# Introduction to the Protein Structure Prediction using the Global Optimization

Jinwoo Lee [1]

1) *Department of Mathematics, Kwangwoon University, Seoul 139-701, KOREA*

Corresponding Author : Jinwoo Lee, e-mail : jinwoolee@kw.ac.kr

## INTRODUCTION

Proteins are essential parts of all living organisms and participate in every process within cells, for example, in biochemical reactions, cell signaling, and immune responses. Their cellular functions are carried out through structural charicteristics. Thus the revealing the structure of proteins can provide important clues to understand functions and mechanisms of life.

Proteins are large molecules made of amino acids arranged in a linear chain. The sequence of amino acids in a protein is uniquely specified by nucleotides encoded in a gene. In the 1960's, C. B. Anfinsen established the "Thermodynamic Hypothesis", which states that the native conformation of a protein is uniquely determined by its amino acid sequence. Since then, researchers have made efforts to determine protein structures from their sequences.

One way to determine the structure of a protein is to use experimental methods like X-ray crystallography, or NMR spectroscopy. In order to do that a protein shoud be purified as a soluble form, or crystalized. However, it takes at least several months to several years. In many cases, like membrane proteins, it is almost impossible, and thus only few membrane strucures are determined.

Protein structure prediction seeks to develop computational methods to determine the native structure of a protein of which structure has not been determined experimentally. The prediction is categorized as two types, template based modeling(TBM) and free modeling(FM). TBM build models based on "template" structures with sequence similarity to the protein being modeled. If template structures could not be found, we should try FM.

The first difficult step in TBM is to find appropriate templates which share structural similarity through the sequence of the protein being modeled. The second step, and the main challenge in TBM is to align template sequences and a query sequence simultaniously, so called multiple sequence alignment(MSA).

Performing MSA is classified as an NP-hard, combinatorial optimization problem. Thus most practical methods use heuristic algorithms, like progresive alignment. It aligns two sequences first, and add one by one. By its construction, errors occured initially can not be fixed. Many variants to overcome the demerit are proposed, but the demerit can not be resolved essentially.

In every step of protein structure prediction, we should define a score function, and optimize it. It is again a difficult combinatorial optimization problem. For example, consider a protein composed of 100 amino acids, and suppose that one amino acids can form 10 different conformations. Then total number of conformations is $10^{100}$. Thus finding global minimum

conformation in such a huge space is quite challenging. Moreover, we do not know the exact energy function to be optimized.

Prof. Jooyoung Lee proposed a global optimization method called conformational space annealing (CSA). The CSA adopts advantages of genetic algorithms, and simulated annealing, a popular global optimization method. It searches the whole conformational space initially, and gradually concentrates on regions with low energy maintaining the conformational diversity.

In recent CASP7, we proposed a new prediction method based on the CSA. In every stage of the prediction, consistency-based energy functions are defined, and they are systematically optimized by using the CSA. Submitted models by the method ranked in top group in TBM and high accuracy template based modeling categoriies (TBM/HA). In this talk, I will introduce these things in more detail.


## AKNOWLEDGEMENT

I would like to thank Prof. Jooyoung Lee, our group leader and Dr. Keehyoung Joo, a main member of our group in KIAS.