A Methodology in Predicting Protein Tertiary Structure

A Thesis

presented to the Department of Computer Science

of The Chinese University of Hong Kong

in partial fulfilment of the requirements

for the Degree of Master of Philosophy

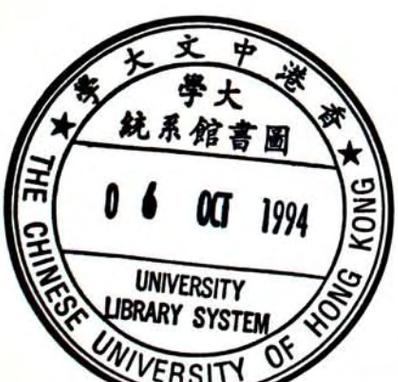
by

Li Leung Wah

February 1993

Acknowledgements

I wish to express my thanks to my supervisor, Dr. Y. S. Moon. Not only does he teach me how to do research but also advises me how to present my thesis. Furthermore, I would also like to thank Dr. H. W. Yeung; he provides much information about the structure of protein to me. Finally, I add special thanks to my family members; they not only encourage me but also provide me financial support during my study.



Abstract :

Due to the contribution of protein modeling, many novel protein molecules and reactants are created. These new products are invented by trial and error. However, this is not an effective and efficient method. In biochemical theory, the function of a protein molecule is dictated by its three dimensional structure which, in turn, is controlled by its amino acid sequence. Useful and practical protein molecules can be invented if the relationship between a protein sequence and its three dimensional structure is discovered. There are many methods trying to predict the three dimensional structure of protein molecules. Nevertheless, the results of prediction are not good enough. A new methodology, based on geometry and the characteristics of the structure of an amino acid, is proposed. In protein secondary structure prediction, the results of prediction of 16 protein molecules are slightly better than other methods. In protein tertiary structure prediction, a reasonably good result has been obtained for a tricosanthin molecule. Moreover, the tertiary structure of a protein molecule having about 200 amino acids can be obtained within 10 minutes. It is a faster and cheaper method, compared with the energy minimization method.

Table of Contents **Page**

Chapter 1

Protein Modeling

In recent years, biotechnology becomes a hot topic all over the world. Twenty years ago, useful enzymes and vaccines were either produced by chemical synthesis or extracted from living organisms. The yield was low and the cost of production was very expensive. Moreover, many side effects were induced when the patients used these materials because of the existence of impurities. Fortunately, the world has changed. The impossible becomes possible. We can easily obtain highly purified enzymes by biochemical methods. Furthermore, biochemical products are widely used in different areas. In the military, there are many chemical and biological weapons. In industry, different kinds of enzymes are used to improve productivity. Moreover, in environmental protection, bacteria can be used to degrade useless materials into useful elements or to extract useful materials from waste. It shows that biochemical products are not only valuable but also profitable. Therefore, many countries put much effort to develop biotechnology.

1.1 Genetic Engineering

There are two areas in biotechnology. The first one is "Genetic Engineering". In this aspect, information on genetics is used to produce useful matter. Insulin is one of the successful substances produced by this method. The

life cycle of bacteria is very short, a generation is about 20 minutes. Hence, in several hours, a large amount of bacteria can be produced if suitable nutriment is provided. This powerful reproduction of bacteria is utilized by scientists. They implant the gene of insulin in bacteria. A great deal of insulin can be made within several hours.

1.2 Protein Engineering

The objective of protein engineering is to modify the function of protein molecules so that the novel protein molecules can be used to improve human life.

1.2.1 The Basic Concept

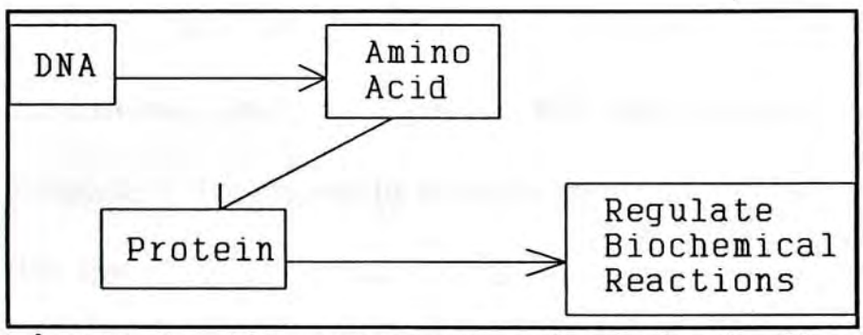


Figure 1. The sequence of control flow from DNA to biochemical reactions.

Inherent information in living organisms is carried by DNA. Three consecutive DNA bases dictate the formation of an amino acid. An amino acid chain forms a protein molecule. Protein molecules in living organisms regulate the biochemical reactions so that living organisms can function normally.

1.2.2 The importance of protein modeling

Protein modeling is another branch of protein engineering. The purpose of protein engineering is to investigate the relationship between structure and function of protein molecules. The chemical reaction between a protein molecule and its reactant is quite different from general chemical reactions. Although a protein molecule is large in size, the reaction only occurs in a specific site which is called active site. Furthermore, the chemical reaction only occurs when the structure between a protein molecule and its reactant is complementary. Hence, due to this specific selectivity, the catalytic performance of protein molecules is more efficient than other chemical catalysts.

In fact, the function of a protein molecule is dictated by its three dimensional structure. However, the three dimensional structure of a protein molecule is determined by its amino acid sequence. From figure 1, it is easily seen that modifying either the structure of a DNA or an amino acid can change the structure of a protein molecule.

There are many advantages if the secrets between the structure and function of protein molecules are discovered.

1. The function of protein molecules can be estimated while their structures are found.
2. The structure of protein molecules can be predicted if their functions are known.
3. Novel protein molecules with special functions can be created as you like.

1.2.3 Applications

Protein engineering, in fact, is employed in different aspects. There is only a brief description of its application in industry and medicine.

1.2.3.1 Industry

Subtilisin is one of the good examples to elucidate the success of protein engineering in industry. It is significant that subtilisin represents the largest industrial enzyme market, primarily as an additive in laundry detergents[73].

Mutants of subtilisin are created in order to investigate the changes of (i) the chemical activities, (ii) the thermal stability and (iii) the resistance to oxidizing agents. Different mutants are invented by replacing only one amino acid from the native subtilisin. As a result, mutants of subtilisin having different characteristics are formed. Some mutants still exert chemical activities in a high temperature environment. Although some mutants would not collapse in strong oxidizing agents, they would not exert chemical activities. Therefore, due to these special characteristics, suitable mutants can be chosen in order to improve productivity at different situations.

1.2.3.2 Medicine

Angiotensin-Converting Enzyme (ACE) is another good example to show the achievement of protein engineering. In biochemical reactions, reactants have intention to join together if their structures are complementary to each other. According to this phenomenon, a strong inhibitor, namely, captopril, is designed to

inhibit the function of ACE [58]. When ACE captures captopril, ACE cannot digest captopril. The normal function of ACE, increases the blood pressure, cannot be exerted. Hence, captopril is good news to hypertensive people.

1.3 The structure of protein molecules

Before we discuss the predicting algorithms in protein structure, the structure of protein molecules should be first introduced.

The basic building block of a protein molecule is amino acid. An amino acid consists of an amino group, a carboxyl group, a hydrogen atom and a distinctive group of atoms which is referred as a side chain group "R". The amino group of an amino acid attaches to the carboxyl group of another amino acid to form a peptide bond. Many amino acids, usually more than a hundred, are joined by peptide bond to form a chain, which is an unbranched structure. A protein chain consists of a regularly repeating part, called the main chain, and a variable part, comprising the distinctive side chains.

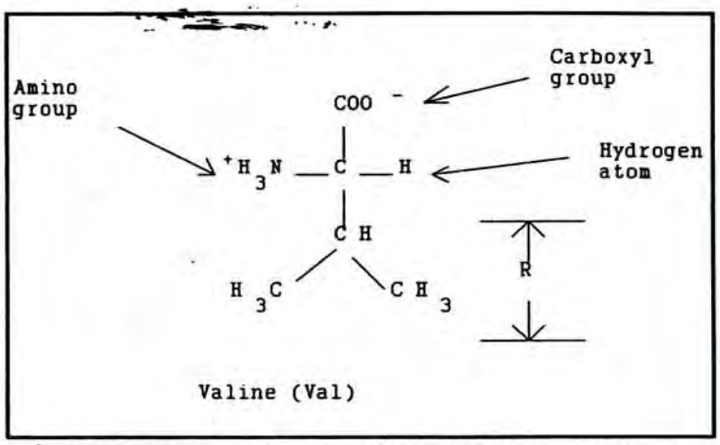


Figure 2. The structure of an amino acid -- Valine.

In protein architecture, it is convenient to refer to four levels of structure. The primary structure is simply the sequence of amino acid. The secondary structure refers to locally well-defined periodic structures, such as α-helix, ß-sheet

Lys-Glu-Thr-Ala-Ala-Ala-Lys-Phe-Glu-Arg-Gln-His-Met-Asp-Ser-Ser-Thr	17
Ser-Ala-Ala-Ser-Ser-Ser-Asn-Tyr-Cys-Asn-Gln-Met-Met-Lys-Ser-Arg-Asn	34
Leu-Thr-Lys-Asp-Arg-Cys-Lys-Pro-Val-Asn-Thr-Phe-Val-His-Glu-Ser-Leu	51
Ala-Asp-Val-Gln-Ala-Val-Cys-Ser-Gln-Lys-Asn-Val-Ala-Cys-Lys-Asn-Gly	68
Gln-Thr-Asn-Cys-Tyr-Gln-Ser-Tyr-Ser-Thr-Met-Ser-Ile-Thr-Asp-Cys-Arg	85
Glu-Thr-Gly-Ser-Ser-Lys-Tyr-Pro-Asn-Cys-Ala-Tyr-Lys-Thr-Thr-Gln-Ala	102
Asn-Lys-His-Ile-Ile-Val-Ala-Cys-Glu-Gly-Asn-Pro-Tyr-Val-Pro-Val-His	119
Phe-Asp-Ala-Ser-Val	124

Figure 3. The primary structure of a protein molecule (bovine ribonuclease).

and ß-turn. The tertiary structure mentions the overall steric arrangement of amino acids in protein molecules. Protein molecules, in nature, would comprise one or more amino acid chains. The three dimensional structure of one single amino acid chain is called a "subunit". Protein molecules containing more than one amino acid chain display an additional level of structural organization, namely, quaternary structure which refers to the way in which the chains are packed together. If a protein molecule is only made of one amino acid chain, this protein molecule would not have quaternary structure.

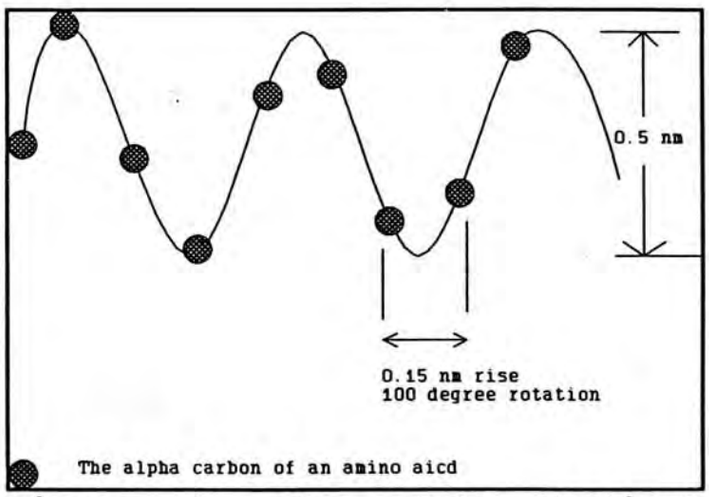


Figure 4. The alpha helical structure of a protein fragment.

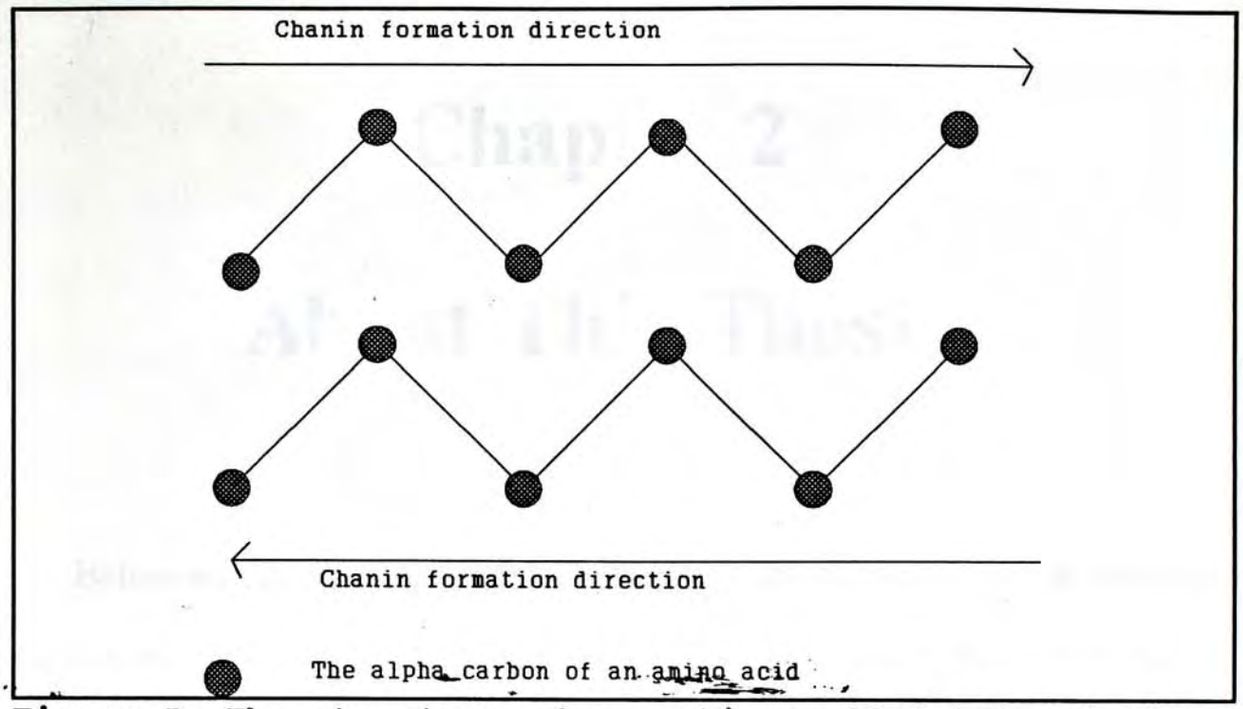


Figure 5. The structure of an anti-parallel beta sheet.

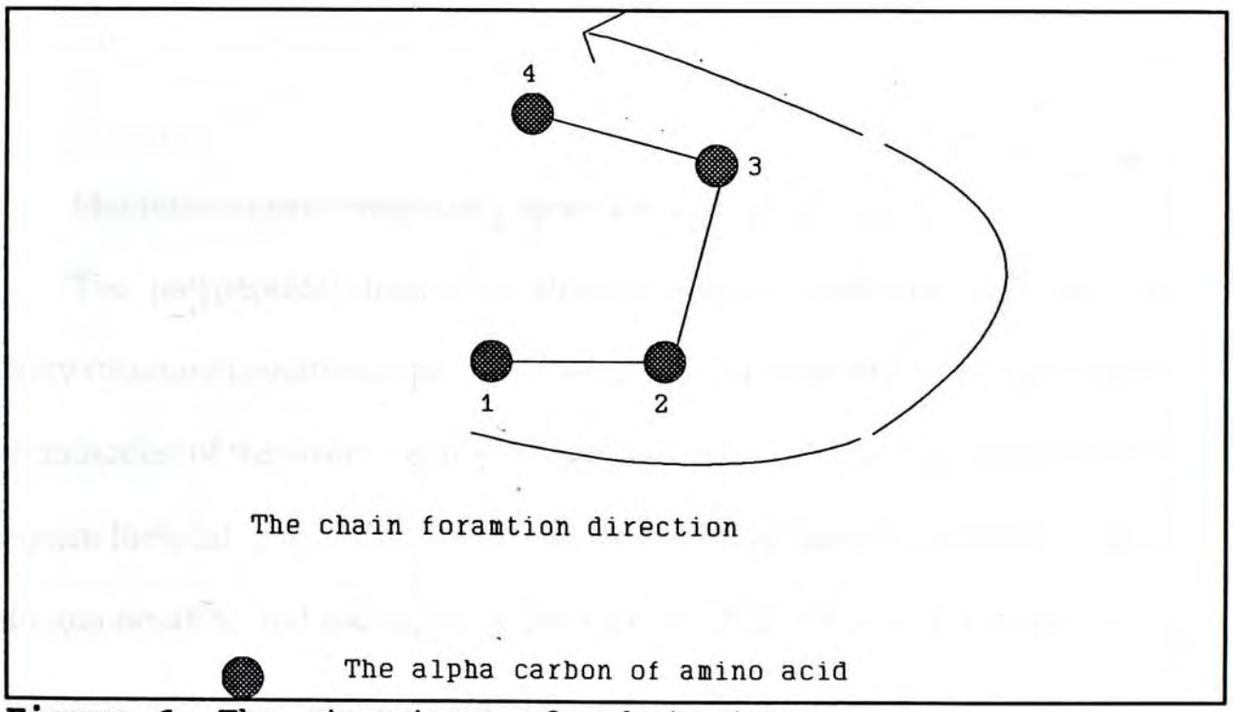


Figure 6. The structure of a beta turn.

Chapter 2

About This Thesis

Before we discuss the objective of this thesis, the current predictive methods in protein tertiary structure are first introduced. Afterwards, the techniques in artificial intelligence and computer graphics are described. Furthermore, there are explanations stating why artificial intelligence and computer graphics are especially suitable for protein modeling.

2.1 Methods on protein tertiary structure prediction

The polypeptide chain of a globular protein molecule is linear, but its tertiary structure is quite compact. This was apparent from the first crystallographic determination of the structure of a protein molecule by Kendrew *et al.* in 1960. The compact form of protein molecules need to satisfy many interactions, such as hydrogen-bonding and hydrophobic interactions. It is difficult to predict the three dimensional structure of a protein molecule since many factors are involved. However, still many predictive methods are proposed, namely, energy minimization, sequence homology and hierarchical assembly method.

2.1.1 Energy minimization method [27]

Free energy is a measurable quantity to represent the state of a molecule. The higher the free energy of a molecule owns, the higher the reaction activity of a molecule possesses. A molecule would occur in different conformations with different free energy values. Among these states, a molecule with the lowest free energy value would have a most stable form. There are two general methods in energy minimization.

a. The first method involves minimization of the overall energy by adjustment of conformational parameters. It can be applied to the full covalent structure of a protein molecule or to idealized representations using modified potential functions. [27]

b. The second one is based on calculations of molecular dynamics. In a definite time domain, this method simulates the motion of a molecule.

Both procedures follow physically realizable paths according to small incremental changes in structure. At the present stage of development, neither approach has been able to fold an extended chain into its known compact native conformation without using constrained potential functions.

2.1.2 Sequence homology method [57]

The assumption of sequence homology is quite simple. The amino acid sequence of a protein molecule determines the higher order structures of a protein molecule. Protein molecules with similar primary structure would have similar

tertiary structure. To find the tertiary structure of a protein molecule, it is sufficient to find a primary structure of a protein molecule. A database containing protein primary structures and their related tertiary structures must be first developed before using the sequence homology method.

There are two approaches in the sequence homology method. First, a protein molecule, let say "protein A", with known primary structure but unknown tertiary structure is compared with all the amino acid sequences in a database. A protein molecule in a database with primary structure most similar to "protein A", let say "protein B", is chosen. It is because the higher the similarity in primary structure, the higher is the similarity in tertiary structure. Therefore, the tertiary structure of "protein A" can be determined by slightly modifying the tertiary structure of "protein B". Nevertheless, there are not enough protein molecules whose tertiary structures are known. Therefore, it is hard to find a protein molecule with primary structure highly similar to an unknown protein molecule.

An alternative method is to use a statistical approach to find the protein tertiary structure. Parts of an amino acid sequence, let say "Protein C" are compared with records in a database so as to find the most similar protein segments. The similar segments obtained from comparison may come from different protein molecules. These segments with their corresponding tertiary structures are then assembled into the final tertiary structure of "Protein C". However, the results of this method are often not acceptable.

2.1.3 Hierarchical assembly method [24]

In contrast to these methods, another approach, called hierarchical assembly method, disregards any explicit consideration of energy terms and concentrates solely on probable geometry, based on examples from the known structures of protein molecules. This method is not related to the physically realizable process. It is only a heuristic procedure dependent heavily on analyses of the known protein structures. This method is conveniently divided into three stages :

a. Determine the secondary structures of a protein molecule from its primary structure, e.g. Chou and Fasman [7,8,9], Lim [48,49], Garnier [32,33].

b. Determine an approximate tertiary fold by packing the secondary structures, e.g. Ptitsyn and Rashin, Cohen [24].

c. Calculate the native conformation by refining the tertiary fold.

The predictive rules are obtained from a cluster of protein molecules which are not related to each other. This is the advantage of this predicting method. However, the results of prediction are not good enough. It is because this kind of method involves three steps. Error occurring in any one of these steps would cause incorrect results of prediction. In protein secondary structure prediction, the best result is not more than 70% [24]. Thus, low predicting results are unavoidable.

2.2 Artificial intelligence and molecular modeling

Solving general problems instead of traditional algorithmic problems is the aim of "Artificial Intelligence". The main difference between artificial intelligence

and arithmetic processing is symbol manipulation. Artificial intelligence can manipulate different symbols but arithmetic algorithms cannot. Many artificial intelligent programs including heuristic rules instead of algorithms can solve some problems which have not definite answer. These problems must be solved by previous experiences or practices.

The structure of a protein molecule is very complicated. A protein molecule always consists of more than a hundred of atoms. Atoms in a protein molecule forms a special structure which provides a definite biochemical function. Many chemical restraints govern the formation of a protein molecule. These restraints are:

a. distance between two atoms
b. rotation angle between two atoms
c. dipole-dipole interaction
d. Van der Waals force interaction
e. overall free energy

Therefore, it seems difficult to predict the structure of a protein molecule using traditional arithmetic algorithms. However, the technique "Constraint" [68,69] in artificial intelligence is especially suitable for protein modeling. Constraints can be formulated into rules which are arranged in a descending order of their importance. Rules are selected from this ordered list. The most important one is first used to see whether it is suitable for solving the problem or not.

2.3 Computer graphics and molecule display [28]

The radius of an atom is about 10^{-10}m. This size cannot be visualized by electronic microscope, let alone the human eyes. How can scientists investigate molecules without seeing them? A molecule is composed of atoms and chemical bondings. Traditionally, a molecule is represented by the ball-and-stick model. Atoms are represented by balls and chemical bondings are represented by sticks. It is time-consuming to construct a physical model of a molecule by hand. Furthermore, it is difficult to investigate a chemical reaction between two molecules using a rigid molecular model.

Fortunately, techniques in computer graphics can solve the above-mentioned problems. The graphic image of a molecule can be displayed on a screen in a few minutes. This whole graphic image can be manipulated by the user. He/she can translate, rotate the whole object, enlarge or reduce its size. Furthermore, he/she can change the subtended angle between two chemical bondings or the relative positions of atoms. Moreover, the process of molecular reactions can be displayed on a screen by the animation techniques. The information on the change of reaction energy, the change of the shape of molecules or the possible products of a reaction can be collected by scientists. Hence, computer graphics is a useful tool to investigate molecular structure.

2.3.1 Molecular model in computer graphics

Electrons of an atom move around the atomic nucleus. In Bohr's model, only the electronic density of an atom can be estimated. The exact position of an

electron cannot be determined. Hence, no one knows the "real" picture of an atom.

Use of computer graphics here is, not to display the actual picture of an atom, but to describe the characteristics of a molecule, namely, electronic density, bond angles, relative positions of atoms and so on. However, a picture cannot show all the characteristics of a molecule. Different kinds of pictures depict different kinds of molecular properties. Therefore, different molecular models are established to reveal different kinds of characteristics of a molecule.

In general, there are four molecular models, namely, ball-and-stick, isoelectric surface, vector and amino acid residue.

a. Ball-and-Stick model

In the ball-and-stick representation, balls with different sizes and different colours represent different atoms. The bond length, relative positions of atoms and the overall steric structure of a molecule can be captured by scientists. Scientists would feel comfortable since this picture looks like the physical model of a molecule. However, the graphic image on a screen is so complicated that this picture is not suitable for interactive manipulation.

b. Isoelectric surface model

The electronic density of an atom is the main theme to be displayed in this model. The electronic density is represented by combinations of dots.

Hence, a surface rather than a sphere is shown. The electronic density of atoms are additive, therefore, the overall electronic surface of a molecule can be displayed. Some special structures such as groove, cleft, hole or channel can be distinguished. The reaction sites of two molecules are always identified by this method.

c. *Vector model*

In this kind of picture, there are only lines on a screen. A line represents covalent bond and the intersection point between two lines represents an atom. The information provided by this model is inadequate. It only describes the rough overall structure of a molecule. Nevertheless, the graphic image is quite simple, it is especially suitable for interactive manipulation and animation.

d. *Amino acid residue model*

In protein modeling, the secondary structures of a protein molecule are important parts to be investigated. It is difficult to extract the information of a protein secondary structure from previous methods. It is because the vector model is too simple but the ball-and-stick model is too complicated. The amino acid residue representation is the best solution. This model is a compromise between the characteristics of the ball-and-stick model and the vector model. The ball in the amino acid residue model represents an amino acid residue rather than an atom. The stick represents the linkage between two amino acid residues rather than the covalent bond.

The whole picture of a protein molecule is simplified. The skeleton, especially the secondary structures, of a protein molecule is revealed.

2.3.2 Interactive graphic operations

An object with three-dimensional (3D) structure cannot be displayed on a screen comprehensively. It is because a screen only has a two-dimensional display surface. A 3D object, such as a molecule, must be transformed into a 2D view when displayed on a screen. In mathematics, there are many functions to map coordinates from three-dimensional into two-dimensional. Some characteristics of a 3D object are lost when it is compressed into a 2D view. Part of this object must be obscured by the other parts of this object. There are operations that can be used to manipulate objects so that all the properties of objects can be displayed on a screen. The operations are : translation, rotation, scaling, zooming in and zooming out.

The height and width of a screen are fixed. If a molecule is too large to be displayed on a screen, part of this molecule is missing on a screen. Fortunately, a molecule can be moved along X- or Y- direction. Similarly, a molecule can be rotated about X-, Y- and Z-axis. The hidden part of a molecule can be moved to a suitable position so that it can be displayed on a screen. It is very easy to operate the rotation and translation function by specifying the axis and the value to be rotated or translated respectively.

However, translation and rotation cannot alter the size of a molecule. There must exist an operation to change the size of an object. Scaling is a function that can enlarge or reduce the size of an object.

In research, scientists are not always interested in the whole molecule. He/she may only be concerned with part of a molecule such as the reaction site. This reaction site is possibly composed of only several atoms. Hence, a zoom-in operation must be provided. He/she can select a particular area in a screen and then magnify it. The advantage is that the user can concentrate to study the desired part that they are interested in. Similarly, a zoom-out command can be used to display the whole molecule on a screen.

2.4 The objective of this thesis

The best way to describe the structural information of a protein molecule is the three dimensional coordinates of its atoms. During the past ten years, many predicting algorithms in protein tertiary structure were proposed and implemented. However, their predictive results were not acceptable. The most successful one was 70% accurate. Furthermore, these algorithms used a lot of computer time, say, several hours of CPU time on supercomputer.

The objective of this project is to predict the tertiary structure of protein molecules in an improved way and to develop a fast method for the prediction.

After studying the hierarchical assembly, energy minimization, sequence homology methods for protein tertiary structure prediction, we have decided to adopt the idea of the hierarchical assembly method. The reasons are :

i The concept of the hierarchical assembly method is similar to the protein folding mechanism. In protein folding, secondary structure fragments are first formed. Afterwards, these fragments are joined together and finally a compact form of a protein molecule is developed.

ii There are many forces to stabilize a protein tertiary structure in nature. However, it is difficult to formulate the forces. Moreover, there are many unknown factors affecting the protein tertiary structure. In addition, the energy minimization method is quite an expensive method since large amount of CPU time is involved.

iii In the sequence homology method, a database containing the primary and tertiary structures of protein molecules must be set up first. It is because our research group has no information on the tertiary structure of protein molecules, the sequence homology method is not considered.

Our methodology in protein tertiary structure prediction is composed of two steps. The hydrophobicity and the secondary structures of a protein molecule are first determined. Then, the compact form of a protein molecule is assembled by the protein secondary structures. The algorithm in hydrophobicity determination and

protein secondary structure prediction are described in detail in Chapter 3. The methodology that assembles a protein molecule from its secondary structures to its tertiary structure is depicted in Chapter 4. In Chapter 5, the equipment, special features, memory and data structures used by our algorithms are fully discussed. The results of our predictions are reported in Chapter 6. Furthermore, the conclusion, advantages, disadvantages and further development are discussed on Chapter 7.

Chapter 3

Algorithms For Protein Secondary Structure Prediction

3.1 Hydrophobicity

A protein molecule is composed of amino acids. The properties of an amino acid would directly affect the characteristics of a protein molecule. Therefore, investigating the properties of an amino acid is an important step to estimate the tertiary structure of a protein molecule.

Solubility is one of the properties of an amino acid. An amino acid has different solubility in different solvents. In chemistry, there is an empirical rule called "like dissolves like". It means that molecules which have similar properties dissolve in each other. For example, oily materials dissolve in carbon tetrachloride (CCl_4) but form an oily layer in water.

A quantity which measures the solubility of a substance in polar solvent (such as water) is called hydrophobic value. The word "hydrophobic" means "hate

water". An amino acid which has a high hydrophobic value indicates that it has a low solubility in polar solvent.

Based on solubility, protein molecules can be classified into two categories: globular protein and membrane protein molecules. Globular protein molecules usually dissolve in polar solvent whereas membrane protein molecules dissolve in non-polar solvent. The environment of globular protein molecules is polar. Amino acids which possess high hydrophobic values are surrounded by other amino acids with low hydrophobic values. However, the situation is reversed in membrane protein molecules. The environment of membrane protein molecules is non-polar. Amino acids with high hydrophobic values are located on the surface of membrane protein molecules. The distribution of amino acids in such a way maintains the protein structure and minimizes the unstable factors.

The primary structure of a protein molecule is linear. Amino acids are linked one by one. The hydrophobic value of an amino acid, in fact, is affected by other amino acids near it. The hydrophobic value of each amino acid is determined by chemical experiments. However, the hydrophobic value of an amino acid in a protein chain is calculated by the following equation [24].

$$H_i = \frac{\sum_{j=i-n}^{i+n} H_j}{2n+1} \qquad (1)$$

The hydrophobic value of the *i th* amino acid is equal to the mean value of hydrophobic value from *i-n* th to *i+n th* amino acids, where n is an arbitrary value. The value *(2n+1)* is called window size. The hydrophobic value of an amino acid in a protein chain changes if the size of window varies. According to different experiments, better results of prediction are obtained when the window size is 5, 7 or 9 [24,36].

Three protein molecules that have membrane affiliation are illustrated in Figure 7. In the upper panel, amino acids from 75-94 are recognized as membrane-spanning segment. Similarly, amino acids from 110-130 in the middle panel and amino acids from 470-490 in the lower panel are also clearly classified as membrane-spanning fragment. From the results in Figure 7, the location of amino acids in a protein molecule can roughly be estimated by the hydrophobic values.

3.2 Algorithms for protein secondary structure prediction

In protein molecules, groups of amino acids would form a locally well-defined secondary structure such as helix or sheet. In the folding process of a protein molecule, these special structures are first formed in order to reduce the free energy of a protein molecule. Then, the protein tertiary structure is assembled by these secondary structures. Therefore, before developing a method in protein tertiary structure prediction, an algorithm for protein secondary structure prediction must be created first.

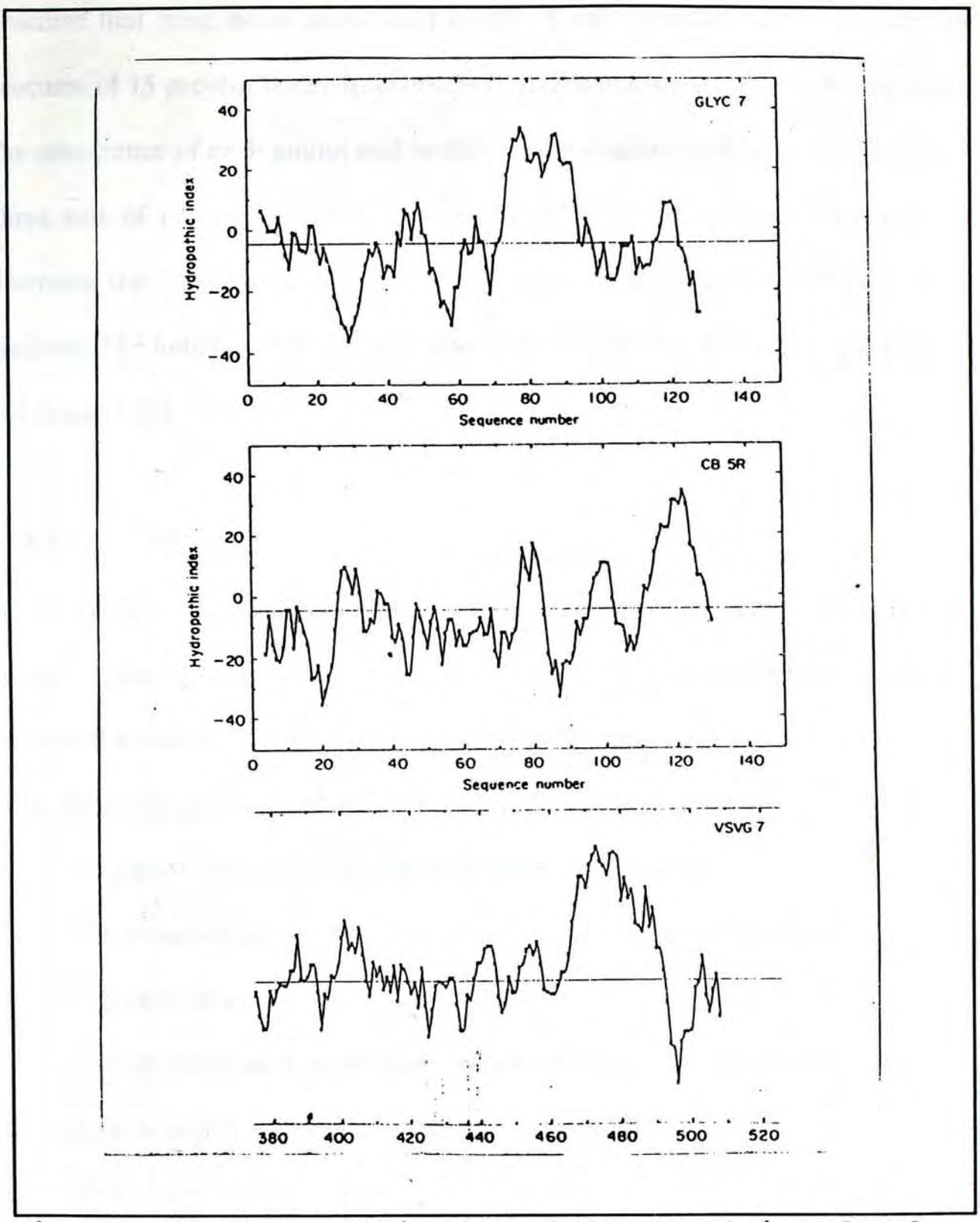


Figure 7. The hydrophobic plot of three protein molecules. The window size is 7.

3.2.1 The Chou and Fasman Method

The most famous method in protein secondary structure prediction is proposed by Chou and Fasman. Their method is based on the protein tertiary

structure that have been determined by the X-ray crystallography. The tertiary structure of 15 protein molecules with 2473 amino acids were carefully examined. The occurrence of each amino acid in different secondary structures was recorded. Three sets of parameters were deduced by statistical method. Furthermore, by observing the characteristics of different secondary structures, some rules were finalized. The helical, sheet and turn structures were predicted by these parameters and rules [7,24].

3.2.1.1 Method

Amino acids in helical and sheet structure can be divided into six groups, strong former, former, weak former, indifferent former, weak breaker and strong breaker. Two weak formers were treated as one former.

Rules in helical structure prediction

1. At least 4 helical formers exist within 6 amino acids.
2. According to the helical parameters, the mean value of these 6 amino acids is greater than or equal to 1.03.
3. Helical structure is terminated by helical breaker or strong helical breaker.
4. Amino acid "Proline" cannot occur in helical structure.

Rules in sheet structure prediction

1. At least 3 sheet formers exist within 5 amino acids.
2. According to the sheet parameters, the mean value of these 5 amino acids is greater than or equal to 1.00.

Rules in turn structure prediction

1. According to the turn parameters, the product of 4 consecutive amino acids is greater than $0.75 * 10^{-4}$.

(Algorithm in Appendix B)

3.2.1.2 Results

Chou and Fasman claimed that their prediction method had 80%, 86% and 90% accuracy in helical, sheet and turn structures respectively. However, the results of the computerized Chou and Fasman method was not as high as that described by Chou and Fasman. The overall results of prediction from computer program was only 55% in accuracy [56].

In Chou and Fasman method, an amino acid in a protein molecule might be predicted as helix, sheet and turn independently. However, an amino acid in a protein molecule could only belong to one of the secondary structures. If the state of an amino acid in a protein molecule is predicted in more than one secondary structures, this phenomenon is called "regional overlapping problem". In Chou and Fasman method, they solved this problem by their knowledge and experience. However, these knowledge and experience were not formulated into rules. Hence, their results of prediction was better than that of their computerized method. Therefore, Nishikawa regarded that the Chou and Fasman method was qualitative rather than quantitative [56].

3.2.2 The GOR method

Garnier and his colleagues developed another prediction method using information theory which was based on probability considerations. This method is called GOR method for the sake of convenience.

3.2.2.1 Theory

Each amino acid has different probabilities to occur in different secondary structures. However, an amino acid in a definite position of a protein molecule should belong only to a particular secondary structure. The probability of an amino acid that belonging to different secondary structures were calculated. For example, the probabilities of an alanine occurring in helical, sheet and turn structures of a particular molecule were predicted as 0.6, 0.3 and 0.1 respectively. Thus, in the GOR method, this alanine would be predicted to occur in a helical structure.

3.2.2.2 Method and results

Garnier and his colleagues believed that the state of an amino acid was affected by other amino acids near it. (Just like the hydrophobic value of an amino acid in a protein sequence). For each amino acid, they defined that 16 amino acids occurring on its two sides might be considered simultaneously. The occurring probabilities of each amino acid in different secondary structures were derived from 26 protein molecules with about 4500 amino acids [33]. (Algorithm in Appendix C)

The results of the GOR prediction method is 63% [56]. Simple theoretical basis and easy implementation on computer are the advantages of the GOR method. Moreover, the individual results of the GOR method do not conflict with each other. This is the great difference between the Chou and Fasman method and the GOR method. In real life, the helical structures of a protein molecule are always composed of at least 4 amino acids. Turn and sheet structures are usually composed of at least 3 amino acids. However, the results of the GOR method do not always match these phenomena. This is the major weakness of the GOR method.

Lim did not use statistical approach to predict the protein secondary structures. He, based on the physio-chemical properties of amino acids, proposed another kind of prediction method. The secondary structures of a protein molecule were predicted according to the charge, size, shape and other properties of amino acids. He announced that the results of his method in helical and sheet structures were 80% and 85% respectively [48,49].

Nishikawa in detail analyzed the regional overlapping problem of the Chou and Fasman method. He described four types of regional overlap. He suggested that when a regional overlap occurs, we should not only concern the overlapping region, but the whole structure of a helix and a sheet in the overlapping region should be considered.

For example, the type II in Figure 8 indicates that the entire region from a to d must be carefully examined rather than the region from b to c.

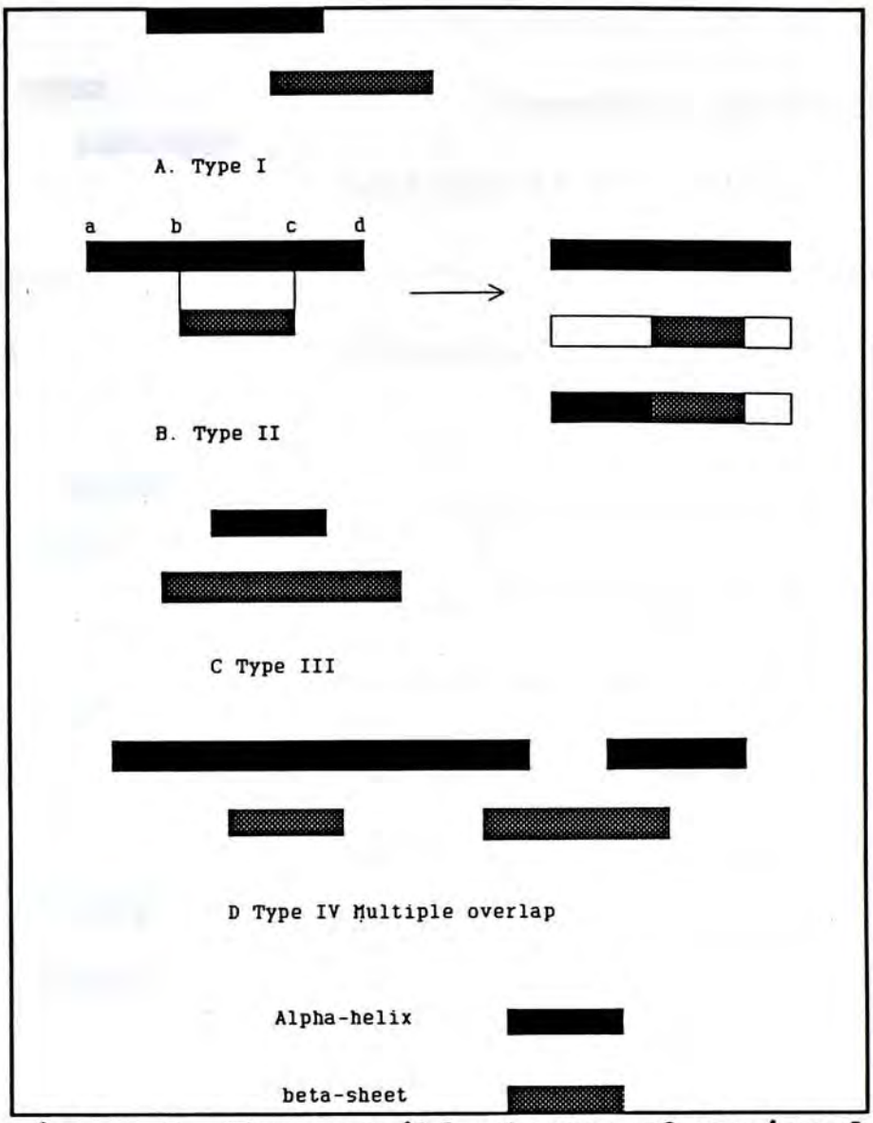


Figure 8. The possible types of regional overlap occur between helical and sheet region.

3.3 A proposed algorithm

After studying these prediction methods, some rules are formulated to predict the protein secondary structures.

First, the Chou and Fasman method is used to predict protein secondary structure individually. The regions predicted as turn structure are selected regardless of regional overlapping problem. Second, our algorithm would select the

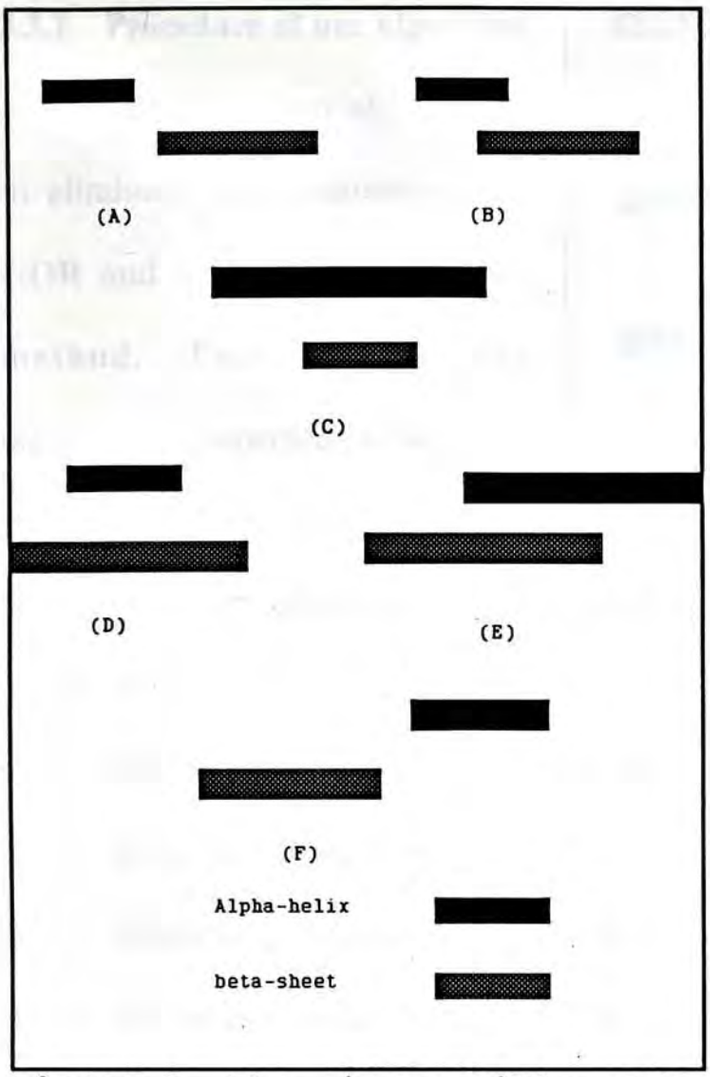


Figure 9. The six possible cases that two regions are overlapped with each other.

secondary structure which has a higher independent predictive value than the other structures. (The cases of regional overlap are illustrated in Figure. 9)

Third, a function is created to join two helical or sheet structures when they are separated by a random coil structure. The joined structure must obey the prediction rules in Chou and Fasman method.

Finally, a helical fragment which has no more than 4 amino acids and a sheet structure which has less than 3 amino acids are predicted as random coil structure. This rule is to eliminate the fatal weakness of the GOR method.

3.3.1 Procedure of our algorithm

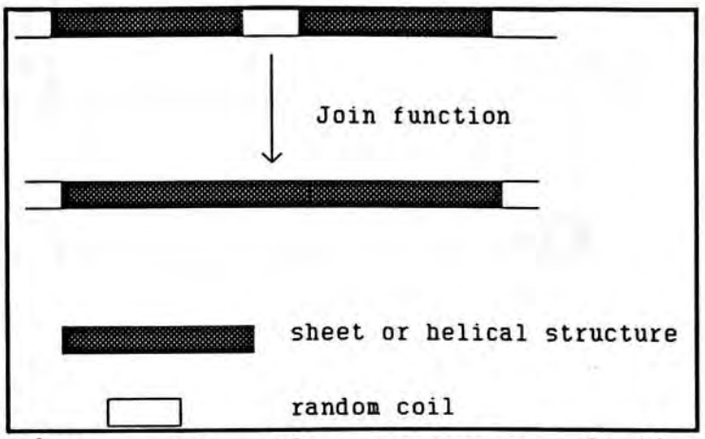


Figure 10. The purpose of the join function.

The aim of our algorithm is to eliminate the weakness of the GOR and the Chou and Fasman method. Furthermore, our algorithm is expected to be easily implemented on computer. The procedure of our prediction algorithm can be summarized as follows :

1. Find the protein secondary structures individually.
2. Select turn structures.
3. Select other secondary structures in the rest of the protein chain.
4. Try to join helical and sheet fragments if possible.
5. If overlap regions occur, try to solve them.
6. Try to join helical and sheet fragments again.
7. Ineligible helical and sheet fragments are assigned as random coil.
8. Amino acid residues without assigned as helical, sheet or turn structure are assigned as random coil.

Chapter 4

A protein tertiary structure prediction method

The ultimate concern of protein modeling is to determine the positions of atoms in a protein molecule.

Amino acids are added one by one to form a linear chain which is the primary structure of a protein molecule. The linkage between two amino acids is called a peptide bond. Some amino acids fold into a special structure such as helix, sheet or turn. These special structures are called secondary structures of a protein molecule. These secondary structures are finally packed into a compact structure which is called protein tertiary structure.

According to the above-mentioned characteristics, this chapter describes how to determine

(i) the coordinates of atoms in a peptide bond,

(ii) the coordinates of atoms in the protein secondary structures, and

(iii) the coordinates of atoms in a protein molecule.

4.1 The Linkage between two amino acids.

Each amino acid consists of an amino group, a carboxyl group and a side chain group which is always denoted by "R". Different kinds of amino acids would have different structures in their side chain group "R". The side chain group of glycine only contains a hydrogen atom. In alanine, a methyl group (CH_3) instead of a hydrogen atom is in side chain group "R".

Atoms in an amino acid can rotate freely. But due to repulsive force of other atoms in the same amino acid, the bond length and subtended angle between two atoms are fixed. The basic skeleton of an amino acid is shown in Figure 11.

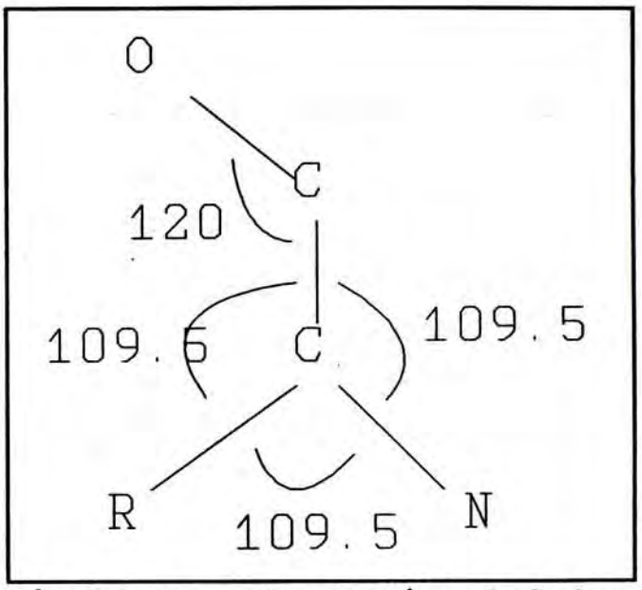


Figure 11. The basic skeleton of an amino acid.

The amino group of the first amino acid combines with the carboxyl group of the second amino acid to form a peptide bond. At the same time, a water molecule is produced. Figure 12 illustrates atoms involved in a peptide bond formation. Furthermore, atoms with label 1 to 5 are lying on the same plane which is called "peptide plane".

There are some assumptions for us to easily determine the coordinates of atoms in a peptide plane.

(i) The relative positions of atoms in different amino acids must be fixed first.

(ii) The peptide plane is assumed lying on the X-Y plane, i.e. the Z-coordinate of atoms are equal to zero.

(iii) The alpha carbon atom of the first amino acid is on the origin (0,0,0), and

(iv) the nitrogen atom of the first amino acid is on the X-axis (X,0,0).

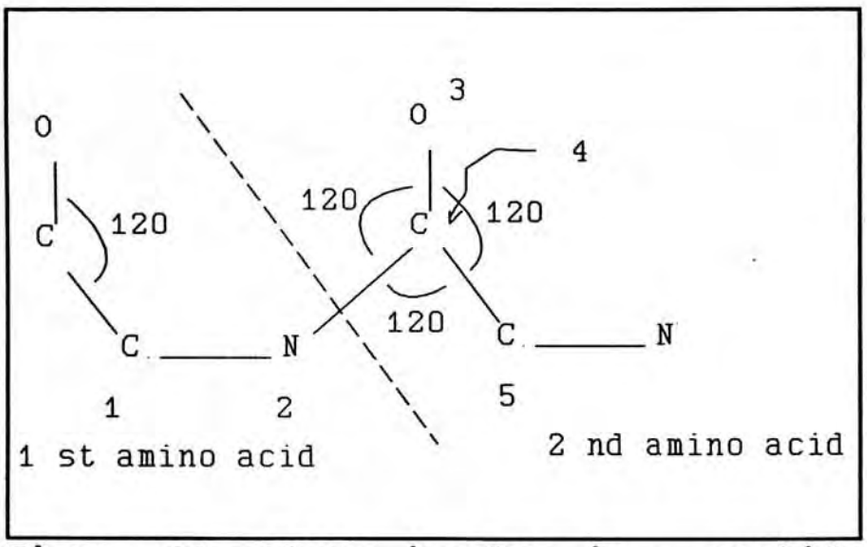


Figure 12. Atoms involve in a peptide bond formation.

The bond length between a nitrogen atom and a carbon atom in a peptide bond is 1.32 Å (1 X 10^{-10} m). The subtended angle of a nitrogen atom between two carbon atoms is 120 degrees. In Figure 13, the coordinates of the carbon atom (X1, Y1) can be calculated by the following equations

$$X1 = X + 1.32 \cos(60^\circ)$$

$$Y1 = 1.32 * \sin(60^\circ)$$

The bond length between a carbon and an oxygen atom in a peptide plane is 1.22 Å. The distance between a carbon and a carbon atom is 1.38 Å.

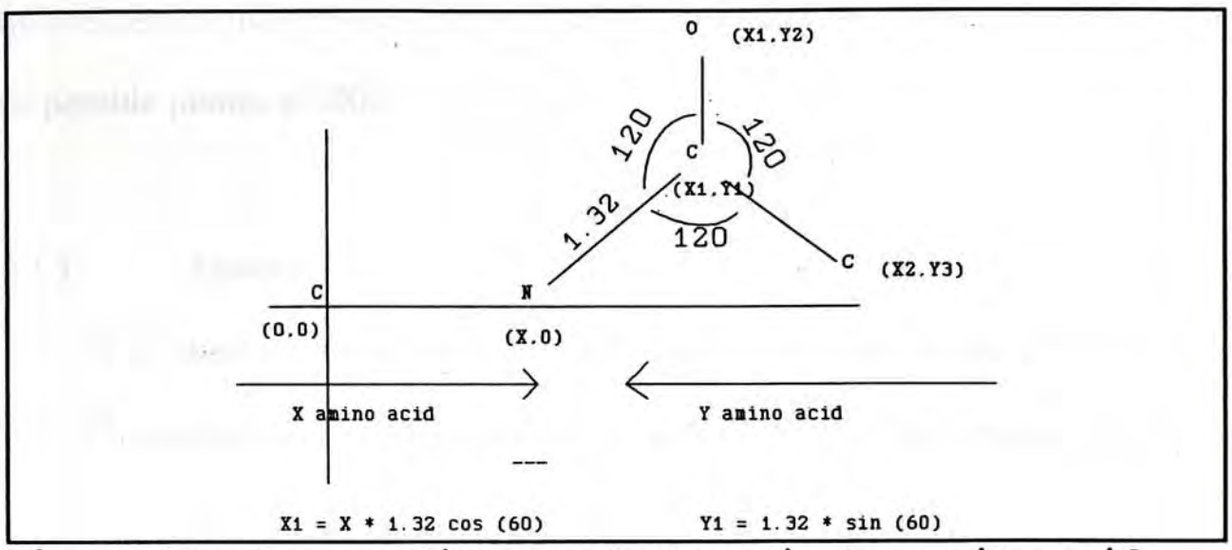


Figure 13. The coordinates of atoms in two amino acids.

Thus

$$Y2 = Y1 + 1.22$$

$$X2 = X1 + 1.38 * \cos(30°)$$

$$Y3 = Y1 - 1.38 * \sin(30°)$$

Similarly, the coordinates of other atoms in these two amino acids can be determined.

4.2 Rotation angle between two peptide planes

When the rotation angle between two peptide planes changes periodically, different protein secondary structures are formed. The rotation angle between two peptide planes is an important factor to determine the coordinates of atoms in a protein secondary structure.

4.2.1 Helical structure

The alpha carbon atom of the first amino acid in a helical structure is 100° apart from the alpha carbon atom of the second amino acid. The perpendicular displacement of these two atoms is 1.5Å. It implies that the rotation angle between two peptide planes is 100Å.

4.2.1.1 Concept

In a helical structure formation, the first peptide plane is placed on the X-Y plane. The second peptide plane is placed on the positive X-Z plane, making an angle 109.5° to the first peptide plane. This picture is shown in (a) of Figure 14.

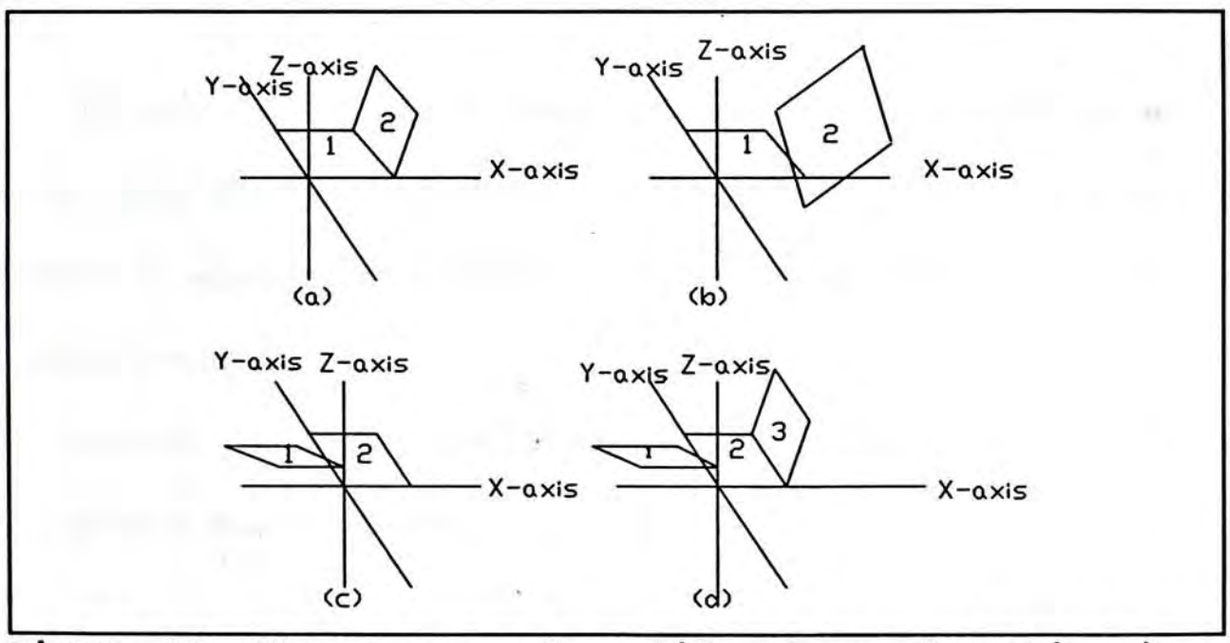


Figure 14. The process of peptide planes formation in a helical structure.

Afterwards, the second peptide plane is rotated 100° about the X-axis. Before the third peptide plane is attached to the second peptide plane, the second peptide plane must be moved to the X-Y plane first. At the same time, the first peptide plane is also moved to the corresponding position. Figure 14 (b) and figure 14 (c) show the above-mentioned phenomena.

The third peptide plane is then connected to the second peptide plane with a subtended angle 109.5°. The third peptide plane is rotated 100° about the X-axis. Similarly, this procedure continues until all the peptide planes are put in the suitable positions.

4.2.1.2 Procedure

We only have the relative positions of atoms in an amino acid but not a peptide plane. Therefore, to determine the coordinates of atoms in a helical structure, an algorithm is established to implement the concept in a helical structure formation.

(i) According to the procedure of peptide plane formation, the first peptide plane is assumed forming on the X-Y plane.

(ii) The alpha carbon atom of the second amino acid is then translated to the origin (0,0,0).

(iii) The coordinates of atoms in amino acids are then moved to the corresponding positions.

(iv) The atoms in amino acids are then rotated 100° about X-axis. This procedure is to simulate the rotation angle between two peptide planes.

(v) Afterwards, the whole structure is rotated and translated in order to move the nitrogen atom of the second amino acid to the X-axis. However, the alpha carbon of the second amino acid is still located at the origin (0,0,0).

(vi) Then the third amino acid is added to the second amino acid according to the process of peptide plane formation.

(vii) Go back to step (ii), replace the second amino acid with the third amino acid. This procedure continues until all the amino acids in a helical structure are moved to suitable positions.

4.2.2 Sheet structure

The sheet structure in a protein molecule is not as smooth as a paper. The sheet structure is somewhat like a paper which is fold several times.

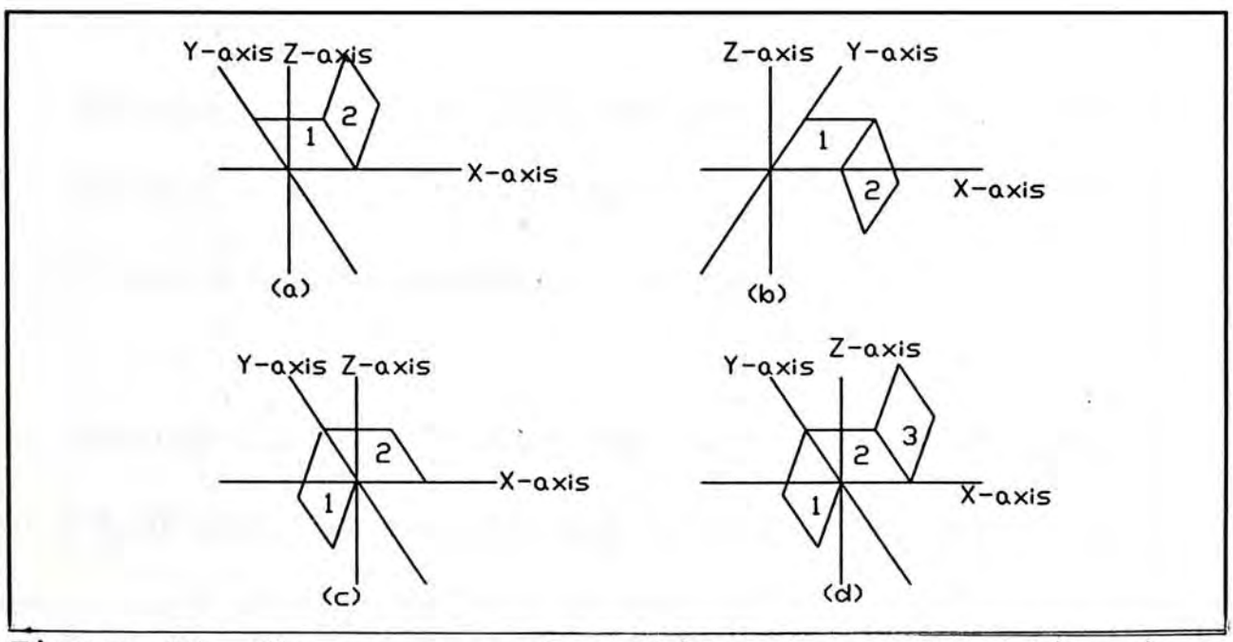


Figure 15. The process of peptide planes formation in a sheet structure.

In a sheet structure, the rotation angle between two peptide planes changes alternatively from 180° to 0°. In (a) of Figure 15, a peptide plane is first placed on the X-Y plane. The second peptide plane is then attached to the first peptide plane. Afterwards, the second peptide plane is rotated 180° about the X-axis, just like (b) in Figure 14. This two peptide planes are then moved until the second peptide plane is placed on the X-Y plane. The third peptide plane is then attached to the second peptide plane. However, the third peptide plane does not rotate about any axis. However, the whole structure is then translated until the third peptide plane is placed on the X-Y plane. Afterwards, the fourth peptide plane is connected to the third peptide plane. The fourth peptide plane is then rotated 180° about the X-axis, just like the second peptide plane. This procedure repeats until all the peptide planes are connected.

4.2.3 Turn structure

The meaning of "turn" is that the direction of the peptide bond formation turns 180° from its original direction. The turn structure in a protein molecule is always formed by four amino acids.

In a turn structure, the rotation angle between two peptide planes is 0°. Thus, it is not necessary to rotate any angle between two peptide planes. The first peptide plane is placed on the X-Y plane first. The second peptide plane is then attached to the first peptide plane. The subtended angle between two peptide planes is 109.5°.

Afterwards, the second peptide plane is then moved to the X-Y plane. At the same time, the first peptide plane also moves to the corresponding position. The third peptide plane is then connected to the second peptide plane. The formation process proceeds until the turn structure is formed.

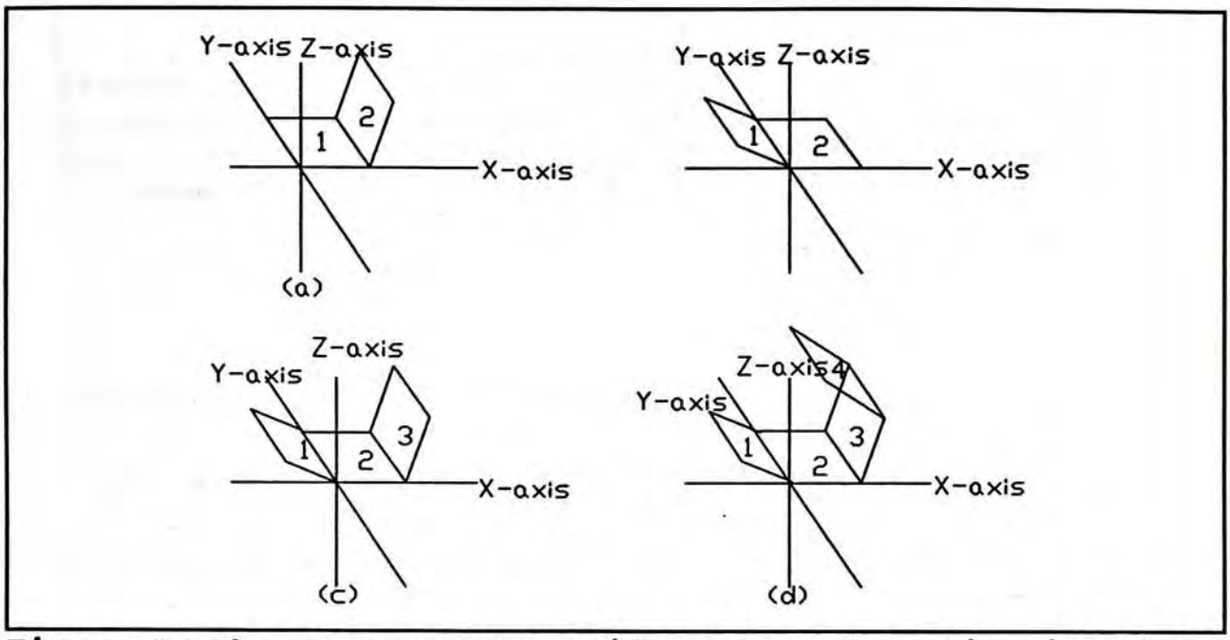


Figure 16 The process of peptide planes formation in a turn structure.

In Figure 17, the picture shows that it is not necessary to alter the rotation angle among the first four peptide planes. However, the rotation angle between the fourth and the fifth peptide plane in a turn structure must be 180°. Otherwise, the atoms in the first and the fifth peptide plane will collide together.

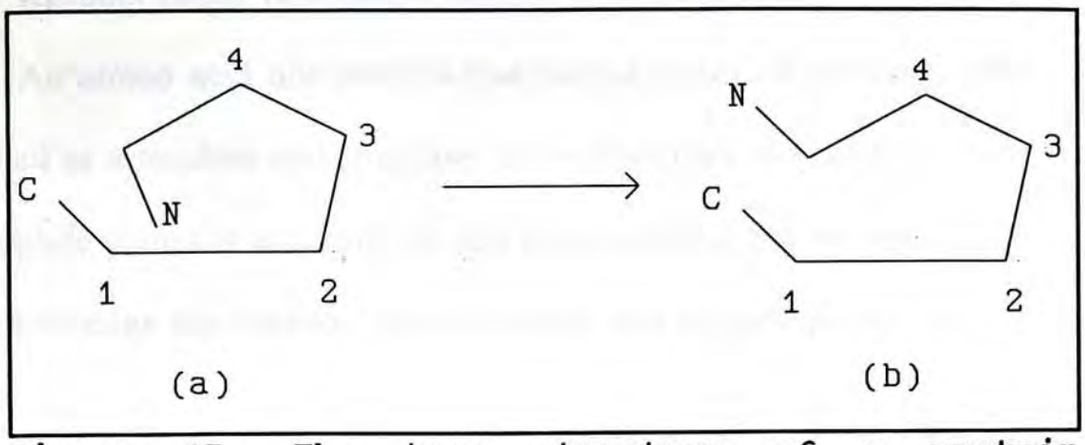


Figure 17. The turn structure of a protein molecule. (a) The direction of peptide planes among the first four amino acids. (b) The rotation angle between the fourth and the fifth peptide plane is 180°.

4.2.4 Anti-parallel sheet and turn structure

Two sheet structures run in similar direction are called a parallel sheet. If they run in opposite direction, the whole structure is called an anti-parallel sheet. The anti-parallel sheet structure is always linked by a turn structure.

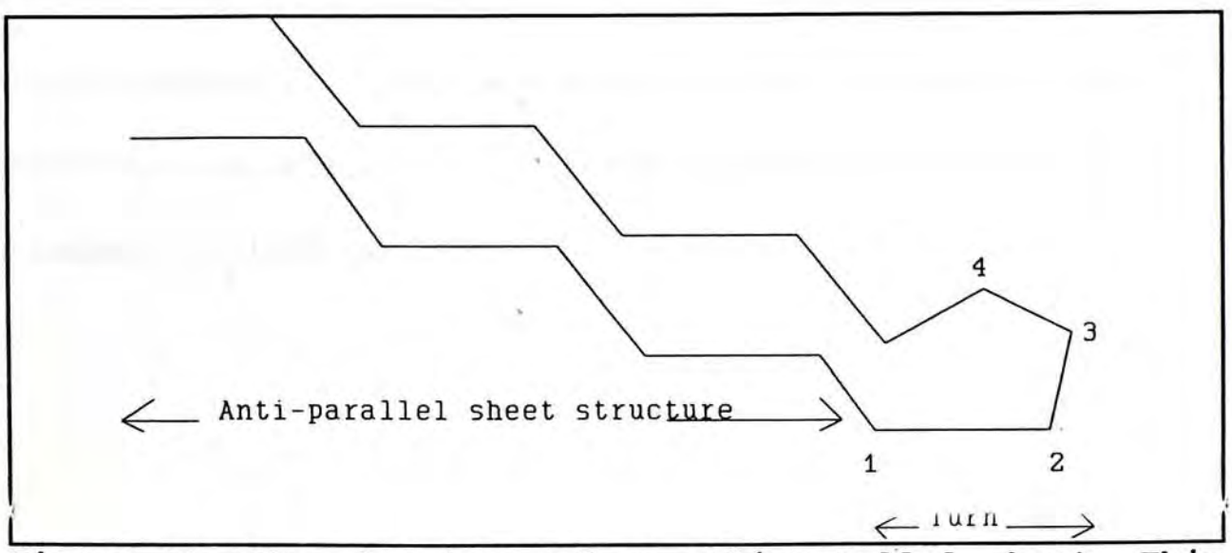


Figure 18. The structure of an anti-parallel sheet. This picture also shows that an anti-parallel sheet structure is always linked by a turn structure.

4.3 Random factor in rotation angle of peptide planes

An amino acid not predicted as helical, sheet or turn structure is finally classified as a random coil structure. In random coil, the rotation angle between two peptide planes is arbitrary. At this stage, there is not enough information for us to determine the rotation angle between two peptide planes in random coils. The simplest way for solving this problem is to create a function which generates a random value falling in the range from 1 to 360.

4.4 Atomic size

Although an atom is so small in size, it still occupies space. Different atoms have different radii. For example, the radius of a hydrogen atom is 0.037 nm, a carbon atom is 0.074 nm and a nitrogen atom is 0.074 nm. When the distance between two atoms is close enough, a covalent bond is formed. Nevertheless, in a protein molecule, atoms in different amino acids would not form a covalent bond except those peptide bonds and disulphide bonds. Therefore, a concept "atom collision" is defined. The radii of two atoms are the best information to determine whether the atoms collide or not. If the distance between two atoms is smaller than the summation of their radii, "atom collision" occurs.

4.5 Tertiary structure prediction algorithm

In my protein tertiary structure prediction system, two main rules must be obeyed. The first rule is that two atoms cannot be too close to each other; that is, "atom collision" is prohibited. Secondly, the process of solving the collision of atoms must not break the secondary structures which are predicted previously. The best way for solving the "atom collision" problem is to change the rotate angle between two peptide planes which are belonged to a random coil. It is because changing the structure of random coils would not affect the secondary structures of a protein molecule.

The protein tertiary structure prediction algorithm is described as follows :

```
ptr      --   pointer points to an amino acid sequence.
copy(aa) --   The coordinates of atoms of an amino acid is copied to
              memory locations which contain the final coordinates of
              atoms of a protein molecule
ptr = 0;
copy(aa);
rotate a nitrogen atom on the X-axis;
ptr = 1;
Do while ptr < = the number of amino acid in a protein molecule {
        copy(aa);
        calculate the position of the alpha-carbon in ptr->amino acid;
        shift the alpha-carbon of ptr->amino acid to that position;
        check whether atoms in ptr->amino acid are collided with other
        atoms;
        if atom collision occurs {
                solve atom collision problem until no atom collision is
                occurred;
        }
        move the whole molecule so that the coordinates of the alpha carbon
        of ptr->amino acid = 0,0,0
        if ptr->amino acid = random coil {
                if previous amino acid = turn {
                        rotate atoms in ptr->amino acid about X-axis with 180
                        degrees;
                }
                bond_angle = generate a random angle between 1 to 360;
                rotate atoms in ptr->amino acid about X-axis with
                bond_angle;
        }
        if ptr->amino acid = helical structure {
                rotate atoms in ptr->amino acid about X-axis with 100
                degrees;
        }
        if ptr->amino acid = sheet structure {
                if ptr->amino acid is the even number amino acid in the
                sheet structure. {
                        rotate atoms in ptr->amino acid about X-axis with 180
                        degrees;
                }
        }
        ptr = ptr + 1;
}
```

```
Solve collision routine :
        back_ptr -- pointer points to an amino acid;
        back_ptr = ptr - 1;
        Repeat {
                time = 0;
                Repeat {
                        rotate atoms in back_ptr->amino acid about X-axis with 15
                        degrees;
                        time = time + 1;
                        check atom collision occurs or not;
                }
                until no collision occurs or time >= 24;
                if collision still occurs {
                        Find another amino acid which is predicted as a random coil.
                }
        }
        until no collision occurs or back_ptr < 0;
        if back_ptr < 0 {
                exit the system and display error message
        }


Check atom collision routine :
        ptr -- pointer points to current amino acid
        Total_amino_acid -- The number of amino acid in a protein molecule
        Collide --      A flag which indicates whether atoms collide or not.

        Collide = 0   /* No atom collision */
        ptr = 1;
        For (counter1 = 1 to Total_amino_acid {
                For     (counter2 = 1 to atoms of an amino acid that pointed by ptr)
                {
                        Find atom radius;
                        For     (counter3 = 1 to all atoms in a protein molecule
                                except those atoms belong to an amino acid that
                                pointed by ptr) {
                                Find atom radius;
                                Sum_radii = Sum atom radii;
                                Dist =          Calculate the distance between two
                                                atoms;
                                If ( Sum_radii > Dist)
                                        Collide = 1;  /* Collision occurs */
                        }
                }
         }
```

Chapter 5

Implementation

Our project mainly includes two parts:

(i) To determine the coordinates of atoms in the tertiary structure of a protein molecule from the primary structure of a protein molecule.

(ii) To display the graphic image of a protein molecule in a monitor, and furthermore, provide facilities so that the user can directly manipulate the graphic image interactively.

In this chapter, primarily, the hardware and software facilities which are used to implement our project are briefly described. The user-defined data types, data structures and some special techniques used to implement or to improve the performance of our project are also discussed.

5.1 Hardware

An IBM AT compatible machine is used to implement our project. This machine uses a Intel 80386 microprocessor. The graphic image of a protein molecule, in general, is quite complicated. A monitor with high capacity in resolution and colour must be adopted to show a complicated picture. The NEC 3D MultiSyn monitor is selected. Furthermore, a VGA (Video Graphic Array) adaptor is used to communicate between the machine and the monitor.

From the view point of software engineering, our project can be classified as a rule-based expert system. This system is composed of many rules to build the secondary structures and the tertiary structure of a protein molecule. Moreover, the tertiary structure of a protein molecule is displayed on a screen. Comparing to Prolog, the C language is weak in rule manipulation. However, its rich set of built-in functions for graphics and image manipulation is good enough for us to implement our project. Therefore, C is used in our software implementation.

Machine (microprocessor)	80380 SX (20 MHz)
Graphic interface	VGA (Video Graphic Array)
Main Memory	2M bytes
Operating system	DOS 3.3
Programming language	Turbo C 2.0

5.2 User-defined data types and data structures

Our system accepts the amino acid sequence of a protein molecule as basic information. Each amino acid, in biochemical notation, can be represented by a capital letter. A capital letter "G" represents amino acid glycine. "H" is used to represent histidine. Therefore, *an array of characters* can be used to store the amino acid sequence.

In protein secondary structure prediction, the state of an amino acid in a protein molecule can be represented by "H", "B", "T" or "C" character. "H" is used to represent helical structure. "C", "B" and "T" represent the random coil, sheet structure and turn structure respectively. Hence, *an array of characters* can be used to contain the results of prediction.

In protein tertiary structure prediction, a list of coordinates of atoms are finally created. Each atom has its X-, Y- and Z-coordinates. Moreover, a character field must be used to represent the atomic type. In this way, a record containing atomic type and its coordinates call "atomic coordinates" can be created for each atom in the tertiary structure.

```
Record of atomic coordinates
     Atom type      : character
     X-coordinate   : real
     Y-coordinate   : real
     Z-coordinate   : real
```

It is because a character occupies a byte and a real number takes four bytes. Therefore, an atomic coordinates record takes thirteen bytes. However, under the Disk Operating System (DOS), the maximum size of a data segment is 64K bytes. If the size of a group of data is greater than 64K bytes, errors may occur. Due to this constraint, our system can handle 5000 atoms only. Therefore, it does not work on a protein molecule which made up of more than 5000 atoms. Of course, this constraint does not occur in bigger computers.

Furthermore, a data file which contains the atomic coordinates of an amino acid is created. It is because an amino acid can be viewed as a single unit in protein tertiary structure prediction. During peptide plane formation, the whole set of the atomic coordinates can be added to the existing structure.

5.3 Technique in molecule displaying

One of the aims in computer graphics is to display objects in realistic sense. There are many methods to display objects with depth sense. They are shading, rendering and ray tracing. In our protein modeling methodology, shading technique is adopted. (The shading algorithm of a sphere is described in Appendix D)

There are different atoms in a molecule. In general, different atoms are represented by different colours and sizes. The radius of a carbon atom is greater than that of a hydrogen atom. If a red sphere represents a carbon atom and a green sphere represents a hydrogen atom, you will see that a red sphere is greater than a green one.

Z-buffer is one of the techniques to remove hidden line or hidden surface [34]. The advantage of the Z-buffer technique is its constant speed in image generation regardless of the complexity of the image. However, this technique uses a lot of memory locations. A picture formed by 640 * 480 pixels and 256 different colours needs 1.2M bytes. Although this amount of memory location is available in many workstations, it is a burden on the IBM micro-computer. After evaluating the speed, memory utilization and complexity of the algorithms, an algorithm that

put images of atoms on a screen is finally employed.

The plane of a screen is defined as the X-Y plane of a three dimensional space. As the positive Z-axis points directly to the user, an object with a small value in Z-coordinate may be obscured by other objects having greater Z-coordinates. Therefore, the Z-coordinates of atoms of a molecule are sorted in an ascending order. The image of an atom with smallest value in Z-coordinate is first put on a screen. The image of other atoms are then put on a screen one by one. The final picture is generated after all the images of atoms are sent to a screen. Some pictures generated by this algorithm are shown on Figures 22, 23, 24, 25 and 26.

5.4 Image processing

Generate the picture of a sphere by the shading technique requires, on average, 2 to 5 seconds. If a molecule has thousands of atoms, the whole picture can take more than one hour to create. It is not feasible in interactive graphic manipulation. Therefore, image processing techniques are adopted in order to improve the performance.

The graphic image of each kind of atom is first generated and saved in a data file. When a particular atom is to be displayed on a screen, the graphic image of this atom is sent to a screen rather than using shading techniques to generate. In this way, only several seconds are required to create the display of a molecule having hundred atoms.

A graphic image, in general, is bounded by a rectangle. In molecule modeling, images of sphere are displayed on a screen. There are two rectangular graphic images rather than one rectangular image are put on a screen in order to generate the desired image of a sphere.

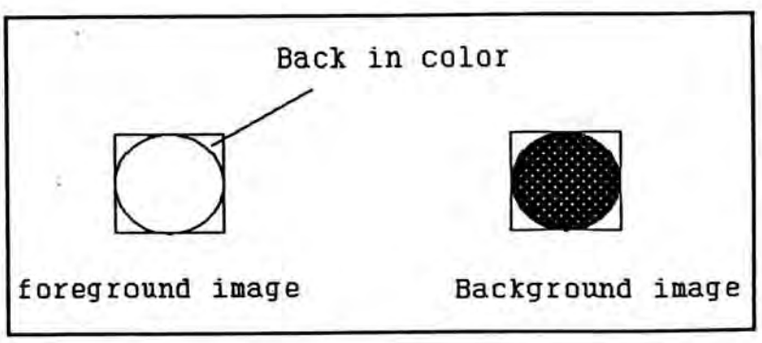
In VGA (Video Graphic Array) standard, a single byte is used to represent a single colour. There are 256 different combinations in a single byte. Hence, 256 different colours can be shown on a screen simultaneously. The colour "black" is always represented by bit map 0. It means that all the bits in a byte are set to 0. However, the colour "white" is represented by bit map 255, i.e. all the bits in a byte are assigned to 1.

"AND" and "OR" operations are always used in image processing. In "AND" operation, a "0" bit is generated if there is only one "0" among those bits. However, in "OR" operation, a "1" bit is gained if there is only one "1" among those bits.

According to this logic and the above-mentioned description, the "black" colour is overridden by any other colour in an "OR" operation. In an "AND" operation, the "black" colour overrides other colours. Similarly, in an "OR" operation, the "white" colour overrides other colours. It should be noted that in an "AND" operation, the "white" colour is overridden.

```
Black     AND  any colour     =>   Black
Black     OR   any colour     =>   any colour

White     AND  any colour     =>   Any colour
White     OR   any colour     =>   White
```



In our project, two images, let say, foreground image and background image are created first. In background image, a circle with "black" colour is embedded in a square which is "white" in colour. Nevertheless, a circle with a desired colour is embedded in a square which is "black" in colour in the foreground image.

In handling the image of a sphere, the background image is first put on a screen by an "AND" operation. According to the bit operations, a circle with "black" in colour is then put on a screen. It is because in an "AND" operation, the "white" colour is overridden. Afterwards, the foreground image is put on a screen with same location of the background image by an "OR" operation. Thus, the "black" colour is overridden. In this way, a sphere with a particular colour is finally shown on a screen.

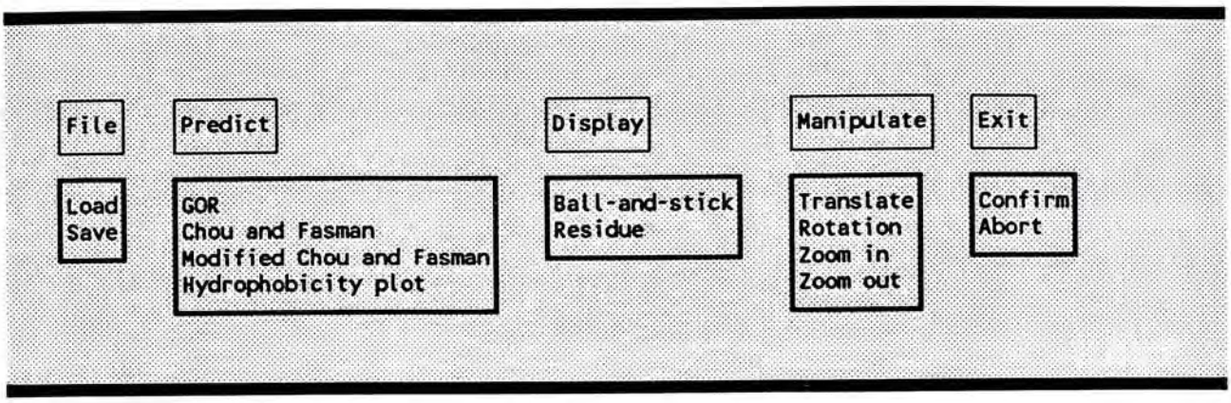
In our project, a pair of images of each atom are first saved in files. When an atom is expected to be sent to a screen, the background and foreground images of this atom are then put on a screen. Thus, no shading technique is required to generate an image every time.

5.5 Options in our program

There are five options under the pull-down menu, namely, "File", "Predict", "Display", "Manipulate" and "Exit". Using the "File" option, a file containing the primary structure of a protein molecule can be loaded into RAM. Furthermore, the coordinates of atoms of a protein molecule to be predicted by our methodology can be saved into a file too.

The protein secondary structure prediction algorithms including the GOR method, the Chou and Fasman method and the modified Chou and Fasman method are put in the "Predict" option. Moreover, the hydrophobicity plot of a protein molecule is also embedded in the "Predict" option. A user can choose

different methods to predict the tertiary structure of a protein molecule so that the results generated by different methods can be compared.



Either the ball-and-stick model or the amino acid residue model can be selected in the "Display" option. Translation, rotation and scaling operations are provided in the "Manipulate" option for a user to manipulate a molecule interactively. A molecule can be translated in the X- and Y-axis. Similarly, a molecule can be rotated in the X-, Y- and Z-axis.

Finally, there are two options in the "Exit" menu, the "Confirm" and "Abort". A user leaves this program when "Confirm" is chosen. The default value of the "Exit" menu is "Abort". It prevents a careless user from choosing the "Exit" menu.

5.6 Steps in protein tertiary structure prediction

Before using this system, an ASCII file containing the primary structure of a protein molecule is created. After our system is activated, the "File" option must first be selected. Using the "Load" option, a user specifies the file previously created. The primary structure of a protein molecule is loaded into RAM. Then you can choose one of the options in the "Predict" column. After confirmation, the tertiary structure of a protein molecule is then predicted. Several minutes later, the final image of a protein molecule is shown on a screen.

The user can select different options under the "Display" menu to show different characteristics of a protein molecule. Furthermore, he/she can also manipulate this molecule by selecting options under the "Manipulate" menu. If the coordinates of atoms of this molecule are to be saved, the "Save" option in the "File" menu must be chosen.

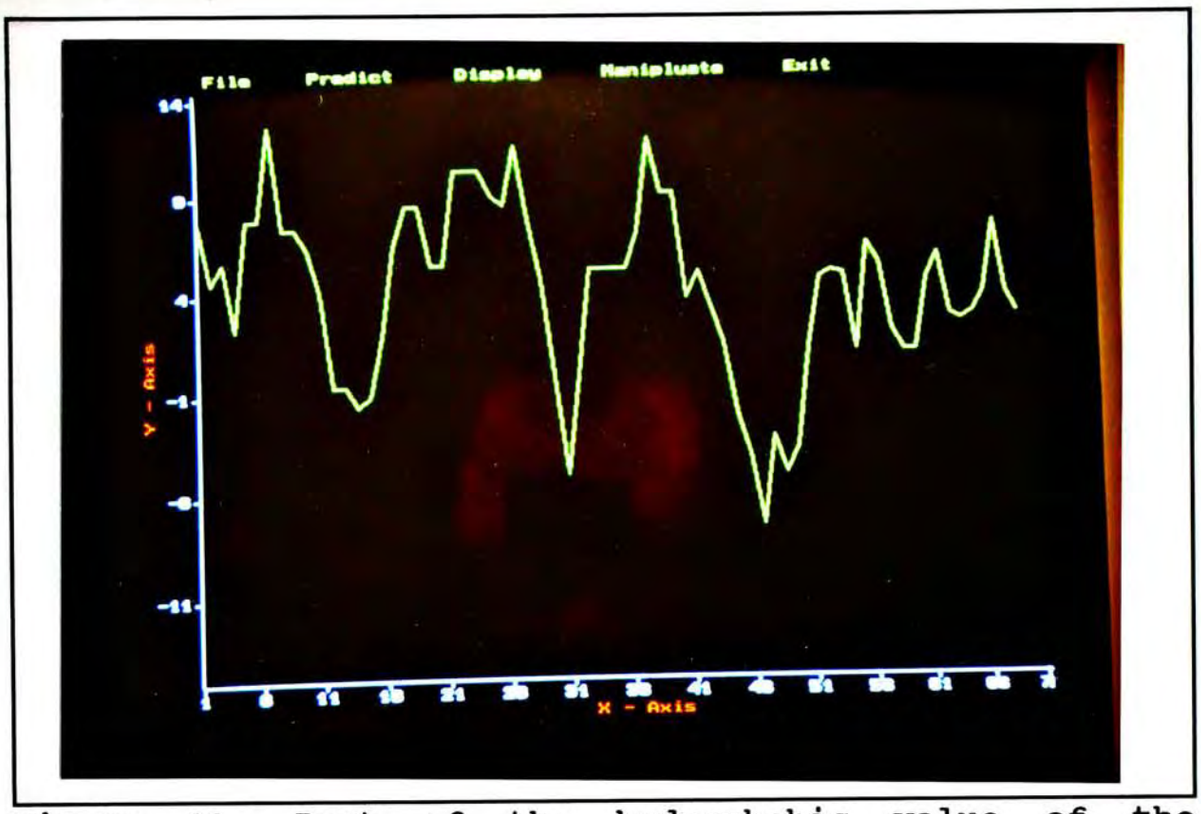


Figure 19. Part of the hydrophobic value of the tricosanthin. (From amino acid 1 to 70)

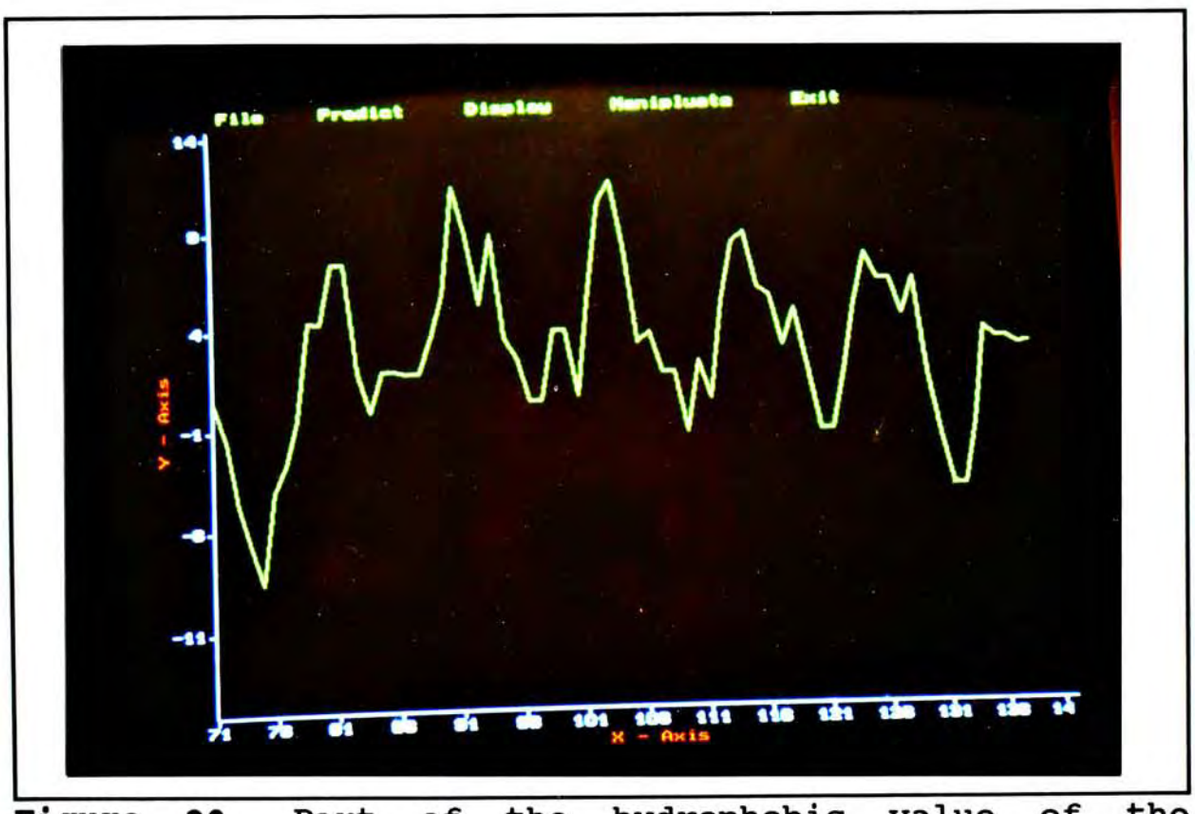


Figure 20. Part of the hydrophobic value of the triconsanthin. (From amino acid 71 to 140)

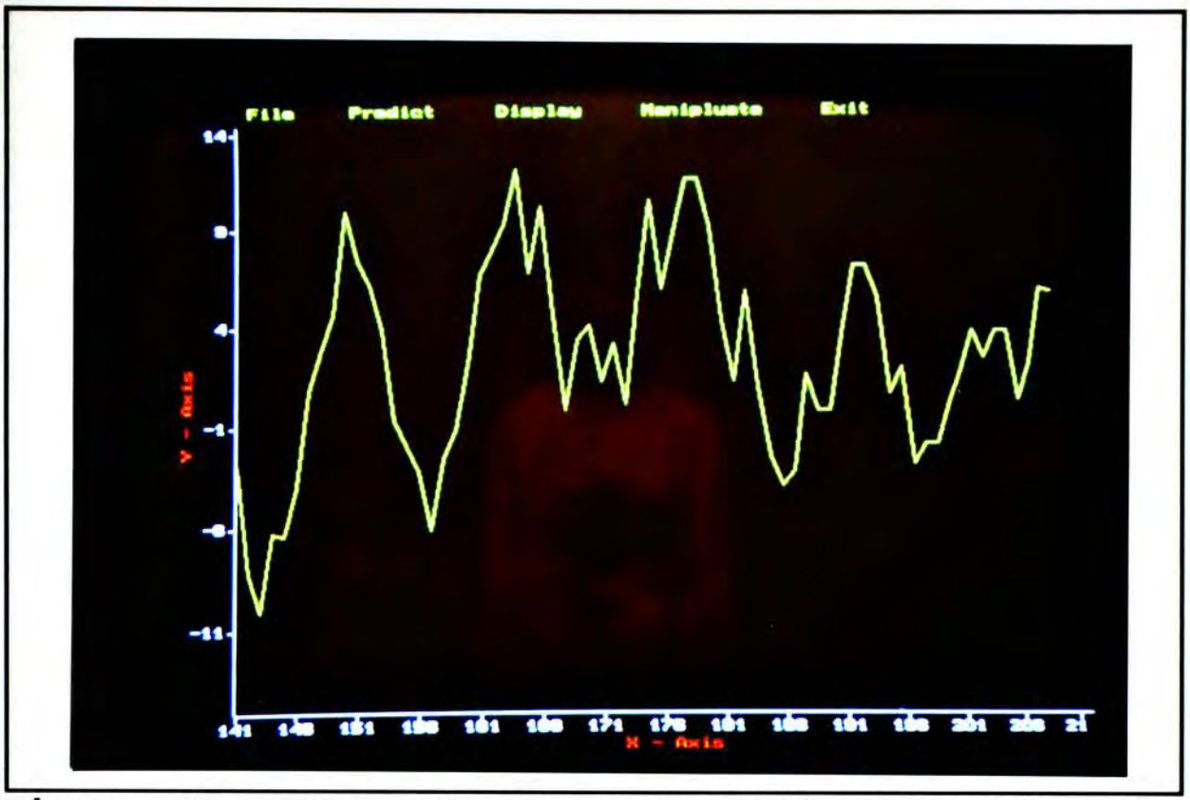


Figure 21. Part of the hydrophobic value of the tricosanthin. (From amino acid 141 to 210)

Figure 22. The amino acid residue representation of a tricosanthin molecule.

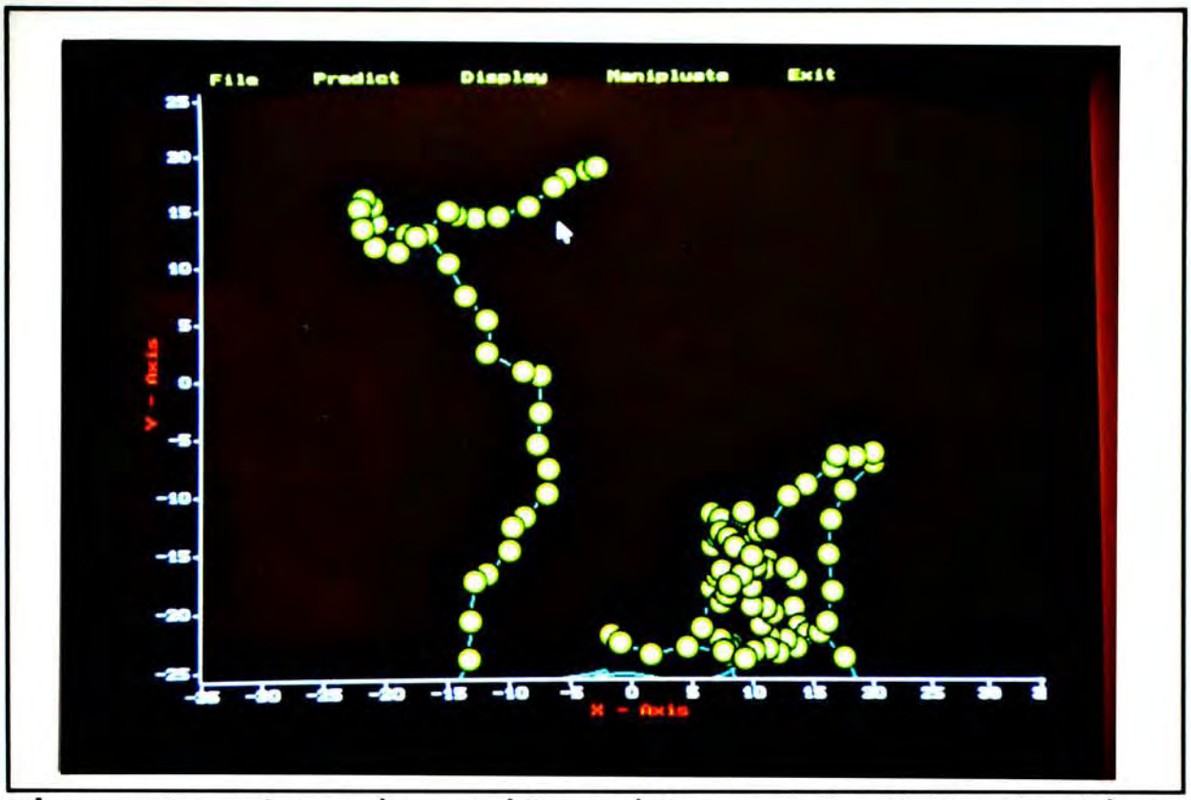


Figure 23. The amino acid residue model of the tertiary structure of a tricosanthin molecule. Sheet structure is pointed by an arrow.

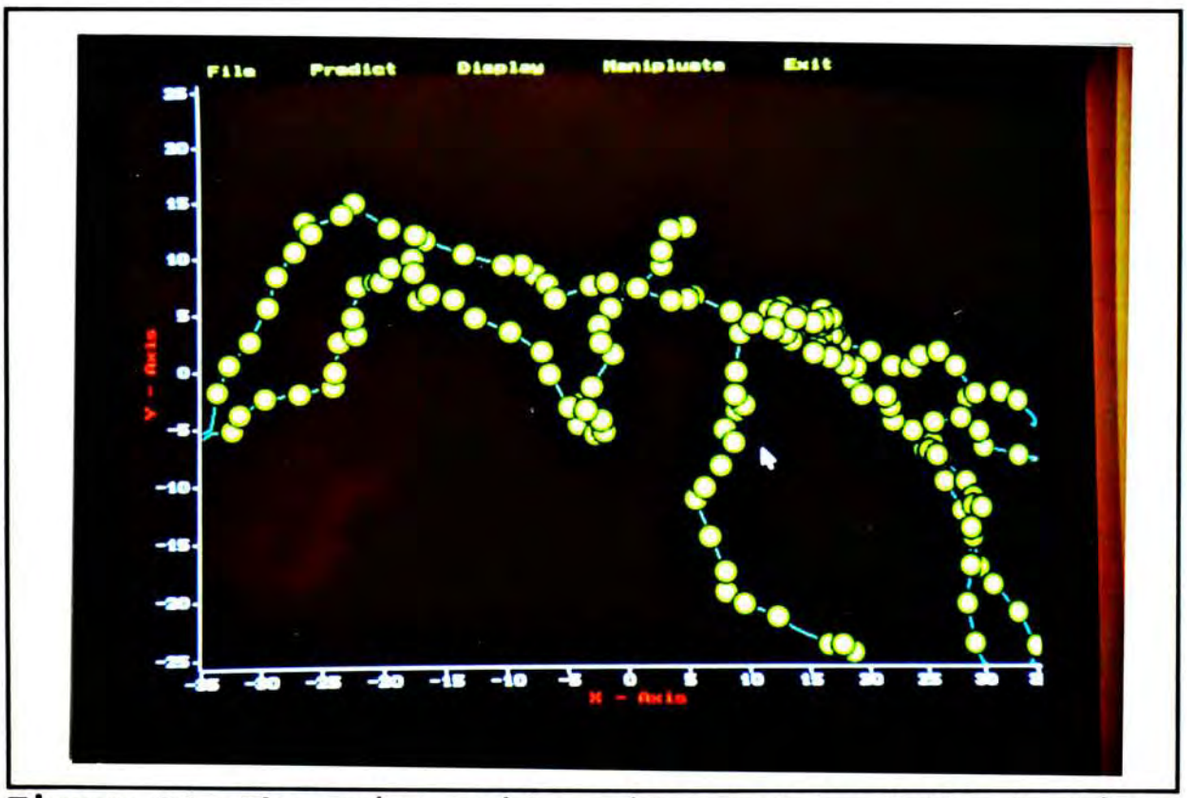


Figure 24. The amino acid residue model of the tertiary structure of a tricosanthin molecule. Helical structure is pointed by an arrow.

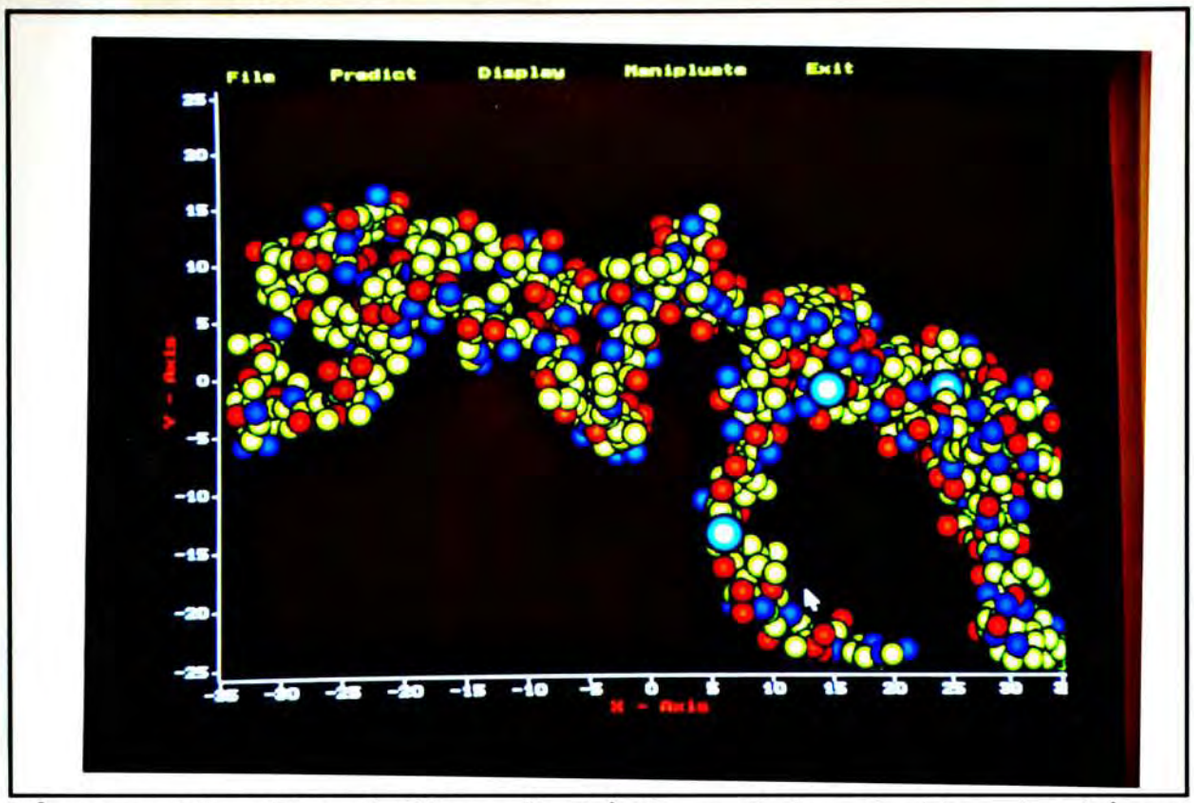


Figure 25. The ball-and-stick model of the tertiary structure of a tricosanthin molecule. Helical structure is near an arrow.

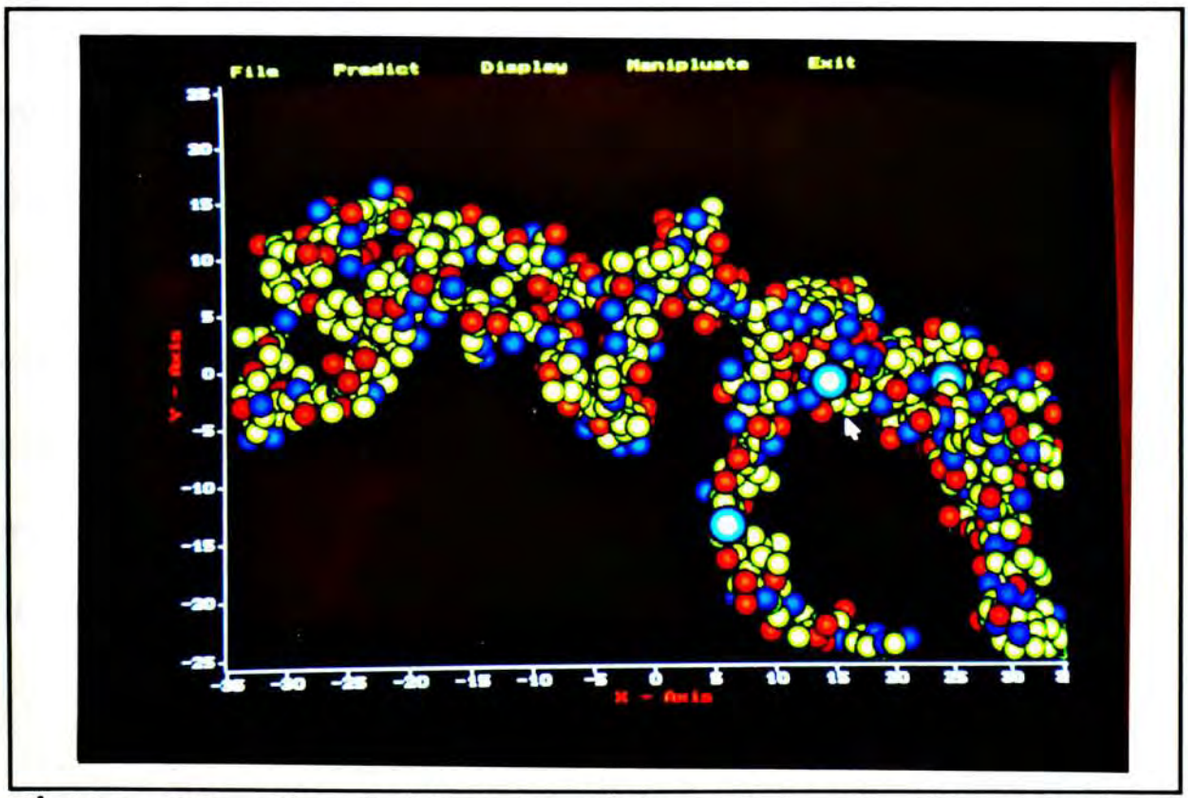


Figure 26. The ball-and-stick model of the tertiary structure of a tricosanthin molecule. The sulphide atom is pointed by an arrow.

Chapter 6

Results

Two prediction algorithms are proposed in this project. The former predicts the secondary structures of a protein molecule while the latter predicts the tertiary structure of a protein molecule.

6.1 The result of protein secondary structure prediction

The structure of a native protein molecule can be determined by X-ray crystallography. As mentioned before, there are four types of secondary structures in protein molecules : helix, sheet, turn and random coil. The state of an amino acid in a protein molecule belongs to only one of these four types of the secondary structures. Suppose that an amino acid in a protein molecule belongs to a helical structure. If the result of prediction also indicates that this amino acid belongs to a helical structure, we can say that the algorithm is successful to predict the state of this amino acid.

```
HHHHHHBBBBBTTTTBBBBBCCCCC      native state

HHHHHBBBBTTTTCCCBBBHHHHH       predicted state

↑↑↑↑↑ ↑↑↑  ↑↑   ↑↑↑

H :  helix
B :  sheet
C :  random coil
T :  turn
```

Figure 27. Amino acids pointed by arrow are successful in prediction.

There are many methods to evaluate the predictive schemes for the secondary structures of a protein molecule. The best and the simplest way to estimate the power of prediction of a predictive scheme is the fraction of amino acids correctly assigned. The power of prediction *f* is represented as

$$f = (\Sigma_S F^+_S) / (\Sigma_S F_S) = \Sigma_S F^+_S / N \qquad [24]$$

where F_S is a number of amino acids observed in a secondary structure S. The summation of amino acids in all the secondary structures S is equal to the total number of amino acids, N. F^+_S is the number of amino acids correctly predicted in the secondary structure S. The fraction *f* can be conveniently transformed on a percentage basis (*f* X 100%).

Different kinds of prediction methods give different prediction results in different protein molecules. For "A" protein molecule, the Chou and Fasman method gives a better result than the GOR method. However, for "B" protein

molecule, the result of the GOR prediction method is better than that of the Chou and Fasman method. The average power of prediction for all protein molecules in a database is the best indicator to determine which method gives the best result of prediction. This evaluation criterion is proposed by Kabsch and Sander [24].

The native state of protein molecules in table I and II are obtained from Lim's paper [49] in which the turn structure is neglected. In Lim's research, he predicted the helical, sheet and random coil structure in protein molecules. Thus, Lim's paper only provides information on three types of secondary structures in protein molecules. In this project, two different methods are used in order to test which one gives a better result. First, like Lim's approach, the prediction for turn structure is not attempted. Hence, this prediction method only predicts helix, sheet and random coil. The power of prediction is 44.4% and the results are shown in table I. In another method, the turn structure prediction algorithm is used. However, amino acids predicted as turn structures are classified as random coil. It is because turn structure is classified as random coil structure in Lim's method. The power of prediction of this method is better than the previous one. The successful percentage is 59.22 and the results are shown in table II.

Table I. The prediction power of 16 protein molecules in 3 states (helix, sheet and random coil).

Protein	Total residues	amino acid in right states	percentage (%)
Cytochrome b_5	93	31	33.33
Cytochrome c	104	59	56.73
Cytochrome c_2	112	31	27.69
Elastase	240	78	32.5
Ferredoxin	54	26	48.15
Hemoglobin alpha-chain	141	88	58.16
Hemoglobin beta-chain	146	68	46.57
Lysozyme	129	67	51.94
Myoglobin	153	97	63.40
Nuclease	149	74	49.66
Papain	212	82	36.68
Ribonuclease S	124	63	50.81
Rubredoxin	54	22	40.74
Subtilisin BNP	275	123	44.73
Trypsin inhibitor	58	35	60.34
Trypsinogen	229	71	31.00
Average	2273	1009	44.40

In another aspect, protein molecules can be classified into four groups depending on their secondary structures. These four groups are *all-helix*, *all-sheet*, *helix/sheet* and *helix + sheet*. Protein molecules containing no sheet structure are included in the all-helix group. Protein molecules do not have any helical structure are included in the all-sheet group.

Table II. The prediction power of 16 protein molecules in 4 states (helix, sheet, turn and random coil).

Protein	Total residues	amino acid in right states	percentage (%)
Cytochrome b_5	93	39	41.94
Cytochrome c	104	67	64.42
Cytochrome c_2	112	67	59.82
Elastase	240	149	62.08
Ferredoxin	54	43	79.63
Hemoglobin alpha-chain	141	64	45.39
Hemoglobin beta-chain	146	64	43.83
Lysozyme	129	75	58.14
Myoglobin	153	71	46.4
Nuclease	149	103	69.13
Papain	212	131	61.79
Ribonuclease S	124	67	54.03
Rubredoxin	54	34	62.96
Subtilisin BNP	275	195	70.91
Trypsin inhibitor	58	34	58.6
Trypsinogen	229	143	62.44
Average	2273	1346	59.22

Protein molecules which contain helical and sheet structures occurring randomly are included in the helix + sheet class. However, a protein molecule is classified as the helix/sheet group if its helical and sheet structures occur alternatively.

15 protein molecules in Table II are divided into four groups as mentioned above. Four protein molecules are included in the all-helix group. Four are classified as members of the all-sheet group. Six and one protein molecules are categorized in the helix + sheet and helix/sheet groups respectively. However, cytochrome c_2 cannot be categorized in any of these four groups since there is not enough scientific information to classify it.

Protein molecules belonging to the all-helix group do not possess any sheet structure. Similarly, in the all-sheet group, protein molecules do not contain any helical structure. Hence, two additional rules can be drawn

Rule 1: There is no need to predict sheet structure of a protein molecule in the all-helix group.

Rule 2: The prediction algorithm for helical structure would not give a better result if a protein molecule is in the all-sheet group.

The secondary structures of protein molecules in the helix + sheet and helix/sheet groups do not have any special characteristic. Therefore, no additional rule can be drawn from the helix/sheet and helix + sheet groups.

Protein molecules belonging to the all helix group can be used to assess the effectiveness of rule 1. The prediction results for this group protein molecules without using rule 1 are first obtained. Afterwards, rule 1 is employed in the same algorithm. Another set of prediction results is then generated.

Table III. The prediction results of protein molecules belonging the all-helix group are predicted by two different methods.

Protein	Rule 1 is not used	Rule 1 is used
Cytochrome c	64.42	69.23
Hemoglobin alpha-chain	45.39	56.74
Hemoglobin beta-chain	43.83	69.18
Myoglobin	46.4	81.05
Average	50.92	69.30

Table IV. Proteins in the all-sheet group are predicted by two different methods.

Protein	Rule 2 is not used	Rule 2 is used
Elastase	62.08	52.08
Ferredixon	79.63	79.63
Rubredoxin	62.96	90.74
Trypsinogen	62.44	73.80
Average	63.95	66.90

In table III, the middle column shows the prediction values without using rule 1. On the other hand, the right hand side column shows the prediction values when rule 1 is employed. It is easily seen that the prediction values for individual protein molecule increase when rule 1 is used. The average prediction measure also increases from 50.92 to 69.30.

Similarly, in table IV, different prediction results for protein molecules in all-sheet group are obtained. Obviously, rule 2 is useful in predicting protein molecules without any helical structure. The average prediction result also improves from 63.95 to 66.90.

The average prediction value in table II increases from 59.22 to 64.85 when both rules 1 and 2 are used.

6.2 The results of protein tertiary structure prediction

A list of coordinates is the usual output of the tertiary structure prediction of a protein molecule. As a measure of their success, most investigators reported the root mean square (r.m.s.) deviation of the atomic positions in their models from those obtained using crystallographical methods.

There are two procedures commonly used to calculate the r.m.s. deviation : the rotation and interatomic distances method. The former is used in investigating questions of evolutionary similarity, and the latter one is normally evaluated in folding studies.

The r.m.s. deviation based on the rotation method is computed as :

$$\Delta r = \sqrt{(\Sigma(x_i - y_i)^2 / n)} \qquad (2)$$

where x_i, y_i are the atomic coordinates of the atoms in the crystal structure and n is the total number of atoms in that protein molecule.

The r.m.s. deviation based on the interatomic distances method is computed as :

$$\Delta d = \sqrt{(\Sigma\Sigma(d_{ij} - e_{ij})^2 / n^2)} \qquad (3)$$

$$d_{ij} = \sqrt{((x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2)} \qquad (4)$$

where d_{ij} are the interatomic distances between atoms in the crystal structure and e_{ij} are the interatomic distances between atoms in the generated structure. The x_j, y_j, z_j are the X-, Y-, Z-coordinates of an atom that is generated by a predictive method. The x_i, y_i, z_i are the X-, Y-, Z-coordinates of an atom that is obtained by crystallographicl methods.

Cohen and Sternberg discovered that the Δd was not a good criterion to assess the result of prediction of a protein molecule. It was because they found that when the generated structure was translated in any direction, Δd would fluctuate very sharply. However, they found that this phenomenon would not occur when the generated structure was rotated[12].

Furthermore, Cohen and Sternberg[12] developed the following equation that relates the number of amino acids in a protein molecule chain "N" to the r.m.s.

deviation ($\hat{\Omega}r$) of a randomly generated structure from its native structure.

$$\hat{\Omega}r = 0.0468\ N + 9.25 \qquad (5)$$

For example, if a protein molecule has 105 amino acids, the $\hat{\Omega}r$ in the above-mentioned equation would have a value 14.16. This value indicates that there is 14.16 rotation deviation between a native protein structure and a randomly generated structure of the same protein sequence.

The prediction algorithm can be evaluated by the $\hat{\Omega}r$ between the randomly generated structure and the predicted structure. The $\hat{\Omega}r$ of a predicted structure can be calculated from equation (2). The $\hat{\Omega}r$ of a randomly generated structure of this protein can be obtained from equation (5). If the $\hat{\Omega}r$ of the predicted structure is less than that of a randomly generated structure, it means that the predicted structure is better than a randomly generated structure. The lower the $\hat{\Omega}r$ value is, the better the result of prediction is.

The coordinates of atoms of a tricosanthin molecule are used to assess our prediction algorithm. Tricosanthin contains 234 amino acids and with totally 1806 atoms. The $\hat{\Omega}r$ of the randomly generated structure of a tricosanthin molecule is equal to

$$0.0468 * 234 + 9.25 = 20.20$$

To calculate the minimum rotation deviation of the tricosanthin, the native structure of a tricosanthin molecule was first fixed. The predicted result of a tricosanthin molecule was rotated around the X-axis and Z-axis in order that the minimum rotation deviation could be obtained. A tricosanthin molecule was rotated around the X-axis and Z-axis from 1 to 360° with 1° increment. Therefore, 129600 (360 * 360) rotation deviations would be generated. The minimum value in these 129600 combinations is selected as the minimum rotation deviation.

The atomic radius was also used as a constraint in protein folding. The atomic radius varies from 0 to 0.6 angstrom (10^{-10}m). The atomic radius of carbon atom is 0.74. Half of this length 0.37 is also used to evaluate the results of prediction.

Table V. The rotation deviation (∆r) of the tricosanthin which was predicted by our algorithm.

atomic radius	minimum rotation deviation
0.0	14.21
0.1	10.85
0.2	21.17
0.3	22.74
0.37	17.85
0.4	17.85
0.5	14.06
0.6	19.64

In Table V, the minimum rotation deviation ∆r of a tricosanthin molecule was 10.85. It is good enough to demonstrate the effectiveness of our prediction algorithm.

Chapter 7

Conclusion

A new methodology in predicting the tertiary structure of a protein molecule has been described here. This methodology consists of two parts, the first part is an algorithm to predict the secondary structures of a protein molecule. The second part is a method which predicts the tertiary structure of protein molecules. This chapter mainly discusses the significance, advantages and disadvantages of the algorithm and the method. Further development of our protein structure prediction method and possible solutions to improve the results are also depicted.

7.1 Comments on the protein secondary structure prediction algorithm

7.1.1 Advantages and disadvantages

Our prediction algorithm for protein secondary structure is based on the Chou and Fasman method. Therefore, this new algorithm still reserves the advantages inherited from that method; that is, fast in speed, simple in theory and good results in prediction. Furthermore, many rules are also added to our algorithm. The purpose is to reduce the errors inherited from the original Chou and Fasman method.

There are at least four amino acids to form a helical structure. Similarly, no less than three amino acids are required to form the sheet and turn structures. The ineligible helical, sheet and turn structures previously mentioned are eliminated by our additional rules. If a protein molecule comprises of helical or sheet structure only, either helical or sheet structure are to be predicted. The advantage of this rule not only solves the regional overlapping problem in the Chou and Fasman method, but also increases the accuracy in prediction.

7.1.2 Further development

Though the prediction results of our algorithm are still not good enough, it is not a dead end. Some possible techniques can be used to improve the prediction. First, the parameters of the Chou and Fasman method are generated by the statistical assumption. These parameters come from not more than 30 protein molecules. There may be bias towards some amino acids due to insufficient data. As more tertiary structures of many protein molecules have been discovered by the X-ray crystallography for ten years, these parameters should be updated in order to minimize the undesirable bias.

Moreover, the parameters are obtained from the observation of the behaviour of a single amino acid. Garnier and his colleagues find that their prediction results increase slightly if the behaviour of pairs of amino acids are considered [24,25]. According to this research direction, three or more consecutive amino acids in a protein sequence may provide more constraints in a protein molecule folding. These constraints may offer a new potential to increase the

prediction accuracy.

7.2 Discussion on X-ray crystallographic data

The coordinates of atoms in a protein molecule are obtained by the X-ray crystallography. These data are now used to assess the results of a predictive scheme in protein tertiary structure prediction. However, some scientists regard that the X-ray crystallographic date cannot reflect the native structure of a protein molecule. They believe that the native structure of a protein molecule changes from time to time.

On the other hand, many scientists believe the X-ray crystallographic data. It is because they observe the followings :

a. A protein molecule can perform its function in a crystal state.

b. Atoms in a crystal state still vibrate in a small range.

c. There are water molecules inside a protein crystal.

Therefore, to a certain limit, the X-ray crystallographic data provide a reasonably good three dimensional structure information of a protein molecule. Nevertheless, the X-ray crystallographic data is the only information to assess different prediction algorithms.

7.3 Comments on the protein tertiary structure prediction algorithm

7.3.1 Advantages and disadvantages

Our algorithm in protein tertiary structure prediction is based on the theory of geometry. The basic structural unit in protein molecules is a peptide bond. Some atoms in peptide bond rest on a same plane. Due to this special characteristic, therefore, using geometrical theories, the relative positions of atoms on this plane can be determined. Furthermore, in a protein secondary structure, the rotation angle between two peptide planes is well defined. The coordinates of other atoms in this secondary structure can thus be deduced. In this way, this algorithm can be easily understood and implemented.

The state of some amino acids in protein molecules are classified as random coil. The rotation angle between two peptide planes in random coil structure cannot be well determined. Hence, a random factor is employed in our algorithm to generate a random rotation angle value. Although it is not a good way to solve this problem, it is better than assigning an arbitrary value. Furthermore, our algorithm also considers the repulsive force between two atoms. If the distance between two atoms is too close, the rotation angle between two peptide planes is modified so that atom collision is impossible.

Our algorithm has a good theoretical basis. Besides, another good point is the performance. The tertiary structure of a protein molecule with 200 amino acids can be constructed within 6 minutes. It is much faster than the energy minimization method.

However, our algorithm in protein tertiary structure prediction still has weakness. Atoms in side chain group would not form covalent bonds with other atoms except the sulphide atoms. Two sulphide atoms in some cases can form a disulphide linkage so that protein molecules becomes more stable. Although disulphide linkage is a valuable information in protein structure, our algorithm at this stage cannot handle this situation.

7.3.2 Further development

7.3.2.1 Rotation angle between two peptide planes

First, the subtended angle between side chain and main chain of a protein molecule can be considered to refine the coordinates of atoms in a protein molecule. Islam and his colleagues find that there is a relationship between side-chain conformations and the secondary structures in globular protein molecules [51]. The side chain group of an amino acid, which forms a secondary structure, always makes a particular angle to the main chain. This angle can help us to adjust the positions of atoms in the side chain group. Furthermore, the main chain of a protein molecule is determined by the planes of peptide bond. Though the rotation angle between two peptide planes has been determined in helical, sheet and turn structure, we still cannot find the rotation angle between two peptide planes in

random coil structure. Thus, to determine the rotation angle between two peptide planes in random coil may be a critical step to solve the weakness in this predictive scheme.

Reference :

(1) Abarbanel R. M., Cohen F. E., Fletterick R. J., Kuntz I. D., Secondary structure assignment for α/ß proteins by a combinatorial approach. (1983) Biochemistry 22, pp. 4894 - 4904

(2) Berendsen H. J. C., Dynamics simulation as an essential tool in molecular modeling. (1988) J. Computer-Aided Molecular Design 2, pp. 217 - 221

(3) Bing D. H., Caporale L.H., Feldmann R. J., Furie R. J., Furie B. C., Potter M., Mainhart C., On the construction of computer models of proteins by the extension of crystallographic structures. (1985) Ann. N. Y. Acad. Sci. 439, pp. 12 - 43

(4) Bohr H., Bohr J., Brunak S., Cotterill R. M. J., Lautrup B., Norskov L., Olsen O. H. and Petersen S. B., Protein secondary structure and homology by neural networks. (1988) FEBS Letter 241, pp. 223 - 228

(5) Bruijn D. P., Hays G. R., Symposium overview the shell conference on computer-aided molecular modelling. (1988) J. Computer-Aided Molecular Design 2, pp. 165 - 178

(6) Catlow C. R. A., Strategies for modelling of catalysts. (1988) J. Computer-Aided Molecular Design 2, pp. 255 - 258

(7) Chou P. Y. and Fasman G. D., Conformation parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. (1974) Biochemistry 13, pp. 222 - 245

(8) Chou P. Y. and Fasman G. D., Prediction of the secondary structure of proteins form their amino acid sequence : Advances in Enzmology. Vol. 47 (1978). John Wiley & Sons, New York, pp. 45 - 148

(9) Chou P. Y. and Fasman G. D., Prediction of protein conformation. (1974) Biochemistry 13, pp. 222 - 245

(10) Clarke B., A molecular graphics suite of programs for a microcomputer to display molecules from Carmbridge Crystallo-Graphic data files and the alpha-carbon backbone of proteins from protein data bank crystal files. (1988) Comput. Chem. 12 No. 1, pp. 65 - 82

(11) Cohen F. E., Clardelli T. L., Epstein L. B., Kosen P. A., Kuntz I. D., Smith K. A., Structure-activity studies of interleukin-2. (1981) Science 234, pp. 349 - 352

(12) Cohen F. E., Sternberg M. J. E., On the prediction of protein structure : The significance of the root-mean-square deviation. (1980) J. Mol. Biol 138, pp. 321 - 333

(13) Connolly M. L., Ferrin T. E., Kuntz I. D., Langridge R., Real-time color graphics in studies of molecular interactions. (1981) Science 211, pp. 661 - 666

(14) Connolly M. L., An application of algebraic topology to solid modeling in molecular biology. (1987) The visual Computer 3, pp. 72 - 81

(15) Craik C. S., Fletterick R., Rutter W. J., Splice junctions : Association with variation in protein structure. (1983) Science 220, pp. 1125 - 1129

(16) Dayhoff M. O., Atlas of protein sequence and structure. (1978) National Biomedical Research Foundation, Washington.

(17) Dearing A., Computer-Aided molecular modelling : Research study or research tool? (1988) J. Computer-Aided Molecular Design 2, pp. 179 - 189

(18) Dixon J. S., Kuntz I. D., Sheridan R. P., Scott K. P., Venkataraghavan R., Amino acid composition and hydrophobicity patterns of protein domains correlate with their structures. (1985) Biopolymers 24, pp. 1995 - 2023

(19) Dollittle R. F., Feng D. F., Johnson M. S., Align acid sequences : Comparison of commonly used methods. (1985) J. Mol. Evol. 21, pp. 112 - 125

(20) Doolittle R. F., Similar amino acid sequence : chance or common ancestry? (1981) Science 214, pp. 149 - 159

(21) Dufton M. J. and Hider R. C., Snake toxin secondary structure predictions. (Structure activity relationships) (1977) J. Mol. Biol. 115, pp. 177 - 193

(22) Ermacora M. R. and Rivero J. L., Secondary structure prediction of 11 mammalian growth hormones. (1988) Int. J. Peptide Protein Res. 32, pp. 223 - 229

(23) Ezzell B., Graphics Programming in turbo C 2.0 (1989) Addison Wesley Publishing Company, Inc., USA.

(24) Fasman G. D. ed., Prediction of protein structure and the principles of protein conformation (1989) Plenum Press, New York.

(25) Fischer M. A., Review letters. (1979) Computer graphics 13, pp. 234 - 236

(26) Feldmann R. J., The design of computing systems for molecular modeling. (1976) Annu. Rev. Biophy. Biogen. 5, pp. 477 - 510

(27) Feldmann R. J., Bing D. H., Potter M., Mainhart C., Furie B., Furie B. C. and Caporale L. H., On the construction of computer models of proteins by the extension of crystallographic structures : Macromolecular structure and specificity : Computer-assisted modeling and application. Ann. N. Y. Acad. Sci., (1985) Vol. 439, New York, pp. 12 - 43

(28) Foley J. D., Dam A. V., Feiner S. K., Hughes J. F., Computer Graphics : Principles and practice 2nd ed. (1990) Addison Wesley Publishing Company, Inc. USA.

(29) Friedland P., Kedes L., Discovery the secrets of DNA. (1985) Computer 10 No. 11, pp. 49 - 69

(30) Fruhbeis H., Klein R., Nallmeier H., Computer-assisted molecular design (CAMD) -- An overview. (1987) Angew. Chem. Int. Ed. Engle. 26, pp. 403 - 418

(31) Fuchs H., Goldfeather J., Hultquist J. P., Spach S., Austin J. D., Brooks F. P., Eyles J. G., Poulton J., Fast spheres, shadows, textures, transparencies, and image enhancements in pixel-planes. (1985) Proceedings of ACM SIGGRAPH, pp. 111 - 120

(32) Garnier J., Gibrat J. F. and Robson B., Further developments of protein secondary structure prediction using information theory. (New parameters and consideration of residue pairs) (1987) J. Mol. Biol. 198, pp. 425 - 443

(33) Garnier J., Osguthorpe D. J. and Robson B., Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. (1978) J. Mol. Biol. 120, pp. 97 - 120

(34) Harrington S., Computer graphics a programming approach 2nd ed. (1988) Mcgraw Hill international editions, Singapore.

(35) Hopp T. P. and Woods K. R., Prediction of protein antigenic determinants from amino acid sequences. (1981) Proc. Natl. Acad. Sci. USA 78, pp. 3824 - 3828

(36) Hopp T. P., Protein surface analysis methods for identifying antigentic determinants and other interaction sites. (1986) J. Immun. Methods 88, pp. 1 - 18

(37) Kaiser E. T., Kezdy F. J., Secondary structures of proteins and peptides in amphiphilic environments. (A review) (1983) Proc. Natl. Acad. Sci. 80, pp. 1137 - 1143

(38) Kabsch W. and Sander C., How good are predictions of protein secondary structure. (1983) FEBS Letter 155, pp. 179 - 182

(39) Karplus M., Molecular dynamics simulations of proteins. (1987) Physics Today, pp. 68 - 72

(40) Kubota Y., Nishikawa K., Takahashi S. and Ooi T., Correspondence of homologies in amino acid sequence and tertiary structure of protein molecules. (1982) Biochim. Biophys. Acta 701, pp. 242 - 252.

(41) Kyte J. and Doolittle R. F., A simple method for displaying the hydropathic character of protein. (1982) J. Mol. Biol. 157, pp. 105 - 132

(42) Lathrop R. H., Smith T. F., Webster T. A., ARIADNE : pattern-directed inference and hierarchical abstraction in protein structure recognition. (1987) Comm. ACM 30 No. 11, pp. 909 - 921

(43) Lathrop R. H., Smith T. F., Webster T. A., Pattern descriptors and the unidentified reading 2 mtDNA dinucleotide-binding site. (1988) Proteins : Structure, Function, and Genetics 3, pp. 97 - 101

(44) Lenstra J. A., Hofsteenge J. H. and Beintema J. J., Invariant features of the structure of pancreatic ribonuclease. (A test of different predictive models) (1977) J. Mol. Biol. 109, pp. 185 - 193

(45) Leung K. N., Leung S. O., Yeung H. W., The immunomodulatory and antitumour activities of trichosanthin - An abortifacient protein isolated from Tian-Hua-fen (*Trichosanthes Kirilowii*). (1986) Asian Pacific J. of Allergy and Immunology 4, pp. 111 - 120

(46) Levinthal C., The formation of 3D biological structures : Computer uses and future needs. (1984) Annu. N. Y. Acad. Sci. 426, pp. 171 - 180

(47) Levitt M., Warshel A., Computer simulation of protein folding. (1975) Nature 253, pp. 694 - 698

(48) Lim V. I., Structural principles of the globular organization of protein chains. A sterochemical theory of globular protein secondary structure. (1974) J. Mol. Biol. 88, pp. 857 - 872

(49) Lim V. I., Algorithm for prediction of alpha-helical and beta-structural regions in globular proteins. (1974) J. Mol. Biol. 88, pp. 873 - 894

(50) Max L. N., Spherical harmonic molecular surface. (1988) IEEE Computer Graphics & Application 8 No. 7, pp. 42 - 49

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

(52) Moews P. C. and Knox J. R., Predicted secondary structures of four penicillin beta-lactamases and a comparison with lysozymes. (1979) Int. J. Peptide Protein Res. 13, pp. 385 - 393

(53) Murata M., An efficient algorithm for comparing two protein sequences : implementation for microcomputers. (1988) Comput. Chem. 12 No. 1, pp. 21 - 25

(54) Nakashima H., Nishikawa K. and Ooi T., The folding type of a protein is relevant to the amino acid composition. (1986) J. Biochem. 99, pp. 153 - 162

(55) Neal M., Getting inside molecules. (1985) IEEE Computer Graphics & Application 5 No. 10, pp. 8 - 14

(56) Nishikawa K., Assessment of secondary-structure prediction of proteins comparison of computerized Chou-Fasman method with others. (1983) Biochim. Biophys. Acta 748, pp. 285 - 299

(57) Nishikawa K. and Ooi T., Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods. (1986) Biochim. Acta 871, pp. 45 - 54

(58) Ondetti M. A., Cushman D. W., Sabo E. F. and Cheung H. S., The design of active-site-directed reversible inhibitors of exopeptidase, Kalman, ed. Drug action and design : Mechanism-based enzyme inhibitors (1979) Elsevier North Holland, Inc., pp. 271 - 287.

(59) 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. (1987) J. Mol. Biol. 193, pp. 775 - 791

(60) Porter T. K., Spherical Shading, (1978) Computer Graphics 12, pp. 282 - 285

(61) Ready M. P., Robertus J. D., Ricin B chain discoidin I share a common primitive protein fold. (1984) J. Biol. Chem. 259, pp. 13953 - 13956

(62) Rich E., Artificial Intelligence, McGraw-Hill, New York, 1988

(63) Richardson J. S., The anatomy and taxonomy of protein structure : Advances in protein chemistry (C. B. Anfinsen, J. T. Edsall, F. M. Richards ed.) (1981) Academic Press. New York. Vol. 34. pp. 167 - 339

(64) Rose G. D., Prediction of chain turns in globular proteins on a hydrophobic basis. (1978) Nature 222, pp. 586 - 590

(65) Rose G. D., Gierasch L. M. and Smith J. A., Turns in peptides and proteins : Advances in protein chemistry (1985) Vol. 35 Academic Press, London, pp. 1 - 109

(66) Scherage H. A., Calculations of the three-dimensional structures of proteins : Macromolecular structure and specificity : Computer-assisted modeling and application. Ann. N. Y. Acad. Sci., (1985) Vol. 439, New York, pp. 170 - 194

(67) Scheraga H. A., Saito N., Wako H., Statistical mechanical treatment of α-helics and extended structures in proteins with inclusion of short- and medium-range interactions. (1983) J. Protein Chem. 2 No. 3, pp. 221 - 249

(68) Stefik M., Planning with constraints (MOLGEN : part 1) (1981) Artificial Intelligence 16 No. 2, pp. 111 - 139

(69) Stefik M., Planning and meta-planning (MOLGEN : part 2) (1981) Artificial Intelligence 16 No. 2, pp. 141 - 169

(70) Stryer L., Biochemistry. (1981) W. H. Freeman and company, New York.

(71) Thornton J. M. and Taylor W. R., Structure prediction : Protein sequencing (a practical approach), ed. J. B. C. Findlay and M. J. Geisow, IRL Press, Eynsham, Oxford, England, pp. 147 - 190

(72) Ulmer K. M., Protein engineering. (1983) Science 219, pp. 666 - 671

(73) Wells J. A., Powers D. B., Bott R. R., Katz B. A., Ultsch M. H., Kossiiakoff A. A., Power S. D., Adams R. M., Heyneker H. H., Cunningham B. C., Miller J. V., Graycar T. P. and Estell D. A., Protein engineering of subtilisin, J. A. Welles *et al.* Protein Engineering (1987) Alan R. Liss, Inc., pp. 279 - 287

(74) Wodak S. J., Computer-aided design in protein engineering. (1987) Annu. N. Y. Acad. Sci. 501, pp. 1 - 13

(75) Yada R. Y., Jackman R. L. and Nakai S., Secondary structure prediction and determination of proteins -- a review. (1988) I Protein Res. 31, pp. 98 - 108

(76) Zhang X., Wang J., Homology of trichosanthin and ricin A chain. (1986) Nature 321, pp. 477 - 478

Glossary

Conformation :
The overall three dimensional structure of a molecule. In general, the conformation of a native molecule always changes.

DNA :
Deoxyribonucleic acid (DNA). The basic unit carries the inherent information.

Enzyme :
Specific proteins which catalyse biological and chemical reactions.

Globular protein :
The conformation of a protein likes a sphere. This kind of protein always dissolves in water.

Homology :
Proteins which came from a common ancestor are called homology protein.

Insulin :
A kind of protein which converts glucose into glycogen.

Membrane protein :
Protein molecules which are embedded in membrane.

Mutant :
Bacterion, virus or substance which changes its nature by x-ray or chemicals is called the mutant of its original species.

Polypeptide :
Many amino acids are joined together by polypeptide bond. This amino acids sequence is called polypeptide chain or protein primary structure.

Reactant :
A substance reacts with enzyme.

Vaccine :
A kind of substance which can kill bacteria or virus.

Appendix A

Algorithm in hydrophobic value determination :

O_H[n] : hydrophobic value of an amino acid
F_H[n] : hydrophobic value of an amino acid in a protein molecule

```
Begin
        read amino_acid_no;
        read window_size;   && (positive odd number)
        half_window_size = (window_size - 1) / 2
        start_position = half_window_size
        end_position = amino_acid_no - start_position

        For i := start_position to end_position {
                temp = 0
                For j := (i - half_window) to (i + half_window) {
                        temp = temp + O_H[j]
                }
                F_H[i] = temp / window_size
        }
End
```

F_H[i] finally contains the hydrophobic value of an amino acid in a definite protein.

Appendix B

Algorithm of the Chou and Fasman method

Helical structure prediction

```
Begin
        Read amino_acid_no
        start_position = 1
        end_position = amino_acid_no - 5
        For i := start_position to end_position {
                helical_former = 0
                probability = 0
                proline = 0
                For j := i to i+5 {
                        helical_former = helical_former +
                                                state_of(amino_acid[j])
                        probability = probability + probability_of(
                                                amino_acid[j])
                        proline = proline + is_proline(amino_acid[j])
                }
                if helical_former >= 4 and proline >= 1 and
                  probability >= 1.03 {
                        from i to i+5 mark as helical structure
                }
        }
End

state_of(amino_acid[j])
if amino_acid[j] = strong_helical_former or helical_former {
        return 1
}
if amino_acid[j] = weak_helical_former {
        return 0.5
}

probability_of(amino_acid[j])
return (the probability that forms a helical structure of the amino acid[j])

is_proline(amino_acid[j])
if amino_acid = proline
        return 1
else
        return 0
```

```
Sheet structure prediction
Begin
        Read amino_acid_no
        start_position = 1
        end_position = amino_acid_no - 4
        For i := start_position to end_position {
                sheet_former = 0
                probability = 0
                proline = 0
                For j := i to i+4 {
                        sheet_former = sheet_former +
                                                state_of(amino_acid[j])
                        probability = probability + probability_of(
                                                amino_acid[j])
                }
                if sheet_former >= 3 and probability >= 1.00 {
                        from i to i+3 mark as sheet structure
                }
        }
End

state_of(amino_acid[j])
if amino_acid[j] = strong_sheet_former or sheet_former {
        return 1
}
if amino_acid[j] = weak_sheet_former {
        return 0.5
}

probability_of(amino_acid[j])
return (the probability that forms a sheet structure of the amino acid[j])
```

```
Turn structure prediction
Begin
        read amino_acid_no
        start_position = 1
        end_position = amino_acid_no_3

        i = start_position
        while (i <= end_position) {
                temp = 1
                For j := i to i+3 {
                        temp = temp * probability_of(amino_acid[j])
                If temp > 0.75 * 10-4 {
                        from i to i+3 mark as turn structure
                        i = i + 3
                }
                i = i + 1
        }
End

probability_of(amino_acid[j])
return (the probability that forms turn structure of the amino acid[j])
```

Appendix C

Algorithm of the GOR method

S[n] : probability of an amino acid in helical structure formation
B[n] : probability of an amino acid in sheet structure formation
T[n] : probability of an amino acid in turn structure formation
C[n] : probability of an amino acid in random coil structure formation
F[n] : final predicted state of an amino acid

```
window_size = 17
half_window_size = 8

Begin
        read amino_acid_no;
        start_position = 0
        end_position = amino_acid_no

        For i := start_position to end_position {
                temp_S = 0
                temp_B = 0
                temp_T = 0
                temp_C = 0
                For j := (i - half_window) to (i + half_window) {
                        if (j > 0 and j < amino_acid_no) {
                                temp_S = temp_S + S[amino acid[j]]
                                temp_B = temp_B + B[amino acid[j]]
                                temp_T = temp_T + T[amino acid[j]]
                                temp_C = temp_C + C[amino acid[j]]
                        {
                        F[i] = Max(temp_S,temp_T,temp_B,temp_C)
                }
        }
End


Max(A,B,C,D)
return (the maximum value of these four values)
```

Appendix D

The shading algorithm of sphere is listed as follows

A sphere with radius R and canter (a,b,c), the circle in xy plane can be written in the form

$$g(x,y) = Ax + By + C - Q = 0 \qquad (1)$$

where $A = 2a$, $B = 2b$, $C = R^2 - a^2 - b^2$ and $Q = x^2 + y^2$. Since the symmetrical property of sphere, the upper hemisphere that $Z > z$ can be seen by the user. The Z value of upper hemisphere can be written as

$$z = c - \sqrt{(R^2 - (x - a)^2 - (y - b)^2)} \qquad (2)$$

The equation (2) can be approximated as

$$z = c - (R^2 - (x - a)^2 - (y - b)^2) / R \qquad (3)$$

The hemisphere becomes a paraboloid. Now let the unit vector of normal N of a point at (x,y,z) of the sphere with radius R at canter (a,b,c) will become as

$$N = (1/R)(x - a, y - b, z - c)$$

$$= (1/R)(x - a, y - b, -\sqrt{(R^2 - (x - a)^2 - (y - b)^2)}) \qquad (4)$$

A light source L at infinity with unit vector (l_1,l_2,l_3). The maximum highlight on the sphere is (Rl_1+a, Rl_2+b, Rl_3+c). If the maximum colour intensity is denoted as CMAX and the minimum colour intensity is denoted as CMIN. Then the colour intensity at point (x,y,z) is

$$\text{Colour}(x,y,z) = \begin{cases} \text{CNIN} + (\text{CMAX} - \text{CMIN})(L \cdot N), & \text{if } L \cdot N \geq 0; \\ \text{CMIN}, & \text{if } L \cdot N < 0. \end{cases} \qquad (5)$$

Using the parabolic approximation of the hemisphere in (3), the L·N can be approximated as

$$L \cdot N \approx (l_1(x - a) + l_2(y - b)) / R$$
$$- l_3(R^2 - (x - a)^2 - (y - b)^2) / R^2 \qquad (6)$$

Since the screen is two dimension, the Z-axis is perpendicular to the screen surface. Hence, the Colour at point (x,y,z) can be written as

$$Colour(x,y) = K(Ax + By + C - Q) + CMIN \qquad (7)$$

where

$$K = - (CMAX - CMIN) / R^2,$$
$$A = - l_1R + l_32a,$$
$$B = - l_2R + l_32b,$$
$$C = l_1Ra + l_2Rb + l_3(R^2 - a^2 - b^2),$$
$$Q = x^2 + y^2.$$

From equation (2) to equation (3), there is an approximation from square root to division. Although the accurate of result is decreased, the performance will be increased.