the peptide bond selected by the spinner is displayed on the text status field on the lower right of the option panel. The potential energy of every conformation is also displayed when it is changed. The results of torsion angle energy calculation are recorded into the single String class object and displayed in the text area component of the result viewer tab. Result viewer tab contains the tabs of phi and psi angle information for a single peptide bond and a tab for summary of information for every residues. This result text contains angle and energy of the local minima derived from the torsional energy space of a peptide bond. It also reports the order of local minima from the lowest minimum. This information is displayed according to each peptide bond and is updated for every calculation of energy landscape of any peptide bond. The long results can be searched by scroll bar. Scroll bar option was implemented with JScrollPane class of javax.swing package.

5.3.10 Typical Process of Generating Cotranslational Peptide Model

The main function of ProtTorter is to generate possible three-dimensional conformation of a protein of a given sequence (Figure 5-3). To make new structural model, one first have to create new work place by clicking “New Chain” menu item of “File” menu. User can also simply type “Ctrl+N” to do this work. After creating new work panel, user can add any amino acid out of 20 possible ones by selecting menu items of “Add Amino Acid” menu of “Model” menu. After the addition of new amino acid, user can select the peptide bond, energy of which torsion angle is to be calculated by changing the value of peptide bond number spinner labeled as “peptide(th):”. By pressing “phi pot.” and “psi pot.” the energy landscape of φ and ψ torsion angle each is calculated. From this energy landscape, local energy minima is found and displayed in the relevant combo boxes. One can browse the listed local minima and move to the desired minimum conformation by selecting specific combo box item and pushing “set” button.

This general scheme of addition of amino acids and finding the energy minimum can be applied in any desired order of algorithms. It is notable that cotranslational folding algorithm is quite convenient for this scheme of modeling. Non-cotranslational folding algorithm with no constriction folding path is more complicated and might need additional autonomous processes to this manual scheme.

5.4 Discussion

This new application offers functionalities of simple structure viewing and backbone torsional modeling. These functions were made possible by the backbone torsion angle modification and energy calculation functions. This application exploited graphical user interface for displaying, creating and modifying protein structures. The results of energy calculation and local minimal conformations are also accessible by graphical interface. Every local minimal conformation could be easily converted into

another by selecting relevant values in the combo boxes and clicking “set” button. The conformation can be represented by either space-fill model or wire-frame model.

User can add new amino acid residues to grow peptide chain. This chain can be folded into compact shape by modifying torsion angles of the backbone. One might refer calculated potentials displayed on the status text field on the right of the option panel during the modification of backbone torsion angle. Automatic calculation of energy minima is also available. With these functionalities one can conduct modeling of protein structure very fast and easily. User can fold the protein by distorting torsion angles one by one following various paths including cotranslational order from the N-terminus.

Molecular conformation changes mainly through the torsional movement of atoms along the axis of covalent single bond. The movement of atoms which changes the length and angle of covalent bonds are very rare. Thus, the torsional system of space representation might be regarded as the regeneration of natural processes. This representation of structural space of conformations enables very succinct descriptions of states. These manners of descriptions lead to the much of amenability in dealing with the structure information for numerous assays and analyses. The validity of this representation system has been strongly proved by the application to the structure pair-wise comparison in chapter 3. As mentioned in chapter 3, this torsional system enables sophisticated alignment algorithms. Torsional system also enables very easy implementation of structural simulation by its succinct representations. Furthermore, torsional system enables structure modifications which are very easy for human interpretations. Using this system for implementation, very lucid protein structure modeler was possible to be constructed.

The computational application of ProtTorter introduced in this study is possible to be utilized as for the protein folding simulation application with the option of concurrent viewing of states. There might be options for the concurrent structure modification during the process of simulations. The concurrent monitoring and modification of the

structure offers broad options of structural modeling which were not possible to be achieved in the past methods. One can drastically change the conformations during the simulation to overcome the depth of the well of the local energy minimum or specifically draw the direction of the simulation to the conformations which are similar to the desired conformation. This new type of amenability might offer drastically new approaches to the protein folding study.

The new modeler application with concurrent monitoring and modification would result very fast and user's object oriented modeling of conformations. From preliminary amino acid sequences of interest, the final energy minima conformations would result in rather shorter spent time. Also, the refinements based on the local modifications of the structures are much more tenable with this new modeler. Drug targeting using massive structural modeling from amino acid sequences might be more plausible with the protein modeler with these new possibilities. Protein engineering which designs new proteins of predetermined desired functions might become more feasible by utilizing the fast speed of simulations.

Further goal of ProtTorter is to implement the functionality of structural modification by distortion of torsion angles from imported PDB files in addition to newly built structures. Also, this application needs graphical result presentation including 2D heat-map of φ and ψ angles or other line plots. Also, the structural presentation with ribbon model is necessary to easily appreciate secondary structure characteristics. The most important future improvements might be the torsional variation of side chain conformation. Now, ProtTorter only considers a single side chain conformation. So, there is no alternate possibility of side chain conformation if a steric hindrance occurs. However, there might be some other side chain conformation with different torsion angles that does not cause steric hindrance. By allowing the torsional variation along the torsion angles of side chain and searching all the possibilities, the more correct conformation might be possible to be obtained. Unfortunately, for the lack of side chain conformation variability, only helical structures from rather smaller size peptide chain are correctly

deducible. The research on these helical structures using ProtTorter will be referred in the following chapter.

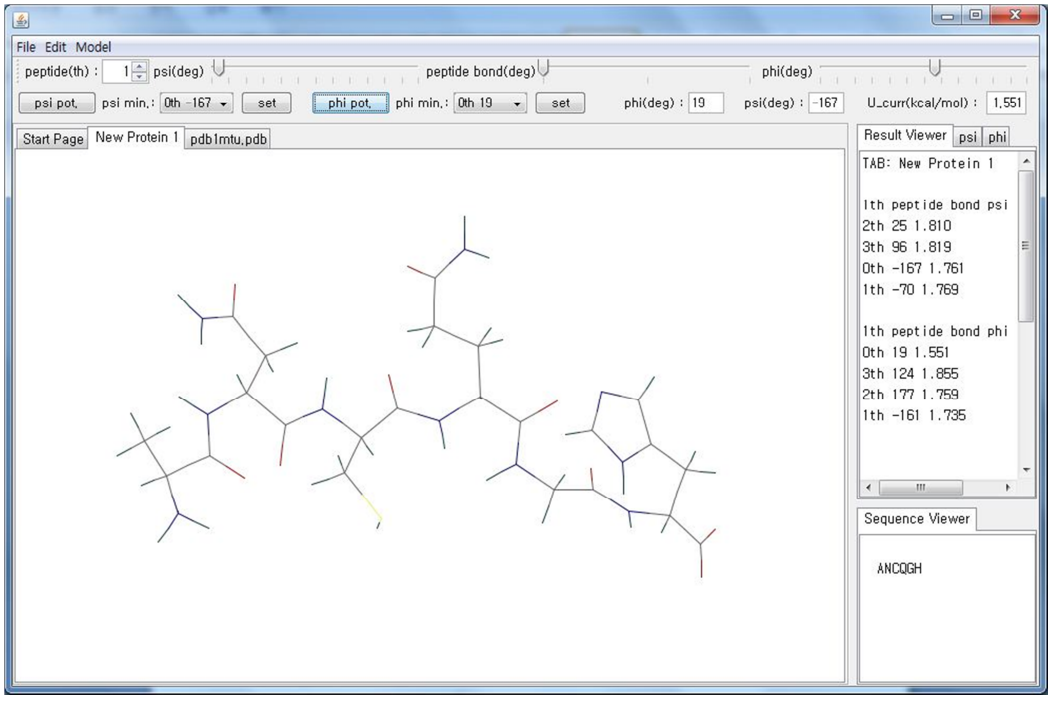

Figure 5-1. Graphical User Interface of ProtTorter. The graphical user interface of ProtTorter is presented. Menu bar is on the top of the application and option panel with torsion angle modification function and status displaying components is located below. Tabbed structure viewer panel, result viewer, and sequence viewer is also shown. Long result and sequence information is searched through scroll bars. Sample structure and energy minima of a peptide with sequence of "ANCQGH" are shown.

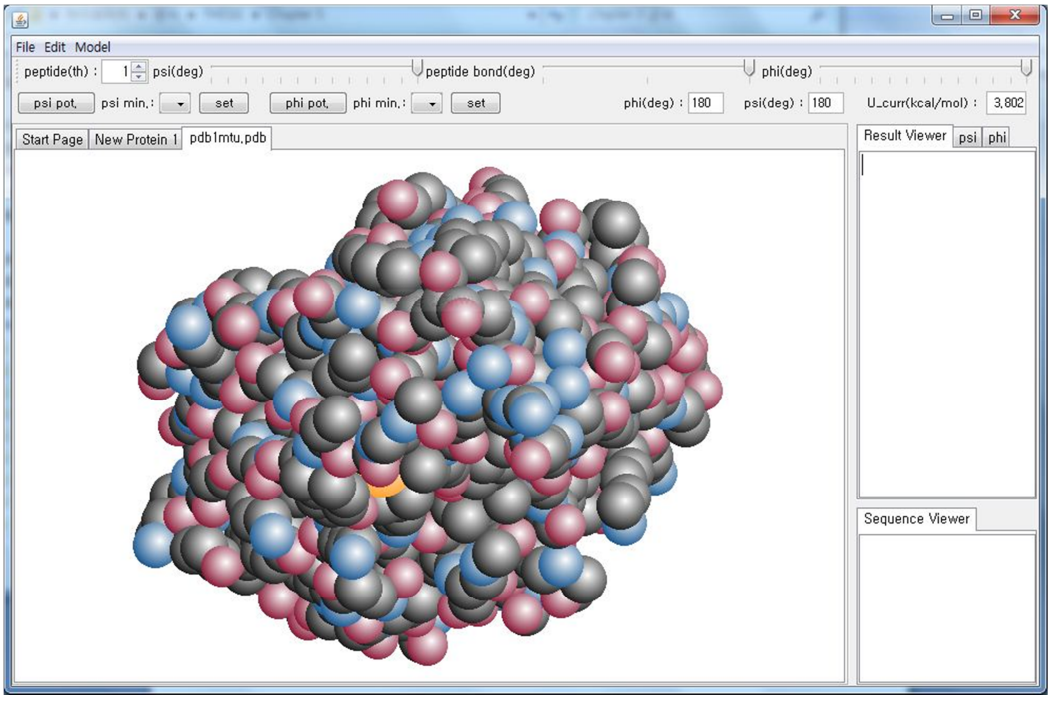

Figure 5-2. Space Fill Model of the Structure of PDB 1mtu. The space fill model of the protein structure of PDB entry 1mtu is shown. The structure was fetched by invoking "Open PDB" menu item of the "File" menu. Carbon (black), nitrogen (blue), oxygen (red) and sulfulr (yellow) atoms are shown in spheres. By selecting either the "Space Fill" or the "Wire Frame" menu item of the "View Mode" menu of the "Edit" Menu, user can change the representing model of the structure.

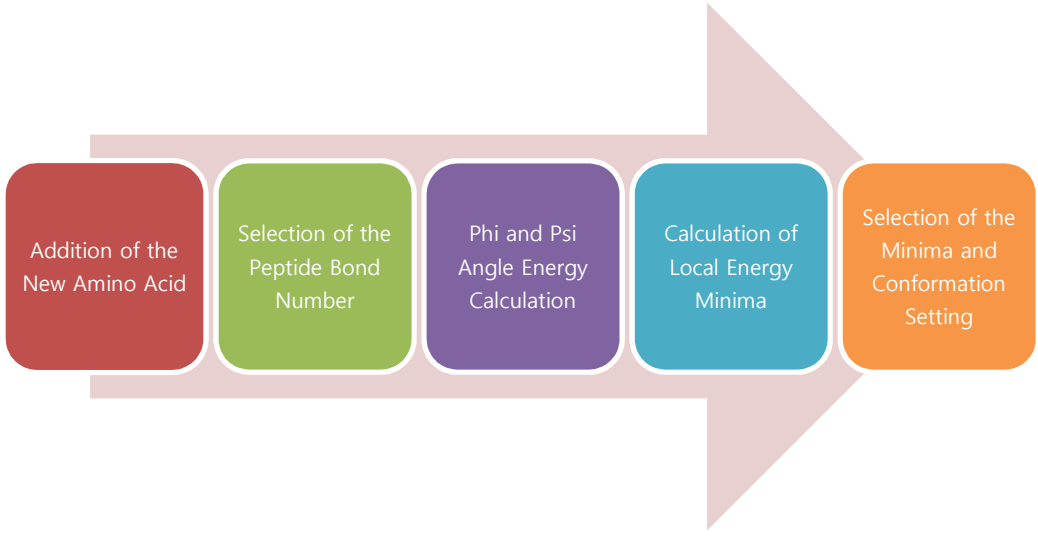

Figure 5-3. Typical Process of the Generation of Cotranslational Conformations. A typical process during the conformational generation according to the cotranslational folding concept is displayed. After the addition of a new amino acid by selecting "Add Amino Acid" menu item of the "Model" menu, user select the peptide bond number of the calculated part using the spinner on the option panel. Then user calculates the energy landscape by selecting "psi pot." or "phi pot." button. After the calculation of the energy the energy minima are displayed in each combo box. User finally can select the wanted energy minima and change the conformation by clicking "set" button next to the combo box.

CHAPTER VI.

Protein Folding of Cotranslational Initial Structure along Torsional Levinthal Path with ProtTorter

6.1 Introduction

The primary problem of protein prediction methods is the overwhelming quantity of the number of possible conformations that nascent protein can adopt. The postulation that protein folding procedure would not try every possible conformation is called as Levinthal paradox. (Levinthal, 1968) There was an argument that there might be a directing pathway to the fold of native state in the protein folding. (Karplus, 1997) Cotranslational folding is one of such possible paths considering the circumstance of protein synthesis from mRNA transcript.

Amino acids are added one by one to the end of the preexisting polypeptide through the formation of peptide bond. After each addition of new amino acids, the newly added amino acids would change their position in the space following torsional movement. The preexisting conformation also would change according to the newly added amino acid. The movement should be described as torsional since the movement along the length of covalent bond or the change of covalent bond angle is quite few in comparison with the torsions along the single covalent bond axis. The torsions result significant change of atoms which are connected through multiple steps of covalent bonds. In other words, each newly added amino acid and preexisting amino acids adopt torsional conformational change after the addition to adopt local conformation. This local conformation could be the global minimal energy conformation along the torsional energy space of all torsion angles. The conformation of the whole protein during the

polypeptide is thus determined by the sequential change of backbone torsion angles of each newly added amino acid residues if one considers the effect of the newly added one is small to the previous local conformation formed without the new one.

The newly synthesized polypeptide extrudes the exit tunnel of the ribosome to be relieved and folded within the solvent environment of the cytosol. In the case of cotranslational translocation of the endoplasmic reticulum proteins, the newly synthesized polypeptide chain penetrates the translocator of the rough endoplasimic reticulum after the guidance of the ribosome to the translocator on the endoplasmic reticulum through the action of signal recognition particle (SRP) and signal recognition particle receptor (SRP receptor). Though local folding would occur during these penetrations of the tunnels, the possible conformations are strictly limited according to the constricted dimensions of the vacuous space. It is known that proteins adopt α-helices during the penetration of the exit tunnel of the ribosome. (Jenni and Ban, 2003; Voss et al., 2006) Thus, it is possible that the significant sequential cotranslational folding would occur after the penetration beyond the membrane translocator. The sequence of local structural folding would follow the order of the release of chaperones in the case of the proteins that chaperone action is strongly important including the cases of cytosolic proteins being imported into membranous organelle lumen. Amino acids specific to the site of the release of the chaperone proteins are, then, free to rotate and adopt energy minimal structure. Though there might be further modifications after these cotranslational folding, this scheme of folding might have strong significance to the simulation studies if the cotranslational folding path has strong impact on the formation of the final native conformation. This limited path of folding strongly reduces the possible candidate conformations by regeneration of the real biological process which was previously regarded as unimportant.

There is a strong discrepancy between the time needed for new amino acid addition to extending polypeptide chain and for synthesized peptide chain fragment to fold into temporary stable structure. Typically, prokaryotes add 15-20 amino acids per second.

(Zhang and Ignatova, 2011) Translation of a single *E. coli* codon takes about 0.05s. (Pedersen, 1984; Krüger et al., 1998; Sørensen et al., 1991; Varenne et al., 1984) Eukaryotes add 1-5 amino acids per second. (Zhang and Ignatova, 2011) In contrast, proteins fold extremely quickly. (Schonbrun and Dill, 2003; Kubelka, 2004) Small proteins fold within microsecond time scale. (Kubelka, 2004) Helical and sheet local structures fold within the lower milisecond time scale. (Roder et al., 1988; Briggs et al., 1992; Lu and Dahlquist, 1992)

This suggests the possibility of the stable structure formation before the complete synthesis of the whole protein. In fact, there is an experimental study that shows the ample time for protein to fold before the completion of synthesis. (Baldwin, 1999) This preliminary structure formation strongly reduces the search space of the later folding for the independence of each N-terminal fragment structure to the folding of rear amino acids. This reduction of conformational space might be the directional pathway following the postulation of Levinthal paradox.

There are evidences of proteins for the attainment of biological structures while they are still in the process of synthesis on the ribosome. (Nicola et al., 1999; Kolb et al., 2000; Cabrita et al., 2010; O'Brien et al., 2011) Also, there is a report that in vitro refolding is slower than in vivo folding (Seckler et al., 1989; Fedorov and Baldwin, 1999) which partly suggests that reduction of conformational space by the cotranslational folding process during the in vivo synthesis strongly affects the formation of the native states. There are other experimental indications which show the possibility of cotranslational folding. β-galactosidase which is bound to ribosome during the synthesis showed enzymatic activity. (Kiho and Rich, 1964) N-terminal fragment of Semliki Forest Virus capsid protein showed self-cleavage enzyme activity during the process of polyprotein synthesis (Nicola et al., 1999) and this rapid cotranslational folding did not require adjuvant cellular components. (Sanchez et al., 2004) First 86 residues out of 147 residues of the whole protein which was truncated from the nascent alpha-globin chain on the ribosome showed specific heme binding activity (Komar et al., 1997) N-terminal

parts of the tandem immunoglobulin domain protein which is bound to ribosome folded into native structure while C-terminal moiety failed to be correctly folded. (Hsu et al., 2007) Disulfide bond formation was observed from nascent immunoglobulin peptides which partly indicate the possibility of the determination of the native conformation during the process of protein synthesis. (Bergman and Kuehl, 1979(1); Bergman and Kuehl, 1979(2)) Also, there was a report of the alpha-helical structure formation while the newly synthesized peptide chain traverses the exit tunnel of the ribosome. (Lim and Spirin, 1986)

Bacterial luciferase is a heterodimer of alpha and beta monomer. It was found that the cotranslational folding of beta monomer in the presence of alpha monomer is faster than the renature of beta monomer in vitro. (Fedorov and Baldwin, 1999) This possibly shows the avoidance of kinetic traps associated with refolding from denatured beta monomer in cotranslational folding. (Evans et al., 2005) Similar phenomenon was observed in firefly luciferase where native-like structure was found during the cotranslational folding. (Frydman et al., 1999) There are numerous other experimental evidences of cotranslational folding. (Fedorov and Baldwin, 1997; Basharov, 2000; Basharov, 2003; Kolb, 2001; Giglione et al., 2009; Koadokura and Beckwith, 2009)

Computational models of cotranslational folding has also been appeared, one of which early attempts was the one of Srinivasan and Rose. (Srinivasan and Rose, 1995) Computational models have provided evidences that nascent chains may adopt partial structures similar to the corresponding parts of the complete protein. (Lu and Liang, 2008) There are also numerical evidences of directional asymmetry for the cotranslational folding. N-terminal region was revealed to be more compact than C-terminal region. (Alexandrov, 1993) Difference of nitrogenous and carboxyl termini in the Gaussian integral which is commonly used in knot theory was observed. (Røgen, 2005) Numerical analysis of the final folds also showed that the most likely path of the folding would be the one from the N-terminus. (Norcoss and Yeates, 2006) Srivastava et

al. (Srivastava et al., 2011) have shown that the correlation between the cotranslational property and the nitrogenous terminal hydrophobicity.

A general trend towards cotranslational folding was observed in the analysis of SCOP database. (Saunders et al., 2011) The trend was most strongly observed in the α/β class where 66% having more centrally oriented N-terminus and 71% having more N-terminal set of core residues. (Saunders et al., 2011) Ellis et al. (Ellis et al., 2010) showed the impact of the consideration of directionality in the protein structure prediction by modified ROSETTA algorithm. Molecular dynamics result of small peptides also showed the possibility of the formation of the native conformation of full protein. (Voelz et al., 2009)

It might be worth to test the validity of cotranslational folding by observing the structural difference of the simulated protein structure from the experimentally determined structure. In this chapter the cotranslational folding has been implemented as the algorithm that generates the initial structure for the whole simulation. However, the structure generated from cotranslational folding might undergo further modifications within the cellular environment. Thus, the consideration of the structural modification after the initial folding is necessary in spite of the ample evidences of the importance of cotranslational folding. A new folding simulation algorithm following cotranslational initial path and further torsional modifying path was suggested in this study.

Protein structures are usually determined from experimental methods including X-ray crystallography and NMR spectroscopy. However, limitations of experimental methods necessitate the theoretical methods of protein structural determination. One of the most critical problems of theoretical anticipations is the colossal magnitude of the possible space that conformations can adopt. Lattice model, Gō model, and many constraints and restraints were developed to overcome this problem. Levinthal also supposed paradox between the sheer overcoming magnitude of the number of possible conformations and the time scale of protein folding. According to Levinthal's postulations, proteins would

have predetermined path of the search of the native conformation in the space of possible conformations. This approach of Levinthal's postulation is, however, not generally used in the current protein folding study. The difficulty of finding the correct folding path might be the one of the most deterring factors of the folding researches based on the predetermined path. Path following the Levinthal paradox generally strongly shrinks the space of possible conformations. Also, the correct postulated path might lead to the accurate simulation of real native structures. Though the scholastic attempt to the methodological development of folding algorithms following Levinthal paradox is worth to make, many researchers are not concerned in this subject mainly for the possible doubts of robustness of any predetermined path.

Instead of searching without following the path of Levinthal paradox, a search which follows the iterative further folding pathway which is determined by heuristic torsional property of amino acid residue after the initial cotranslational folding was conducted in this research. It was supposed that the torsional priority would follow the degree of possible torsional perturbations; i.e. the amino acids with stronger torsional perturbation would tort first while the ones with less perturbation would tort later. Amino acid with side chain of heuristically longer length was supposed to receive stronger disturbances from the collisions with the solvent molecules of the cellular environment. Among the amino acids with similar lengths of side chains, the one with more planarity characters was supposed to receive the higher magnitude of torsional disturbance for the better transduction of torsional collision to the Cα atom of the protein backbone. Following this and other rules, the sequence of energy calculation was determined and converging energy minimum was obtained by iterative optimization.

Protein structure modeling application named ProtTorter which utilizes torsion angle system to reduce the unnecessary degree of freedom has been developed. Using this application which was described in the previous chapter, the prediction of the structures of small peptides was made. The structural similarity between the predicted structure

and experimentally determined structure of the small peptides of 8 amino acids which is obtainable from the Protein Data Bank was made.

6.2 Materials and Methods

To validate the robustness of the cotranslational folding and torsional Levinthal path modification, structure of short amino acid polypeptide was constructed and compared with the real structure determined by solution NMR method. The most representative model among the compiled models of each NMR structure was compared with the predicted structure by the new folding algorithm using backbone torsion angle alignment algorithm developed in the chapter 3. Typical structure alignment algorithms including TM-align (Zhang and Skolnick, 2005) were not used for their movement of frame with insertions and deletions even in the case of the comparison of identical amino acid sequence. The result of the comparison was represented by logPr (Jung et al., 2011) and RamRMSD (Jung et al., 2011) and the graph of torsion angle along the residue number (Figure 6-2). The change of the potential energy during the initial folding and following torsional path optimization is also shown in the graph of figure 6-3. Larger dataset of 15 peptides were also used for the validation of small iteration numbers for energetic convergence and low energy modeling of conformations.

6.2.1 Dataset

Among many possible conformations attainable from the PDB web server, the structure of a single asymmetric chain was chosen for easy analysis of the structures. Structures which contain only the protein and not nucleic acids of DNA, RNA or DNA-RNA hybrid were chosen. The length of the chain was limited to 8 during the generation of PDB web server search query. Structures with more than 90% sequence identity were removed from the final search result. Also structures containing hetero atoms were also removed for the convenience of computational parsing. Using these criteria, structure of 1n9v and 1oeh were obtainable from PDB. We used 1n9v which is the angiotensin

peptide and neglected the fragmented structure of 1oeh. 15 assumptive peptides were also used for further validation of the fast convergence and the simulation of low energy conformations. 4 dipeptide, 4 tripeptide, 4 tetrapeptide and 3 pentapeptide were incorporated in the test set (Table 6-3). The sequence of the peptide was randomly selected from 20 possible cases.

6.2.2 Cotranslational Folding of Initial Structure

There were previous attempts of the implementation of cotranslational folding concept such as SAINT (Ellis et al., 2010) which utilized three-dimensional molecular dynamics. Here, backbone torsion angle system was used for the description of the movement of atoms of amino acid residues. Molecular dynamics treats all the movements within the Cartesian space as possible while employing strong covalent bond stretching and bond angle bending terms in the force field. Though the force field term corrects the unrealistic modeling of movements, the description of typical molecular dynamics method strongly interrupts the insightful explanation of the change of the conformation of proteins which only rotates along the covalent single bonds in practice. This interruption strongly affects the impossibility of human interpretation of the possible mechanisms of the folding of a protein. This is also the reason of the limitation of possible helpful manual intervention of three-dimensional models. Torsion angle system, however, remedies the problems of insightful representation of the conformational change. It is also more realistic for the description of protein conformation than three-dimensional system in lattice models.

Utilizing the benefits of torsion angle representation of conformational space, cotranslational folding was implemented to generate preliminary structure by using ProtTorter modeling application of ours. C-terminus of ribosome bound nascent protein is much heavier than free N-terminus. Thus, the potential energies for every possible conformation following the change of φ and ψ angle each were calculated after every addition of a new amino acid residue assuming the priority of torsional movement of more N-terminal residues. The local minima and the global minimum were calculated

from the potential energies of each of the 360 degrees for each torsion angle of the newly formed peptide bond. Only the global minimum angle was chosen to predict the initial structure of the peptide of a specific sequence.

6.2.3 Iterative Optimization of Initial Structure Following Torsional Folding Path

The amino acid which receives stronger torque during the collision within the solution environment was regarded as more torsional. Considering the fast speed of colliding free atoms and the long length of the concatenated polymer, the individual torsional inertia of each residue was considered less important. Residues which are longer from the backbone were considered to receive more torsion from the collisions with environments. The length of the side chain was approximated as the number of bonds of the side chain from the Cα atom to the outermost atom. Among the amino acids with similar length of side chain, amino acids with more planarity characters were heuristically regarded as more torsional for their more conductivity of the collisional momentum to the backbone dihedral angle. In the case of the same degree of planarity and side chain length, the amino acid with more possible paths of bonds to the outermost atom was regarded to be more torsional. In the case of the same number of possible maximal length paths and the same condition of other criteria mentioned above, amino acid with bigger terminal atoms was regarded to be more torsional. Proline was considered to be planar along all the bonds that consist of the loop structure and the size of γ-oxygen and γ-sulfur atoms were compared in the case of cystein and sulfur. The torsional property of 20 amino acids could be ordered as "W(1)-Y(2)-R(3)-F(4)-K(5)-H(6)-Q(7)-M(8)-E(9)-N(10)-L(11)-I(12)-P(13)-D(14)-V(15)-T(16)-S(17)-C(18)-A(19)-G(20)" where the smaller numbers in the brackets designate the higher torsional priority.

Using the criteria above, the 8 amino acids of the peptide 1n9v were ordered by their torsional property as "D(7)-R(2)-V(8)-Y(1)-I(5)-H(4)-P(6)-F(3)" where smaller number indicates the more priority of the calculation of the potential energy. The order of the

bond for the energy calculation followed the order derived from torsional propensity. Thus, bond in the vicinity of Tyr(Y) was calculated first. The sequence of calculation among two bonds in the vicinity to an amino acid was determined from the sum of the priorities of the connected amino acids by the bond, where smaller sum receives the higher priority. Thus, the 4th peptide bond between Tyr(Y) and Ile(I) was calculated first (1+5=6) than the bond between Tyr(Y) and Val(V) (1+8=9). The order of priority of calculation of energy is "D-(3)-R-(4)-V-(2)-Y-(1)-I-(6)-H-(7)-P-(5)-F" where the number in the round brackets signifies the order of the linkage between the amino acids next to the bracket. Between the dihedral angles (φ and ψ) next to the peptide bond in inter-residue linkage, the dihedral angle near to the higher torsional priority was calculated first. Calculation along the torsion angles of peptide bond was ignored. Using this order of energy calculations, the initial structure obtained from cotranslational concept was iteratively optimized. The iteration continued until the convergence of potential energy was obtained. In the test case of 1n9v PDB entry, 6 times of the iteration of torsional path optimization induced rather conserved modified structure from the initial structure. The results of the case of 15 assumptive peptides are shown in Table 6-3.

6.3 Results and Discussion

The structure of 1n9v predicted by our cotranslational initial structure with iterative optimization along torsional path using the application of chapter 5 is shown in figure 6-1. The generated structure from angiotensin peptide showed loop structure for the most of the 8 residues except the first aspartate residue. The (φ, ψ) dihedral angles usually located around the range of (25º-30º, 0º-5º). This spiral like structure of 1n9v was formed only from the electrostatic interaction of atoms and the torsional barrier of rotatable bonds. The motive force to the more compact conformation might shorten the length of the loop by inducing more strong turns among the residues by occupying other

possible local minima. Considering that α-helical structure was observed from lattice model without the consideration of any detailed electrostatic or torsional potential energy (Leach, 2001), it might be possible that additional restraints of non-electrostatic interaction would induce modifications to the current type of loop structure into well known helices.

The simulated structure was determined from very small search space. As shown in table 6-1, initial structure from cotranslational folding was determined from about 52 possible local minima conformations. Usually, about 70-90 local minima structures were required for the determination of a representative structure for this 8 residues peptide. Maximum of 86 local minima conformations were required for each iteration. This requirement is rather strongly efficient when compared with the typical molecular dynamics which needs myriad number of trajectories. Details of the values of angles and numbers of considered local energy minima for each torsional bond and for each iterations of optimization including the initial structure generation is displayed in table 6-1. The generated structure showed difference to the experimentally determined structure (Figure 6-2). While experimental structure showed rather regular oscillation along the 0º line, simulated structures mainly showed positive angles in the N-terminal region. Also, optimized structure showed strong deviation in the C-terminal region.

Alignment of the structure was also made to assess the similarity among simulated and experimentally determined structures with the backbone torsion angle alignment algorithm introduced in chapter 3 (Table 6-2). Typical structure alignment program including TM-align (Zhang and Skolnick, 2005) was not utilized for their changing of alignment frame by incorporating insertions and deletions for the structures of identical amino acid sequence. We modified our algorithm of backbone dihedral angle alignment to only consider the non-gapped identical frame.

logPr value signifies the difference of the two compared structures with weight imposition on the more closer similarity. The logPr values of simulated structures

showed increasing tendency toward the pairs of later iterations signifying the converging property during the iterative optimization. The lowest logPr value of -15.18 was observed from the pair of 5^{th} optimization and 6^{th} optimization. Pairs of nearer iterations showed lower logPr values than the pairs of farther iterations. The highest logPr value among the pairs of direct modification was -4.62 of initial structure and the first optimized structure. This might reflect the different property of the paths of folding for the cotranslational path and the torsional propensity path. RamRMSD shows the RMS deviation of the position of a pair of backbone torsion angles of a residue on the Ramachandran plot. The values of RamRMSD showed similar tendency to logPr values including the growing similarity for the pairs of later iterations. Pairs of nearer iterations showed lower RamRMSD values than the pairs of farther iterations. The highest RamRMSD value among the pairs of direct modification was 47.17 of initial structure and the first optimized structure possibly reflecting the different scheme of the path of the folding between the initial structure generation and iterative optimization.

The structure determined from NMR spectroscopy showed rather different conformation from simulated structures. The average logPr values of experimental structure from all the structures of iterations ranged from -0.80 to -1.01. RamRMSD varied from 116.68 to 136.28. Though the simulated structure was quite appreciable regarding the low and negative potential energy of -1.704 (kcal/mol), it showed rather strong deviation from experimentally determined structure. This negative potential energy signifies that this structure is stable in the vacuum environment. Regarding that this structure is an energy minima following the path of torsional propensity Levinthal path, energetically stable structure might suppose the possible real energy minima structure which exist out of all possible conformation space.

A possible reason for the deviation of the simulated structure from the experimental structure is the utilized force field. NMR spectroscopy and general process of protein folding is conducted within an aqueous solution while this simulation is conducted under the supposition of vacuous environment. The neglect of the interaction of solvents

to the protein molecule might have introduced difference of the simulated structure to the NMR structure. Hydrophobic effect or free energy related to solvent accessible surface which was determined empirically in previous research could be applied to the generation of the simulated structures. Other possible reason for the difference might be the versatility of NMR models which arises from the motile behavior of the conformation of a small peptide. Thus, in a single result of NMR structure, multiple energy minimal conformations exist. Though we exploited the representative model with the lowest potential energy among the scores of models, it might possible that real structure is more similar to the one of the non-representative models. One of the most reliable reference of the folding research of small peptides is thus to compare the modeled structure with the modeling results from molecular dynamics simulation using general force fields. Therefore, the general unreliability of NMR structure might be one of the reasons for the difference of the modeled structure to the experimentally observed structure.

The difference of the structure from the experimental structure might be due to the falsity of the path which we supposed to be the one for the protein folding process. This point, however, is intriguing considering the fast convergence of the iterations. Finding converging structure for six iterations is quite fast which partly proves the validity of this path, though there is a possibility of falsity and of any other correct folding paths. A further validation set of fifteen assumptive peptides were modeled for the proof of this path (Table 6-3). Among all the cases, the necessary iteration number for the energetic convergence was in the range from one to thirteen. This range of iteration number is possible to be regarded as strongly small considering the complexity of typical molecular dynamics or Monte Carlo methods. Dipeptides showed the least number of iterations of the range from one to two, partly representing the small space of possible conformations. Mean number of iterations of 1.25 was necessary for the convergence of dipeptides. Tri-, tetra- and pentapeptides also showed rather small average number of iterations of 3.25, 6.00 and 4.67 each. Mean number of 3.73 iterations were needed for

convergence among all the cases. For every case of peptides, minimum numbers of iterations were in the range from one to three. These numbers which are smaller than three indicate that this method results the converging conformations very fast in possible cases of different magnitude of search spaces. Potential energy ranged from -2.620 (kcal/mol) to 2.088 (kcal/mol) and showed mean of -0.388 among all the cases of assumptive peptides. Eight cases out of fifteen ones showed negative energy conformations. Negative energy conformations were also observed in every cases of the length of peptides possibly indicating that this method is effective in the derivation of minimal energy conformations. Mean potential energies of 0.238, -1.037, -0.333 and -0.433 (kcal/mol) were observed for the case of di-, tri-, tetra- and pentapeptide each. This low mean potential energy partly validates the robustness of modeling of our new method. Minimum potential energy showed similar values in all the range of the length of the peptides and no correlations with the peptide length. The lowest potential energy of the eight residue peptide conformations was also -1.704 (kcal/mol) which is not far different from the cases of 2-5 residue peptides. This rather uniform distribution of potential energy might suggest that there is no strong difference in the stability of the most stable conformations from different magnitude of the space of possible conformations.

The change of energy during the folding of initial structure and following iterations of optimization of angiotensin octapeptide is shown in figure 6-3. The potential energy showed rather drastic fluctuation during the generation of initial structure using cotranslational folding. This partly indicates that addition of amino acid is either favorable or unfavorable for each different circumstance. This fluctuation is different from following iterative optimizations reflecting the strong effect of the change of configurations. During the iterations of optimizations, the potential energy decreased with saltatory tendency. This indicates the possibility of additional modifications of the initial structure generated by simple cotranslational path of folding. This also indicates that there are a few critical bonds which strongly influence the potential energy of the

whole molecule. After three iterations of optimizations, the potential energy remained as being rather conserved demonstrating the fast convergence of the algorithm to the local energy minima.

One of the causes of the difference of the simulated structure from NMR structure might be the absence of the consideration of the solvent effects. Consideration of solvent effects might include optimized dielectric constants into the force fields and solvation free energy term into the potential energy calculation. Eisenberg and McLachlan developed simple system for the representation of such solvation free energy which depends on the exposed solvent accessible surface area and atom type (Eisenberg and McLachlan, 1986). Considering the prevalence of neutral atoms including typical carbons, the solvation energy might be approximately correlated with the exposed solvent accessible surface. Interestingly, the minimal surface area structure is a sphere which was validated to be the general structure of proteins. Thus, globular structure could be used as a folding criterion also for the representation of the solvent effects. Further study based on this argument might be appreciable.

6.4 Conclusion

The simulated structure showed rather stable local minima structure with the energy of -1.704 (kcal/mol) while the comparison of the predicted and experimental structure of 1n9v showed structural difference with RamRMSD and logPr. The difference might be originated from the use of *in vacuum* force field and unreliable versatility of the NMR structures. The falsity of the postulated folding path might also have been one of the reasons of the difference. The generated structures from initial structure preparation and further optimizations showed the most correlations with the structures of direct modifications while this correlation increased for the later rounds of iterations. The energy showed saltatory drops during the folding simulation. Especially the energy is rather conserved from the 3rd iterations to the final round of iteration.

One intriguing point of this method is to find the convergent conformation after very few rounds of iterations with the consideration of small number of possible conformations. Less than 86 conformations were regarded for every rounds of conformational optimization for octapeptide. It only took 5 rounds of iterations to reach the local minimal energy conformation for angiotensin octapeptide. Mean number of 3.73 iterations were required for 15 peptides of the length from two to five. This very fast convergence of the modeling structure possibly indicates the distinguished robustness of the method. This process of folding was possible to be handled entirely manually and visually through ProtTorter introduced in the chapter 5. By this simple process, it was possible to obtain energetically stable local minima structure which is still smaller than that of the structure (13.363 (kcal/mol)) which is identical in backbone dihedral angles to the experimental structures.

The partial validity of this method of cotranslational initial conformation generation and torsional Levinthal path optimization was proved by stable negative energy conformation which possibly suggests the correct representation of the method of the folding process. This folding algorithm resulted negative energy of -1.704 (kcal/mol) after 5 rounds of iterations. The structure with the same backbone dihedral angles but with different side chain conformations to the experimental representative structure showed energy of 13.363 (kcal/mol). Though the side chain conformation might be different, this strong positive value should be regarded as unstable if one accepts the utilized force field is quite valid. Negative mean potential energy of -0.388 (kcal/mol) and negative minimum potential energy of -2.620 (kcal/mol) were also observed for 15 cases of peptides of different lengths. The negative potential energy also signifies the strong stability of simulated conformation which partly supplies the correctness of the new method. This validity also possibly suggests that the cotranslational formation of three-dimensional structure have rather profound influence as an initial structure on the formation of the native conformation of a protein. The partial validity of the optimization through torsional propensity based path possibly suggests the real folding

pathway which would strongly reduce the immense amount of possible conformational candidates. The success of protein modeling using backbone torsion angle system of this chapter also suggests the proper utility of this system in the representation of possible space of conformations.

Also, the successful result of this algorithm suggests the new possibility of ample human interactions and interpretations during the protein modeling process. Though it is doubtful that this algorithm is purely applicable for further larger proteins, this folding algorithm is quite intriguing because it strongly suppresses the number of considered conformations and suggests possible new Levinthal path which might be partly validated by very fast convergence. It needs much further validations to prove that this algorithm does represent the real path of folding which might reduce the myriad possibilities during simulations. There is also a possibility of the utilization of inappropriate force field and folding pathway. Though some difference was observed from the experimental structure, this trial for the generation of energy minimal structure from theoretical Levinthal path might be appreciable considering the fast converging property and stable negative potential conformation. Future study might embellish this research of first attempt to be acceptable by implementing solvent related criteria.

Table 6-1. Backbone Dihedral Angle and Number of Energy Minima for Various Iterations

	init. †	opt. 1‡	opt. 3	opt. 4	opt. 5	opt. 6
1st bond	ang.(min.#)					
ψ	-171(4)	-171(4)	-171(4)	-171(4)	-171(4)	-171(4)
φ	26(4)	27(4)	26(4)	26(4)	26(4)	26(4)
2nd bond						
ψ	2(2)	1(5)	5(5)	5(5)	5(5)	5(5)
φ	35(5)	31(6)	30(6)	30(6)	30(6)	30(6)
3rd bond						
ψ	1(2)	0(10)	1(8)	1(8)	1(8)	1(8)
φ	31(4)	27(9)	26(8)	26(8)	26(8)	26(8)
4th bond						
ψ	1(2)	2(6)	4(8)	3(7)	3(7)	3(9)
φ	-150(7)	-39(10)	-35(12)	-34(11)	-34(11)	-34(10)
5th bond						
ψ	-153(12)	-24(5)	-179(11)	-179(11)	180(11)	180(11)
φ	-179(8)	-169(5)	-166(5)	-167(11)	-169(11)	-170(11)
6th bond						
ψ	4(3)	-154(8)	-176(8)	-175(8)	-174(8)	-174(8)
φ	-28(4)	-67(5)	-69(5)	-68(5)	-68(5)	-68(5)
7th bond						
ψ	-176(3)	180(3)	6(5)	6(3)	6(3)	6(3)
φ	168(5)	168(5)	34(5)	35(8)	35(8)	35(7)
Search Space	52	72	81	86	86	86

† initial structure generated from cotranslational folding with torsional energy calculation

‡ optimized structure following the folding path determined by the torsional propensity

Table 6-2. Structural Similarity among Structure Models from Various Iterations

	init.	opt. 1	opt. 2	opt. 3	opt. 4	opt. 5	opt. 6	exp. s.
init.†								
logPr	-16.00	-4.62	-3.50	-4.80	-4.84	-4.84	-4.85	-0.80
RamRMSD	0.00	47.17	49.75	78.61	78.39	78.28	78.27	119.67
opt. 1‡								
logPr		-16.00	-6.75	-3.88	-3.98	-4.84	-4.02	-0.80
RamRMSD		0.00	29.93	61.63	61.27	61.25	61.26	134.82
opt. 2								
logPr			-16.00	-7.02	-5.29	-5.25	-5.23	-1.01
RamRMSD			0.00	45.08	45.55	45.73	45.81	116.68
opt. 3								
logPr				-16.00	-11.08	-10.14	-10.13	-0.84
RamRMSD				0.00	0.68	1.00	1.11	135.88
opt. 4								
logPr					-16.00	-13.50	-13.47	-0.82
RamRMSD					0.00	0.41	0.52	136.08
opt. 5								
logPr						-16.00	-15.18	-0.82
RamRMSD						0.00	0.13	136.23
opt. 6								
logPr							-16.00	-0.82
RamRMSD							0.00	136.28
exp. s.*								
logPr								-16.00
RamRMSD								0.00

† initial structure generated from cotranslational torsional folding

‡ optimized structure following the sequence of folding based on the torsional propensity

* experimentally determined structure

Table 6-3. Number of necessary iterations and potential energy of the simulation of the conformations of 15 assumptive peptides

sequence	iteration number	*mean*	*minimum*	potential energy (kcal/mol)	*mean*	*minimum*
DR	1			0.516		
VF	2			-1.980		
DP	1			2.088		
YV	1			0.327		
		1.25	1		0.238	-1.980
DHR	5			-0.723		
HVG	2			-2.453		
YIH	3			-0.852		
YDR	3			0.120		
		3.25	3		-3.908	-2.453
IYVR	3			-0.719		
PFIH	3			-0.473		
YVRD	5			-0.902		
DRPF	13			0.762		
		6.00	3		-1.332	-0.902
AGDYH	6			-2.620		
PFDGR	5			0.267		
PRYGD	3			1.054		
		4.67	3		-0.433	-2.620
total		3.73	1		-0.388	-2.620

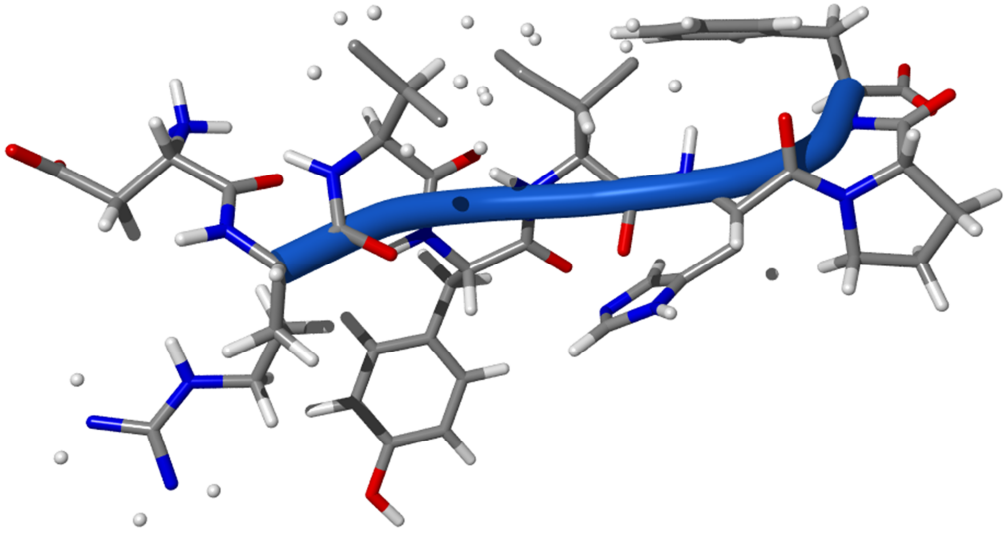

Figure 6-1. Simulated Structure of 1n9v PDB Entry Sequence. The three-dimensional stick model of the structure of the angiotensin peptide (PDB entry 1n9v; "DRVYIHPF") is displayed. Loop structure was found among the most of the peptide (from 2nd to 8th residue) and drawn with blue cylinder. Carbon atoms are shown with gray color, nitrogen atoms with blue color, oxygen atoms with red color, and hydrogen atoms with white color. Some hydrogen atoms are shown as single spheres ignoring the bondage information.

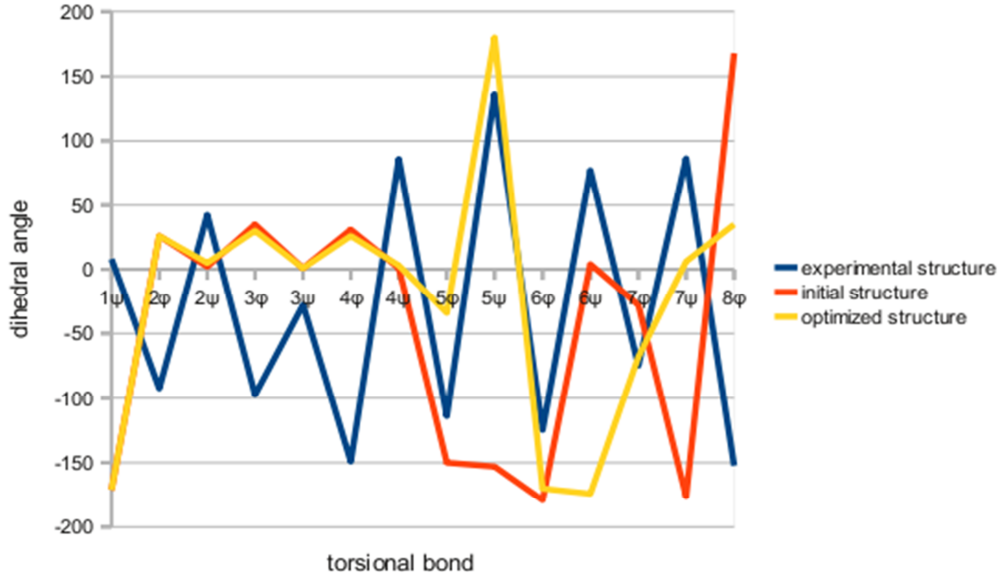

Figure 6-2. Dihedral angles of Torsional Bonds of Simulated Structures and Experimentally Determined Structure. Dihedral angles of the most representative structure of model 21 of NMR experiment were shown with blue line. Dihedral angles of initial structure from the sequential search of torsional global energy minima following the concept of cotranslational folding were shown with orange line. Dihedral angles of the final structure from the iterative optimization according to the sequence following torsional property were shown with yellow line. While experimental structure showed rather regular oscillation along the 0º line, simulated structures mainly showed positive angles in the N-terminal region. Though initial structure and final structure are similar, optimized structure showed strong deviation in the C-terminal region.

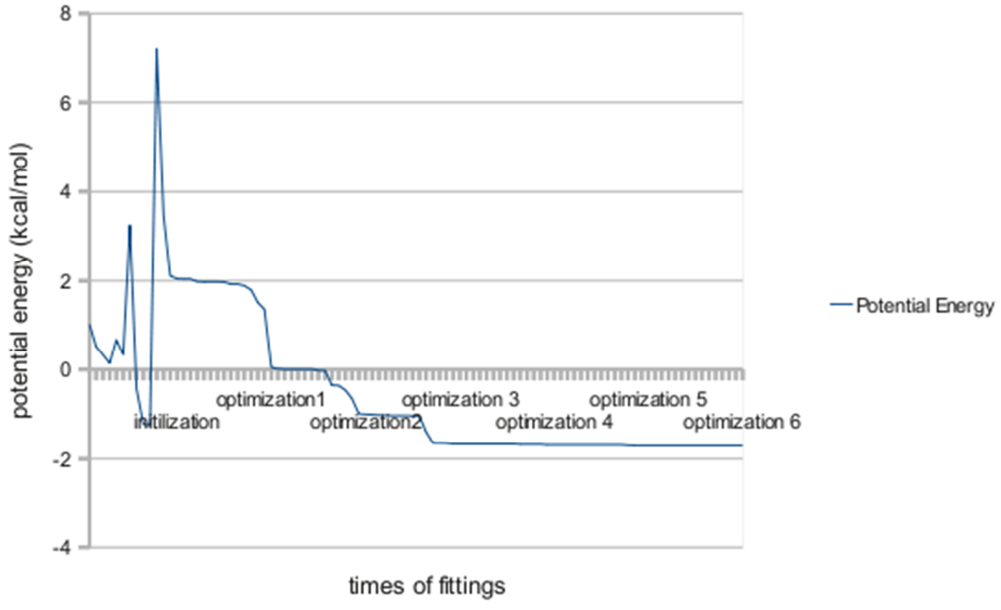

Figure 6-3. Change of Potential Energy during the Initialization and Optimizations. The change of potential energy during the initial structure generation and further iterative optimization is shown in blue line. The potential energy shows drastic fluctuations during the initialization of cotranslational folding which partly indicates that addition of amino acid is not always either favorable or unfavorable. This fluctuation is different from the following iterative optimizations reflecting the strong effect of the change of configurations. During the optimization processes, the potential energy decreased with saltatory tendency. After three iterations of optimizations, the potential energy remained as being rather conserved indicating the fast convergence of the algorithm to the local energy minima. The potential energy perfectly converged after 6 iterations.

CHAPTER VII.

Summary

Proteins are primary component of biological organisms and performs various chemical and physical functions within the cells and tissues of organisms. The structure of a protein is strongly related with the function of the protein and is believed to adopt a single conformation in a native state. Though X-ray crystallographic methods and NMR spectroscopical methods are applied for the determination of the structures, delicate cases of possible observations are still limited. Experimental studies would cost much experimental and temporal resource. The importance of the structural information, however, necessitates the development of genuine and accurate theoretical methods. While *ab initio* folding methods are being developed and continuously imporved, the major problem of folding study in regard of Levinthal paradox still remains. In here, structural globularity of proteins were analyzed as possible criterion of folding algorithms. The validity of the more native like torsion angle system was partly validated in the structural alignments. This torsional system was exploited in the construction of structure analysis web server and modeling stand-alone application. A simple algorithm of folding was attempted using the developed modeling applications.

General background knowledge of protein composition, biosynthesis, and basic structural characteristics were referred in the introduction chapter. Schematic description of general principles and limitations of widely used current experimental method which determines the three-dimensional structure of protein has also been made in chapter 1. The necessity of the research of protein structure model calculation and various protein structure modeling methods including homology modeling, threading, and *ab initio* modeling was also introduced. Various *ab initio* methods including molecular dynamics,

Monte Carlo method, energy minimization method, simulated annealing, and replica exchange method were schematically described. The impossibility of exhaustive search of conformational space which is the fundamental problem of protein folding simulation was also mentioned in the introduction chapter. Simplified approach of lattice model, Gō model, and the necessity of constraints and restraints were also described.

Universal structural feature of native proteins were concerned as the possible restraints for the protein structure simulations. The globularity of proteins was recently incorporated as a folding simulation criterion, and globularity expressed in the radius of gyration was used to improve and validate NMR protein structures. Globularity was also successfully applied to assess the quality of models submitted to the Critical Assessment of Techniques for Protein Structure Prediction center, but an analysis of the globularity of proteins covering a whole database of protein structures had not been conducted.

In chapter 2, this author investigated if globularity is a general characteristic of proteins and whether they can be applied as a valid constraint in protein structure simulations with approximated measurements named Gb-index. Unexpectedly, most of the proteins showed strong structural globularity with only a few percent of proteins being outliers. Mode of approximately 76% similarity to the perfect globe was observed. Small proteins tended to be significantly non-globular ($R^2 = 0.79$) and the minimum Gb-index showed a logarithmic increase with the increase in protein size ($R^2 = 0.62$), strongly implying that the non-globular characteristics might be more acceptable for smaller proteins than larger ones. The distribution of the degree of globular structures of 7131 proteins was confirmed. The strong perfect globe-like character and the relationship between small size and the loss of globular structure of a protein found in chapter 2 may imply that living organisms have mechanisms to aid folding into the globular structure to reduce irreversible aggregation. This also implies the possible mechanisms of diseases caused by protein aggregation, including some forms of trinucleotide repeat expansion-mediated diseases.

Among many possible reliable constraints that reduce the degree of freedom, torsion angle approach was focused. Torsion angle constraint mimics natural process of conformational change of proteins which lacks significant movement of the positions along covalent bonds and bond angles. However, if the constraint that regenerates the process of protein folding would also be helpful for the analysis and prediction of protein structure was still doubtful. In chapter 3, the torsion angle system was applied to structural similarity assessment to investigate the robustness and validity.

Previous researches already noticed that a 3D backbone structure can be mathematically represented with a 1D φ and ψ dihedral angle array. However, performance of the backbone dihedral angle alignment was not supported with sufficiently large test sets to be quantified; i.e. only 2 pairs or 4 pairs of proteins were analyzed. In chapter 3, this author showed that it is more effective to accurately anticipate homology among 1891 pairs of proteins of 62 different proteases and 1770 pairs of 60 proteins of kinases and proteases with the string of φ and ψ dihedral angle array than famous 3D structural alignment tool TM-align showing the robustness of the torsion angle system. It is even more evident considering that gapless global alignment between protein structures was conducted to validate the effectiveness of performing structural alignment with strings of backbone torsion angles. Representation of 3D structure by 1D torsion angle strings allows local alignment, profile construction, hidden Markov models to be implemented with minor modifications and with almost no loss of speed compared with sequence alignment. By the further validation from the previous small-scale studies, the utility of backbone dihedral angle method became more evident.

Protein structure is hierarchically classified as primary, secondary, tertiary and quaternary structures. The secondary structure of a protein might be considered as the local structure with repetitive hydrogen bonds which consists the tertiary structure when combined. Many prediction algorithms and applications were developed to predict the secondary structure from amino acid sequences. There also exist protein fold classification databases where secondary structure is one of the most important criteria

that are used for the categorization. In chapter 4, this author referred the construction of a secondary structure database from PDB and SCOP entries based on the simple classification scheme according to the backbone torsion angles. The database introduced here offers functions of secondary structure database searching, secondary structure calculation, and pair-wise protein structure comparison.

Graphical representation of three-dimensional protein structure is quite valuable for various structural analyses considering the complexity of three-dimensional information. Visualization during the process of the protein folding simulation is quite interesting regarding the fast apprehension of the states. The direction-dependent or cotranslational protein folding algorithm offers very broad option of interference of manipulation compared to the previous algorithms such as molecular dynamics. In this context, computational application which visualizes three-dimensional conformation, calculates the potential energy, and supplies the user interface for backbone torsion angle manipulation seemed to be valuable. ProtTorter, the newly developed application of protein three-dimensional structure viewer and modeler based on the backbone torsions, was introduced in chapter 5. This new application offers functionalities of simple structure viewing and backbone torsion angle modeling. Plausible candidate conformations could be derived fast and easily from the combination of local minima of each torsion angle by utilizing the application. Unfortunately, for the lack of side chain conformation variability, only the structures from rather smaller size peptide chains are correctly deducible.

Cotranslational and additional torsional folding path method was also concerned in the context of Levinthal paradox. The time needed for the amino acid residues to be fold is much shorter than the time needed for the addition of new amino acid to the growing string of peptide chain. This temporal discrepancy possibly indicates that the native conformation is strongly influenced by cotranslational folding of the initial three-dimensional structure though the modification after the translation might still exist. As a possible path following the Levinthal paradox, cotranslational folding of initial structure

strongly reduces the space of possible conformation which results from further modifications. Several trials were previously made to exploit cotranslational folding principle. In chapter 6, the validity of the folding method of cotranslational initial structure with torsional Levinthal path was investigated using the test sets of small peptides. This method was suggested as the possible solution of the protein folding problem following the nature of Levinthal paradox. Torsion angle system was employed to manipulate conformations and AMBER02 force field was used for the calculation of potential energy. Using the user interface of the computational application introduced in chapter 5, initial conformation of a peptide was constructed by sequential process of determination of the backbone torsion angles and the addition of new amino acid. Optimizations through folding process following the sequence of torsional perturbations were iteratively conducted. Positive result for the possibility of this method as the correct path following the Levinthal paradox was obtained as the stable negative local energy minimal structure after small times of optimization iterations for octamer peptide. Analysis of 15 peptides of different amino acid lengths also showed negative mean potential energy of -0.388 (kcal/mol) and very small times of mean iterations of 3.73.

The fact that this method is informative for the prediction of protein structure is quite intriguing for it have resulted the conformation from less than 86 candidate conformations for each round of modifications. Considering the positive results of the stable conformation of the method, the possible heuristic folding path signifies that actual folding mechanism might be somewhat heuristic that follows biological principle rather than being purely mathematic suggesting an implication of the importance of qualitative insights for solving sensitive physical problems. The deviation of the simulated structure from experimental structure might have been originated from inappropriate force field or path of folding. Consideration of solvent effect might help for the better predictions. The development of the more correct path of folding from initial cotranslational structure might be a valuable subject of future research. Also, the

constraints derived from general structural characteristics of native proteins might be helpful in the determination of the final structure.

The globular character of protein structure could be used to reflect solvent influence and folding criteria by minimizing surface area. Torsional system might strongly shorten the span of conformational search by imitating native like movements. Co translational and predetermined paths following Levinthal paradox would signify valuable unrecognized novel insights to the field of protein folding study. With the genuine findings from the general structural property of globular structure and the validity of torsional system on protein structure analysis to the positive results of the cotranslational algorithms for protein folding simulations, it is hoped to make novel investigations of the correct manners of the protein structure analysis. With the new information and scrutinized background knowledge illustrated in the thesis, the trial to regenerate the real folding pathway in the native environment is hoped to be tenable in further studies. The one partly validated in here might need further rigorous validations. Using the concept of the path of Levinthal paradox, torsional degree of freedom, and globularity restraints along with previously established methods, this author intends to do genuine research on protein folding which is important for both the elucidation of biological function and for the design of new molecule with specific function.

BIBLIOGRAPHY

Alexandrov, N., Structural Argument for N-terminal Initiation of Protein Folding, *Protein Science*, 2:1989-1991, 1993.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215: 403-410. 1990.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J., Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs, *Nucleic Acids Research*, 25:3389-3402, 1997.

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G., SCOP Database in 2004: Refinements Integrate Structure and Sequence Family Data, *Nucleic Acids Research*, 32(database issue): D226–D229, 2004.

Baker, D., and Sali, A., Protein structure Prediction and Structural Genomics, *Science*, 294: 93-96, 2001.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., and Nielsen, H., Assessing the Accuracy of Prediction Algorithms for Classification: An Overview, *Bioinformatics*, 16: 412-424, 2000.

Baldwin, T.O., Protein Folding in vivo: The Importance of Ribosomes, *Nature Cell Biology*, 1:154-155, 1999.

Basharov, M.A., Cotranslational Folding of Proteins, *Biochemistry (Moscva)*, 65:1380-1384, 2000.

Basharov, M.A., Protein Folding, *Journal of Cellular and Molecular Medicine*, 7:223-237, 2003.

Baxevanis, A.D. and Ouellette B.F.F., Bioinformatics: A practical Guide to the Analysis of Genes and Proteins, 3rd ed., John Wiley & Sons, p.199, pp.236-238, p.240, p.242, p.245, 2005.

Bergeron, J.J., Brenner M.B., Thomas D.Y., and Williams D.B., Calnexin: a Membrane-bound Chaperone of the Endoplasmic Reticulum, *Trends Biochem. Sci.*, 19:124-128, 1994.

Bergman, L.W. and Kuehl, W.M., Formation of an Intrachain Disulfide Bond on Nascent Immunoglobulin Light Chains, *Journal of Biological Chemistry*, 254:8869-8876, 1979(1).

Bergman, L.W. and Kuehl, W.M., Formation of Intermolecular Disulfide Bonds on Nascent Immunoglobulin Polypeptides, *Journal of Biological Chemistry*, 254:5690-5694, 1979(2).

Bonini, N.M., RNA toxicity is a component of ataxin-3 degeneration in Drosophila, *Nature*, 453: 1107-1111, 2008.

Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C.E., and Baker, D., Rosetta in CASP4: Progress in *ab initio* Protein Structure Prediction, *Proteins*, 45(Supplementary issue 5):119-126, 2001.

Bowie, J.U., Luthy, R., and Eisenberg, D., A Method to Identify Protein Sequences that Fold into a Known 3-dimensional Structure, *Science*, 253:164-170, 1991.

Brenner, S.E., Koehl, P., and Levitt, M., The ASTRAL compendium for sequence and structure analysis. *Nucleic Acids Research*, 28:254-256, 2000.

Briggs, M.S., and Roder, H., Early Hydrogen-bonding Events in the Folding Reaction of Ubiquitin, *Proceedings of National Academy of Science USA*, 89:2017-2021, 1992.

Bryant, S.H., and Lawrence, C.E., An Empirical Energy Function for Threading a

Protein Sequence Through a Folding Motif, *Proteins*, 5:92-112, 1993.

Cabrita, L.D., Dobson, C.M., and Christodoulou, J., Protein Folding on the Ribosome, *Current Opinion in Structural Biology*, 20:33-45, 2010.

Carpentier, M., Brouillet, S., and Pothier, J., YAKUSA: A Fast Structural Database Scanning Method, *Proteins*, 61:137-151, 2005.

Chan, H.S. and Dill, K.A., The Protein Folding Problem, *Physics Today*, Feb:24-32, 1993.

Chandonia, J.M., Hon, G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S.E., The ASTRAL Compendium in 2004, *Nucleic Acids Research*, 32:D189-D192, 2004.

Chandonia, J.M., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S.E., ASTRAL Compendium Enhancements. *Nucleic Acids Research*, 30:260-263, 2002.

Cieplak, P., Caldwell, J., and Kollman, P., Molecular Mechanical Models for Organic and Biological Systems Going Beyond the Atom Centered Two Body Additive Approximation: Aqueous Solution Free Energies of Methanol and N-Methyl Acetamide, Nucleic Acid Base, and Amide Hydrogen Bonding and Chloroform/Water Partition Coefficients of the Nucleic Acid Bases, *Journal of Computational Chemistry*, 22:1048-1057, 2001.

Constantini, S., Macchiano, A.M., and Colonna, G., Evaluation of the Structural Quality of Modeled Proteins by Using Globularity Criteria, *BMC Structural Biology*, 7:9, 2007.

DeBartolo, J., Colubri, A., Jha, A.K., Fitzgerald, J.E., Freed, K.F., and Sosnick, T.R., Mimicking the Folding Pathway to Improve Homology-free Protein Structure Prediction, *Proceedings of National Academy of Science USA*, 106:3734-3739, 2009.

Dunbrack, R.L.Jr. and Karplus, M., Backbone-dependent Rotamer Library for Proteins: Application to Side-chain Prediction, *Journal of Molecular Biology*, 230:543-574, 1993.

Ellis, J.J., Huard, F.P.E., Deane, C.M., Srivastava, S., and Wood, G.R., Directionality in Protein Fold Prediction, *BMC Bioinformatics*, 11:172, 2010.

Evans, M.S., Clarke, T.F., and Clark, P.L., Conformations of Co-translational Folding Intermediates, *Protein and Peptide Letters*, 12:189-195, 2005.

Fedorov, A.N., and Baldwin, T.O., Co-translational Protein Folding, *Journal of Biological Chemistry*, 272:32715-32718, 1997.

Fedorov, A.N., and Baldwin, T.O., Process of Biosynthetic Protein Folding Determines the Rapid Formation of Native Structure, *Journal of Molecular Biology*, 294:179-586, 1999.

Fischer, D. and Eisenberg, D., Protein Fold Recognition Using Sequence-derived Predictions, *Protein Science*, 5:947-955, 1996.

Fisher, D., Elofsson, A., Rice, D., and Eisenberg, D., Assessing the Performance of Fold Recognition Methods by Means of a Comprehensive Benchmark, *Pacific Symposium of Biocomputing*, pp. 300-318, 1996.

Frydman, J., Erdjument-Bromage, H., Tempst, P., and Hartl, F.U., Co-translational Domain Folding as the Structural Basis for the Rapid *de novo* Folding of Firefly Luciferase, *Nature Structural Biology*, 6:697-705, 1999.

Giglione, C., Fieulaine, S., and Meinnel, T., Cotranslational Processing Mechanisms: Towards a Dynamic 3D Model, *Trends in Biochemical Sciences*, 34:417-426, 2009.

Godzik, A., Kolinski, A., and Skolnick, J., *de novo* and Inverse Folding Predictions of Protein Structure and Dynamics, *Journal of Computer-Aided Molecular Design*, 7:397-

438, 1993.

Güntert, P., Mumenthaler, C., and Wüthrich, K., Torsion Angle Dynamics for NMR Structure Calculation with the New Program DYANA, *Journal of Molecular Biology*, 273:283-298, 1997.

Habelka, W.A., Henderson, R., and Oesterhelt, D., 3-Dimensional Structure of Halorhodopsin at 7 Å resolution, *Journal of Molecular Biology*, 247:726-738, 1995.

Henderson, R., Baldwin, J.M., Ceska, T.A., Zemlin, F., Beckmann, E., and Downing, K.H., Model for the Structure of Bacteriorhodopsin Based on High-resolution Electron Cryo-microscopy, *Journal of Molecular Biology*, 213:899-929, 1990.

Hilga, R.H., Togawa, R.C., Montagner, A.J., Palandrani, J.C., Okimoto, I.K., Kuser, P.R., Yamagishi, M.E., Mancini, A.L., and Neshich, G., STING Millennium Suite: Integrated Software for Extensive Analysis of 3d Structure of Proteins and Their Complexes, *BMC Bioinformatics*, 5:107, 2004.

Hogue, C.W.V., Cn3D: A New Generation of Three-Dimensional Molecular Structure Viewer, *Trends in Biochemical Science*, 22:314-316.

Holm, L., and Sander, C., Protein Structure Comparison by Alignment of Distance Matrices, *Journal of Molecular Biology*, 233: 123-138, 1993.

Höltje, H.D., Sippl, W., Rognan, D., and Folkers, G., Molecular Modeling: Basic Principles and Applications, 3rd ed., Wiley-VCH, p.32, 2008.

Hsu, S.T.D., Fucini, P., Cabrita, L.D., Launay, H., Dobson, C.M., and Christodoulou, J., Structure and Dynamics of a Ribosome-bound Nascent Chain by NMR Spectroscopy, *Proceedings of National Academy of Science USA*, 104:16516-16521, 2007.

Huang, X., and Powers, R., Validity of Using the Radius of Gyration as a Restraint in

NMR Protein Structure Determination, *Journal of American Chemical Society*, 123:3834-3835, 2001.

Hubbard, T.J., Ailey, B., Brenner, S.E., Murzin, A.G., and Chothia, C., SCOP: A Structural Classification of Protein Database, *Nucleic Acids Research*, 27: 254–256, 1999.

Humphrey, W., Dalke, A., and Schulten, K., VMD: Visual Molecular Dynamics, *Journal of Molecular Graphics*, 14:33-38, 1996.

Jenni, S. and Ban, N., The Chemistry of Protein Synthesis and Voyage Through the Ribosomal Tunnel, *Current Opinion in Structural Biology*, 13:212-219, 2003.

Jones, D.T., Taylor, W.R., and Thornton, J.M., A New Approach to Protein Fold Recognition, *Nature*, 358:86-89, 1992.

Jung, J., and Lee, B., Protein Structural Alignment Using Environmental Profiles. *Protein Engineering*, 13: 535-543, 2000.

Jung, S., Bae, S., and Son H.S., Validity of Protein Structure Alignment Method Based on Backbone Torsion Angles, *Journal of Proteomics and Bioinformatics*, 4:10, 2011.

Kabsch, W., A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors, *Acta Crystallographica section A*, 34:827-828, 1978.

Kabsch, W., A Solution for the Best Rotation to Relate Two Sets of Vectors, *Acta Crystallographyica section A*, 32:922-923, 1976.

Kabsch, W., and Sander, C., Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-bonded and Geometrical Features, *Biopolymers*, 22:2577-2637, 1983.

Kadokura, H., and Beckwith, J., Detecting Folding Intermediates of a Protein as It Passes Through the Bacterial Translocation Channel, *Cell*, 138:1164-1173, 2009.

Kaplan, W., and Littlejohn, T.G., Swiss-PDB Viewer (Deep View), *Briefings in Bioinformatics*, 2:195-197, 2001.

Karpen, M.E., Haseth, P.L., and Neet, K.E., Comparing Short Protein Substructures by a Method Based on Backbone Torsion Angles, *Proteins*, 6: 155-167, 1989.

Karplus, K., Barrett, C., and Hughey, R., Hidden Markov Models for Detecting Remote Protein Homologies, *Bioinformatics*, 14:846-856, 1998.

Karplus, K., Katzman, S., Shackleford, G., Koeva, M., Draper, J., Barnes, B., Soriano, M., and Hughey, R., SAM-T04: What Is New in Protein-structure Prediction for CASP6, *Proteins*, 61(Suppl 7): 135-142, 2005.

Karplus, K., Karchin, R., Draper, J., Casper, J., Mandel-Gutfreund, Y., Diekhans, M., and Hughey, R., Combining Local-structure, Fold-recognition, and New-fold Methods for Protein Structure Prediction, *Proteins* **53**(Suppl 6): 491–496, 2003.

Karplus, M., The Levinthal Paradox: Yesterday and Today, *Folding and Design*, 2:569-575, 1997.

Kendrew J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., and Wyckoff, H., A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis, *Nature*, 181:662-666, 1958.

Kihara, D., and Skolnick, J., The PDB Is a Covering Set of Small Protein Structures. *Journal of Molecular Biology*, 334:793-802, 2003.

Kiho, Y. and Rich, A., Induced Enzyme Formed on Bacterial Polyribosomes, *Proceedings of National Academy of Science USA*, 51:111-118, 1964.

Kim, M., Sun, C., Kim, J., and Yi, G., Whole Genome Alignment with BLAST on Grid Environment, *Proceedings of The Sixth IEEE International Conference on Computer and Information Technology (CIT'06)*, 2006.

Koadokura, H., and Beckwith, J., Detecting Folding Intermediates of a Protein as It Passes Through the Bacterial Translocation Channel, *Cell*, 138:1164-1173, 2009.

Kolb, V.A., Cotranslational Protein Folding, *Molecular Biology*, 35:584-590, 2001.

Kolb, V.A., Makeyev, E.V., and Spirin, A.S., Co-translational Folding of an Eukaryotic Multidomain Protein in a Prokaryotic Translation System, *Journal of Biological Chemistry*, 275:16597-16601, 2000.

Komar, A.A., Kommer, A., Krasheninnikov, I.A., and Spirin, A.S., Cotranslational Folding of Globin, *Journal of Biological Chemistry*, 272:10646-10651, 1997.

Kovacs, H., Mark, A.E., and Gunsteren, W.F.v., Solvent Structure at a Hydrophobic Protein Surface, *Proteins*, 27:395-404, 1997.

Kraulis, P.J., MOLSCRIPT: A Program to Produce both Detailed and Schematic Plots of Protein Structures, *Journal of Applied Crystallography*, 24:946-950, 1991.

Krieger, E., Nabuurs, S.B., and Vriend, G., Homology Modeling, *Methods of Biochemical Analysis*, 44:509-523, 2003.

Krissinel, E., Henrick, K., Secondary-structure Matching (SSM), a New Tool for Fast Protein Structure Alignment in Three Dimensions, *Acta Crystallographica Section D Biological Crystallography*, 60:2256-2268, 2004.

Krüger, M.K., Pedersen, S., Hagervall, T.G., and Sørensen, M.A., The Modification of the Wooble Base of tRNAGlu Modulates the Translation rate of Glutamic Acid Codons in vivo, *Journal of Molecular Biology*, 284:621-631, 1998.

Kubelka, J., The Protein Folding 'Speed Limit', *Current Opinion in Structural Biology*, 14:76-88, 2004.

Kuszewski, J., Gronenborn, A.M., and Clore, G.M., Improving the Packing and Accuracy of NMR Structures with a Pseudopotential for the Radius of Gyration, *Journal of American Chemical Society*, 121:2337-2338, 1999.

Leach, A., Moelcular Modeling: Principles and Applications, 2nd ed., Prentice Hall, p.513,516, 519, 2001.

Lee, M.M., Chan, M.K., and Bundschuh, R., Simple Is Beautiful: A Straightforward Approach to Improve the Delineation of True and False Positives in PSI-BLAST Searches, *Bioinformatics*, 24:1399-1343, 2008.

Levinthal, C., Are There Pathways for Protein Folding?, *Journal de Chimie Physique*, 65:44-45, 1968.

Levinthal, C., In Debrunner, P., Tsibris, J.C.M., and Munck, E., Mössbauer Spectroscpoy in Biological Systems, Proceedings of a Meeting held at Allerton House, Monticello, Illinois, University of Illinois Press, Urbana, p.22, 1969.

Levitt, M., and Gerstein, M., A Unified Statistical Framework for Sequence Comparison and Structure Comparison, *Proceedings of National Academy of Science USA*, 95: 5913-5920, 1998.

Lim, V.I. and Spirin, A.S., Stereochemical Analysis of Ribosomal Transpeptidation: Conformation of Nascent Peptide, *Journal of Molecular Biology*, 188:565-574, 1986.

Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G., and Chothia, C., SCOP: A Structural Classification of Proteins Database, *Nucleic Acids Research*, 28: 257–259, 2000.

Lo Conte, L., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G., SCOP database in 2002: Refinements Accommodate Structural Genomics, *Nucleic Acids Research*, 30:264–267, 2002.

Louie, A.H., Somorjai, R.L., and Klug, A., Differential Geometry of Proteins: Helical Approximations, *Journal of Molecular Biology*, 168: 143-162, 1983.

Louie, A.H., Somorjai, R.L., Differential Geometry of Proteins: A Structural and Dynamical Representation of Patterns, *Journal of Theoretical Biology*, 98:189-209, 1982.

Lu, H.M., and Liang, J., A Model Study of Protein Nascent Chain and Cotranslational Folding Using Hydrophobic-polar Residues, *Proteins*, 70:442-449, 2008.

Lu, J., and Dahlquist, F.W., Detection and Characterization of an Early Folding Intermediate of T4 Lysozyme Using Pulsed Hydrogen Exchange and Two-dimensional NMR, *Biochemistry*, 31:4749-4756, 1992.

Madej, T., Boguski, M.S., and Bryant, S.H., Threading Analysis Suggests that the Obese Gene Product May Be a Helical Cytokine, *FEBS letters*, 373:13-18, 1995.

Mannhold, R., and Waterbeemd, H.v.d., Substructure and Whole Molecule Approaches for Calculating log*P*, *Journal of Computer Aided Molecular Design*, 15:337-354, 2001.

Marsden, R.L., Lewis, T.A., and Orengo, C.A., Toward a Comprehensive Structural Coverage of Completed Genomics: A Structural Genomics Viewpoint, *BMC Bioinformatics*, 8:86, 2007.

Martz, E., ProteinExplorer: Easy yet Powerful Macromolecular Visualization, *Trends in Biochemical Sciences*, 27:107-109, 2002.

Matthews, B.W., Comparison of the Predicted and Observed Secondary Structure of T4

Phage Lysozyme, *Biochemica et Biophysica Acta*, 405: 442-451, 1975.

McGuffin, L.J., Bryson, K., and Jones, D.T., The PSIPRED Protein Structure Prediction Server, *Bioinformatics*, 16:404-405, 2000.

Miao, X., Waddell, P.J., and Valafar, H., TALI: Local Alignment of Protein Structures Using Backbone Torsion Angles, *Journal of Bioinformatics and Computatioanl Biology*, 6:163-181, 2008.

Michalak, M., Corbett, E.F., Mesaeli, N., Nakamura, K., and Opas, M., Calreticulin: One Protein, One Gene, Many Functions, *Biochemical Journal*, 344:281-292, 1999.

Microsoft Corporation, MS office 2007, Redmond, WA, USA 2007.

Moult, J., Fidelis, K., Zemla, A., and Hubbard, T., Critical Assessment of Methods of Protein Structure Prediection (CASP)-round V, *Proteins*, 53:334-339, 2003.

Murakami, Y., and Mizuguchi, K., Applying the Naïve Bayes Classifier with Kernel Density Estimation to the Prediction of Protein-protein Interaction Sites, *Bioinformatics*, 26:1841-1848, 2010.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C., SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures, *Journal of Molecular Biology*, 247: 536-540, 1995.

Needleman, S.B. and Wunsch, C.D., A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins, *Journal of Molecular Biology*, 48:443-453, 1970.

Nicola, A.V., Chen, W., and Helenius, A., Co-translational Folding of an Alphavirus Capsid Protein in the Cytosol of Living Cells, *Nature Cell Biology*, 1:341-345, 1999.

Norcoss, T. and Yeates, T., A Framework for Describing Topological Frustration in Models of Protein Folding, *Journal of Molecular Biology*, 362:605-621, 2006.

Novotony, J., Bruccoleri, R., and Karplus, M., An Analysis of Incorrectly Folded Protein Models: Implications for Structure Predictions, *Journal of Molecular Biology*, 177:787-818, 1984.

O'Brien, E.P., Christodoulou, J., Vendruscolo, M., and Dobson, C.M., New Scenarios of Protein Folding Can Occur on the Ribosome, *Journal of the American Chemical Society*, 133:513-526, 2011.

Ooi, T., Scott, R.A., Vanderkooi, G. and Scheraga, H.A., Conformational Analysis of Macromolecules IV. Helical Structures of Poly-L-alanine, Poly-L-valine, Poly-β-methyl-L-aspartate, Poly-γ-methyl-L-glutamate, and Poly-L-Tyrosine, *Journal of Chemical Physics*, 46:4410-4426, 1967.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M., CATH-A Hierarchic Classification of Protein Domain Structures, *Structure*, 5:1093-1108, 1997.

Ortiz A., Strauss C.E.M., and Olmea O., MAMMOTH (Matching Molecular Models Obtained from Theory): An Automated Method for Model Comparison, *Protein Science*, 11: 2606-2621, 2002.

Palù, A.D., Dovier, A., and Fogolari, F., Constraint Logic Programming Approach to Protein Structure Prediction, *BMC Bioinformatics*, 5:186, 2004.

Pauling, L., Corey, R.B., and Branson, H.R., The Structure of Proteins: Two Hydrogen-bonded Helical Configurations of the Polypeptide Chain. *Proceedings of National Academy of Science USA*, 37:205-234, 1951.

Pearl, F.M.G., Lee, D., Bray, J.E., Sillitoe, I., Todd, A.E., Harrison, A.P., Thornton, J.M., and Orengo, C.A., Assigning genomic sequences to CATH, *Nucleic Acids Research*, 28:277-282, 2000.

Pedersen, S., *Escherichia coli* Ribosomes Translate in vivo with Variable Rate, *EMBO Journal*, 3:2895-2898, 1984.

Plemper, R.K. and Wolf, D.H., Retrograde Protein Translocation: ERADication of Secretory Proteins in Health and Disease, *Trends Biochem. Sci.*, 24:266-270, 1999.

Ponder, J.W. and Richards, F.M., Tertiary Templates for Proteins: Use of Packing Criteria in the Enumeration of Allowed Sequences for Different Structural Classes, *Journal of Molecular Biology*, 193:775-791, 1987.

Rackovsky, S., and Goldstein, D.A., Protein Comparsion and Classification: A Differential Geometric Approach, *Proceedings of Nataional Academy of Science USA*, 85:777-781, 1988.

Rackovsky, S., and Scheraga, H.A., Differential Geometry and Polymer Conformation. 2. Developlement of a Conformational Distance Function, *Macromolecules*, 13:1440-1453, 1980.

Ramachandran, G.N., Ramakrishnan, C., and Sasiekharan, V., Stereochemistry of Polypeptide Chain Configurations, *Journal of Molecular Biology*, 7:95-99, 1963.

Roder, H., Elöve, G.A., and Englander, S.W., Structural Characterization of Folding Intermediates in Cytochrome C by H-exchange Labelling and Proton NMR, *Nature*, 335:700-704, 1988.

Røgen, P., Evaluating Protein Structure Descriptors and Tuning Gauss Integral Based Descriptors, *Journal of Physics: Condensed Matter*, 17:S1523-S1538, 2005.

Rose, J., and Eisenmenger, F., A Fast Unbiased Comparison of Protein Structures by Means of the Needleman-Wunsch Algorithm, *Journal of Molecular Evolution*, 32:340-354, 1991.

Rost, B., Casadio, R., and Fariselli, P., Refining Neural Network Predictions for Helical Transmembrane Proteins by Dynamic Programming, In Fourth International Conference on Intelligent Systems for Molecular Biology (States, D.J., Agarwal, P., Gaasterland, T., Hunter, L., Smith, R., eds.) AAAI, St. Louis, MO, p. 192-200, 1996.

Rost, B., Liu, J., Nair, R., Wrzeszczynski, K.O., and Orfan, Y., Automatic Prediction of Protein Function, *Cellular and Molecular Life Sciences*, 60:2637-2650, 2003.

Rost, B., Sander, C., and Schneider, R., PHD-An Automatic Mail Server for Protein Secondary Structure Prediction, *Bioinformaitcs*, 10:53-60, 1993.

Rost, B., Schneider, R., and Sander, C., Protein Fold Recognition by Prediction-based Threading, *Journal of Molecular Biology*, 270:471-480, 1997.

Sali, A., 100,000 Protein Structures for the Biologist, *Nature Structural Biology*, 5:1029-1032, 1998.

Sanchez, I.E., Morillas, M., Zobeley, E., Kiefhaber, T., and Glockshuber, R., Fast Folding of the Two-domain Semliki Forest Virus Capsid Protein Explains Cotranslational Proteolytic Activity, *Journal of Molecular Biology*, 338:159-167, 2004.

Saunders, R., Mann M., and Deane, C.M., Signatures of Cotranslational Folding, *Biotechnology Journal*, 6:742-751, 2011.

Sayle, R.A., and Milner-White, E.J., RASMOL: Biomolecular Graphics for All, *Trends in Biochemical Science*, 20:374-376, 1995.

Schonbrun, J., and Dill, K.A., Fast Protein Folding Kinetics, *Proceedings of National*

Academy of Science USA, 100:12678-12682, 2003.

Schwede, T., Kopp, J., Guex, N., and Peitsch, M.C., SWISS-MODEL: An Automated Protein Homology-Modeling Server, *Nucleic Acids Research*, 31:3381-3385, 2003.

Scott, R.A. and Sheraga, H.A., Conformational Analysis of Macromolecules III. Helical Structures of Poly-glycine and Poly-L-alanine, *Journal of Chemical Physics*, 45:2091-2101, 1966.

Seckler, R., Fuchs, A., King, J., and Jaenicke, R., Reconstitution of the Thermostable Trimeric Phage P22 Tailspike Protein from Denatured Chains in vitro, *Journal of Biological Chemistry*, 264:11750-11753, 1989.

Sheraga, H.A., The Role of Tertiary Structure in the Reactions of Several Proteins, Gordon Conference, July, 1957.

Sheridan, R.P., Dixon, J.S., and Venkataraghavan, R., Generating Plausible Protein Folds by Secondary Structure Similarity, *International Journal of Peptide and Protein Research*, 25:132-143, 1985.

Shi, J., Blundell, T.L., and Mizuguchi, K., FUGUE: Sequence-structure Homology Recognition Using Environment-specific Substitution Tables and Structure-dependent Gap Penalties, *Journal of Molecular Biology*, 310:243-257, 2001.

Shindyalov, I.N., and Bourne, P.E., Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path, *Protein Engineering*, 11:739-747, 1998.

Sippl, M.J., and Weitckus, S., Detection of Native-like Models for Amino Acid Sequences of Unkown 3D Structure, *Proteins*, 13:258-271, 1992.

Sippl, M.J., and Wiederstein, M., A Note on Difficult Structure Alignment Problems,

Bioinformatics, 24:426-427, 2008.

Sippl, M.J., On Distance and Similarity in Fold Space. *Bioinformatics*, 24:872-873, 2008.

Sklenar, H., Etchebest, C., and Lavery, R., Describing Protein Structure: A General Algorithm Yeilding Complete Helicoidal Parameters and a Unique Overall Axis, *Proteins*, 6:46-60, 1989.

Skolnick, J., Fetrow, J.S., and Kolinski, A., Structural Genomics and Its Importance for Gene Function Analysis, *Nature Biotechnology*, 18:283-287, 2000.

Skolnick, J., Kolinski, A., and Ortiz, A.R., MONSSTER: A Method for Folding Globular Proteins with a Small Number of Distance Restraints, *Journal of Molecular Biology*, 265:217-241, 1997.

Slater, J.C., Atomic Radii in Crystals, *Journal of Chemical Physics*, 41:3199-3205, 1964.

Smith, T.F., and Waterman, M.S., Identification of Common Molecular Subsequences, *Journal of Molecular Biology*, 147:195-197, 1981.

Söding, J., Protein Homology Detection by HMM-HMM Comparison, *Bioinformatics*, 21:951-960, 2005.

Sørensen, M.A., and Pedersen, S., Absolute in vivo Translation Rates of Individual Codons in *Escherichia coli*: The Two Glutamic Acid Codons GAA and GAG Are Translated with a Threefold Difference in Rate, *Journal of Molecular Biology*, 222:265-280, 1991.

Srinivasan, R. and Rose, G., LINUS: A Hierarchical Procedure to Predict the Fold of a Protein, *Proteins*, 22:81-99, 1995.

Srivastava, S., Patton, Y., Fisher, D.W., and Wood, G.R., Cotranslational Protein Folding and Terminus Hydrophobicity, *Advances in Bioinformatics*, 2011:176813, 2011.

Stivala, A.B., Stuckey, P.J., and Wirth, A.I., Fast and Accurate Protein Substructure Searching with Simulated Annealing and GPUs, *BMC Bioinformatics*, 11:446, 2010.

Summers, N.L., Carlson, W.D., and Karplus, M., Analysis of Side-Chain Orientations in Homologous Proteins, *Journal of Molecular Biology*, 196:175-198, 1987.

Ueda, Y., Taketomi, H., and Nobuhiro, Gō, Studies on Protein Folding, Unfolding, and Fluctuations by Computer Simulation, *International Journal of Peptide Research*, 7:445– 459, 1975.

Varenne, S., Buc, J., Lloubes, R., and Lazdunski, C., Translation is a Non-uniform Process: Effect of tRNA Availability on the Rate of Elongation of Nascent Polypeptide Chains, *Journal of Molecular Biology*, 180:549-576, 1984.

Verlet, L., Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules, *Physical Review*, 159:98-103, 1967.

Voelz, V.A., Shell, M.S., and Dill, K.A., Predicting Peptide Structures in Native Proteins from Physical Simulations of Fragments, *PLoS Computationl Biology*, 5:e1000281, 2009.

Voss, N.R., Gerstein, M., Steitz, T.A., and Moore, P.B., The Geometry of the Ribosomal Polypeptide Exit Tunnel, *Journal of Molecular Biology*, 360:893-906, 2006.

Walker, F.O., Huntington disease, *Lancet*, 369:218-228, 2007.

Walther, D., WebMol: A Java Based PDB Viewer. *Trends in Biochemical Science*, 22:274-275, 1997.

Ward, J.J., McGuffin, L.J., Buxton, B.F., and Jones, D.T., Secondary Structure Prediction with Support Vector Machines, *Bioinformatics*, 19:1650-1655, 2003.

Wei, Q., Wang, L., Wang, Q., Kruger, W.D., and Dunbrack Jr., R.L., Testing Computational Prediction of Missense Mutation Phenotypes: Functional Characterization of 204 Mutations of Human Cystathionine Beta Synthase, *Proteins*, 78:2058-2074, 2010.

Wilmot, C.M. and Thornton, J.M., Analysis and Prediction of the Different Types of β-turn in Proteins, *Journal of Molecular Biology*, 203:221-232, 1998.

Yang, A.S., and Honig, B. An Integrated Approach to the Analysis and Modeling of Protein Sequences and Structures. I. Protein Structural Alignment and a Quantitative Measure for Protein Structural Distance, *Journal of Molecular Biology*, 301:665-678, 2000.

Ye, Y., and Godzik, A., Flexible Structure Alignment by Chaining Aligned Fragment Pairs Allowing Twists, *Bioinformatics*, 19:ii246-ii255, 2003.

Zhang, G., and Ignatova, Z., Folding at the Birth of the Nascent Chain: Coordinating Translation with Co-translational Folding, *Current Opinion in Structural Biology*, 21:25-31, 2011.

Zhang, Y. and Skolnick, J. TM-align: A Protein Structure Alignment Algorithm Based on the TM-score, *Nucleic Acids Research*, 33:2302-2309, 2005.

ABSTRACT (Korean)

구형성과 뒤틀림각에 기반한 단백질 구조 분석 방법론 개발

이름: 정 성 훈
소속: 서울대학교 자연과학대학 협동과정 생물정보학전공

단백질의 구조는 단백질의 기능과 아주 밀접한 관계를 가지고 있다. 이러한 단백질의 구조는 실험적으로 X 선 회절 결정학이나 NMR(핵자기공명) 방법을 통해 구한다. 하지만, X 선 회절 결정학에서는 움직이는 단백질의 구조를 찾아내기가 어렵고 결정을 만들기 힘들다는 단점이 있고, NMR 구조는 막단백질이나 크기가 큰 단백질의 구조를 확인할 수 없다는 단점이 있다. 따라서 이러한 현실적인 문제들을 회피하기 위해 단백질 구조 결정의 이론적 연구가 관심을 받고 있다. 상동성, threading, *ab initio* 방법의 이론적 방법이 세가지 대표적인 방법이다. 자연상태의 안정한 단백질의 구조는 가장 위치에너지가 작은 상태의 구조라고 여겨지고 있다. 단백질 접힘 연구에 있어 가능한 모든 종류의 구조를 탐색할 수 없다는 것은 가장 중요한 문제이다. 이 근본적인 문제는 타당한 강력한 제한조건을 필요로 한다. 따라서 단백질의 구형성(globularity)이 단백질의 일반적인 특성이며 따라서 단백질 구조 모사에 적용할 수 있는 제한조건인지에 대한 조사를 수행하였으며, 이로부터 Gb-index 라는 측정치를 개발하였다. 강한 구형성과 작은 크기와 비구형성의 상관관계가 7131 개의 단백질을 통해 관찰되었다. 이는 생명체가 비가역적 응집을 막기 위해 단백질을 구형으로 접히게 하는 기작을 가지고 있을 수 있다는 것을 암시한다. 이는 또 trinucleotide repeat expansion-mediated disease 들의 발병 기작에 대한 단서를 제공 한다. 실제 환경에서 대부분의 경우 공유결합은 길이가 바뀌지 않는다. 이런 점에 있어서 뒤틀림각계(torsion angle system)은 단백질 구조를 현실적으로

표현하는 데 아주 유용하다. 따라서, 뒤틀림각계(torsion angle system)를 단백질 구조 정렬에 적용하여 타당성을 확인하였다. 62 개의 다른 종류의 protease 의 1891 개 쌍의 단백질에서 단백질의 구조 상동성을 예측하는 실험을 수행했을 때, φ 와 ψ 뒤틀림각(torsion angle)의 1 차원 배열을 가지고 수행한 예측이 3D 구조정보를 바탕으로 정렬을 수행하는 널리 쓰이는 TM-align 이라는 프로그램보다 더 정확하였다. 이 뒤틀림각(torsion angle system)에 기반하여 구조 정렬 어플리케이션 서버와 2 차 구조 서버를 PDB 와 SCOP 정보를 바탕으로 구축하였다. 이 database web application 은 단백질 2 차 구조 검색, 2 차 구조 계산, 그리고 쌍체 단백질 구조 정렬을 수행하는 기능을 가지고 있다. 단백질 접힘 모사과정을 시각적으로 확인할 수 있다는 것은 시각적 표현을 통해 단백질의 상태에 대해 빨리 이해할 수 있다는 점을 고려하면 상당히 흥미로운 일이다. Molecular dynamics 와 같은 기존의 접힘 모사 알고리즘들은 모사 도중에 상태를 확인하고 조작하기가 아주 힘들다. ProtTorter 라는 삼차구조를 시각화하고 위치에너지를 계산하며 등뼈뒤틀림각(backbone torsion angle)을 조작할 수 있는 단백질 구조 모델링 전산 어플리케이션을 개발하였다. 이 개발된 어플리케이션을 이용하여 간단한 새 단백질접힘(protein folding) 알고리즘을 고안하였다. 전사동시적(Co translational)이고 뒤틀림(torsional)적인 접힘 경로를 거치는 Levintahl paradox 의 개념을 따르는 알고리즘을 고안하였다. 이 방법을 작은 peptide 들로 이루어진 실험군(test set)에 적용하여 안정적인 음의 위치에너지와 빠른 수렴이라는 긍정적인 결과를 얻었다. 단백질 구조 정렬을 통해 정당성이 입증된 뒤틀림계(torsional system)와 표면적을 최소화 함으로써 용매와의 상호작용을 반영할 수 있는 구형성 제약조건을 도입하여 ProtTorter 를 사용한 단백질 접힘을 수행하는 것이 앞으로 의미 있는 연구가 될 수 있으리라 본다.

표제어: 단백질 구조, 단백질 접힘, 구조적 구형성, 뒤틀림 각 시스템, Levinthal의 역설, 전사동시성 접힘.

학번: 2008-22789

ACKNOWLEDGEMENT

2004 년 2 월에 낙성대에 자취방을 마련하고 처음 서울 생활을 시작한지 벌써 8 년이 넘었다. 28 년이라는 세월에서 8 년이면 상당히 긴 시간인데도 가장 무덤덤했던 3 분의 1 인 것 같이 느껴진다. 나름대로 공부에 전념했기에 어느새 서른 가까이에 접어드는 것도 느끼지 못할 정도로 바쁘게 생활한 것이 아닌가 한다. 낙성대에는 강감찬 장군의 생가 터와 사당이 있고 관악산은 서애(西厓) 류성룡 선생이 젊은 시절 공부하셨던 곳이다. 강감찬 장군이 송나라 사신에게 문곡성(文曲星)이라고 칭해졌던 것을 생각하면 학업을 닦는 곳으로는 인연이 멀지 않은 곳이다. 학부 때는 대부분 자취방에서 학교까지 걸어서 등교하곤 했는데 주변의 풍경이 아름다웠을 뿐 아니라 전원적이기 마저 해서 기분이 상쾌했었다. 산등성이를 타고 기숙사 삼거리까지 올라가 가파른 벼랑길을 내려오곤 했었는데, 다람쥐를 자주 볼 수 있었고 여름에는 갖가지 곤충과 나비도 볼 수 있었다. 마음이 심란하고 학업에 전념하기 어려울 때는 사당에 가서 향을 피워 올리고 이름을 적고 오기도 했다. 신입생 시절, 영문으로 된 입문서를 붙잡고 겨우 몇 페이지를 넘기려고 끙끙댔었는데 지금은 논문을 읽고 요약하여 자료를 만들며 연구 결과를 영문으로 작성하여 간행하고 있으니 많은 발전을 이룬 셈이다.

이와 같은 발전이 있기까지는 가르쳐 주신 손현석 선생님의 지극한 지도와 도움이 절대적이었다. 부실한 학부수준 실력의 어린애를 당당한 독립적인 연구자로 키워내 주신 선생님께 다시 한번 심심한 감사의 말씀을 올리고 싶다. 아울러 물리학에 대한 신실한 가르침을 해주신 전북대영재교육센터 최종범 선생님과 포항고등학교 하삼수 선생님, 일반물리를 가르쳐주신 서울대학교 김두철 선생님께 깊은 감사를 드리고 싶다. 또 처음 학문의 길에 드는 법을 지도해 주신 덕진초등학교 국종섭 선생님, 과학에 대한 흥미를 일깨워 주신 이종섭 선생님, 기초를 튼튼히 해주신 김영애 선생님께 감사 드린다. 전라중학교 신선운 선생님, 과학경시대회를 지도해 주신 이복순 선생님과 이현아 선생님 그리고 수학의 기초를 잘 세워주신 정광수 선생님께도 깊은 감사의 말씀을 드리고 싶다. 포항에 유학 온 전주학생을 잘 돌봐주시고 키워주신 포항고등학교 송창윤 선생님과 생물학을

가르쳐주신 김진성 선생님, 이해와 격려를 아끼지 않아 주신 손승태 선생님께도 깊은 감사의 말씀을 올리고 싶다. 서울대학교 자연과학대학 생명과학부의 제 교수님께도 훌륭한 강의로 생물학 실력을 쌓아 올리게 해 주신데 대해 깊이 감사드린다. 아울러 훌륭한 지도와 조언을 아끼지 않아주신 노경태교수님, 윤창노박사님, 성제경교수님, 김희발교수님께 깊은 감사의 말씀을 올리고자 한다. 여러모로 격려를 아끼지 않아 주신 윤창노박사님께는 부족하고 모자라는 심사지원자의 장점과 가능성을 살펴주신 것에 대한 특별한 감사의 말씀을 올리고자 한다.

실험실 맏언니로서 사려 깊게 보살펴 주신 선배 안인성 박사님께 감사의 말씀을 전하며, 연구에 큰 도움과 지도를 해주신 선배 배세은 박사님께도 깊은 감사의 말씀을 전하고 싶다. 그 밖에 연구실의 선배님들과 후배님들(김보란님, 김하연님, 이지혜님, 장진화님, 제미경님, 조광훈님, 송연정님, 김정훈님, 민해숙님, 박해일님, 김미란님, 이영미님, 윤재문님, 유태곤님, 황지선님)께도 감사 드린다. 여러모로 도움을 받고 조언을 구한 것에 대해 감사하는 마음과 아울러, 한편으로는 선배로서 후배님들께 많은 것을 도와주지 못한 점이 아쉽기만 하다.

전라중학교 동창으로 같은 학교에 다니면서도 자주 만나지 못한 류충석 학우에게 미안한 마음을 전하고 싶으며, 선배 윤석준님과 다른 여러 선배님께도 안부를 전해드리고 싶다. 항상 고향에서 격려해준 황병훈 군에게도 깊은 감사의 마음을 표하고 싶다. 항상 격려를 아끼지 않아준 오현진 양에게 깊은 감사의 마음을 전하며 학부 동기 문형민 양에게도 여러모로 감사의 마음을 표하고자 한다.

그리고 마지막으로, 낳아 주시고 길러 주심과 아울러 항상 걱정 근심으로 아들의 오늘이 있기까지 격려해 주신 고향의 아버님과 어머님께 깊은 감사의 마음을 전해 드린다. 항상 건강하시고 행복하세요.