Summary of Dissertation

Title: Prediction of the folding processes for immunoglobulin-like beta-sandwich and beta-trefoil proteins

Doctoral Program in Advanced Life Sciences
Graduate School of Life Sciences
Ritsumeikan University

アウンプチン パンヤブット
AUMPUCHIN Panyavut

Abstract

Describing the folding mechanism of a protein is one of the goals in the molecular bioinformatics study. Since a protein folds into its specific 3D structure from a unique amino acid sequence, it is interesting to extract as much information as possible from the amino acid sequence. The main purpose of this thesis is to extract the folding mechanisms from an amino acid sequence, especially the structure formation in an initial state of folding, by means of inter-residue average distance statistics methods, Average Distance Map (ADM) and contact frequency analysis (F-value). Moreover, a coarse-grained Gō-model simulation was applied for some proteins to simulate the whole stories of protein folding. The proteins used in this thesis are immunoglobulin (Ig)-like beta-sandwich and beta-trefoil proteins. The results of both proteins extracted from only the amino acid sequences coincided well with the coarse-grained Cα Gō-model simulation results as well as available experimental data. That is, the β strands 3 and 5 are suggested to be the significant strands for the initial folding processes of the Ig-like beta-sandwich proteins, which emphasize the significance of the common structure, namely, the key strands for folding and the Greek-key motif. On the other hand, β strand 5 of a beta-trefoil protein plays an important role in structural construction, according to the result of a conserved hydrophobic residues on β strand 5 is always formed packing with other conserved hydrophobic residues in both regular and irregular beta-trefoil proteins. In summary, it can be confirmed that the average distance statistics methods were able to predict the initial folding event and that a coarse-grained Cα Gō-model simulation can be used to investigate the whole process of protein folding, which corresponded well to available experimental data.

Chapter 1: General Introduction

Proteins are the end products of the decoding process from cellular DNA in all living organisms. Protein folding is a physical process of a polypeptide chain which folds into their biologically active structure. The correct final folded native structures of proteins are key to their functions. Therefore, misfolding can lead to inactive or toxic proteins, and a protein aggregates that cause severe diseases. To study the protein folding, the native structure of each protein is the most important one that requires as input data in various methods. Even though, the native folded structures are complex, but are known in great detail thank to NMR or X-ray crystallography. The multiple conformations of an active protein structure are hinged on the amino acid sequence compositions. However, the mechanisms of protein folding have not been discovered so far.

Chapter 2: Investigation Methods

According to the question how the amino acid sequence determines one particular fold instead of another and how to extract folding properties from only that sequence information. Therefore, the present average distance statistics methods,[1] ADM and F-value, are performed to extract a compact region and residue which frequently form contacts in an initial folding event of target proteins by using only its amino acid sequence information. These methods have been confirmed in previous studies which successfully extract the folding properties that coincide well with the data from experimental analyses. Then, an evolutionary analysis is performed by using a multiple sequence alignment to detect the conserved compact regions and the conserved hydrophobic residues. Finally, a 3D-based coarse-grained Cα Gō model simulation[2] was conducted to investigate the protein folding processes as well as the whole story of protein folding.

Chapter 3: Study of folding mechanisms for immunoglobulin-like beta-sandwich proteins

The initial folding processes of Ig-like beta-sandwich proteins have been investigated in this study by means of inter-residue average distance statistic methods, ADM, and F-value analysis as well as by 3D-based Gō-model simulation. In this present study, the initial folding unit of titin protein, including the Ig domains and FN3 domains, was investigated. Furthermore, an evolutionary analysis also performed by using sequence-based and structure-based multiple sequence alignments for every domain in the titin chain. The conservation of predicted folding units and hydrophobic residues was studied. The present Gō-model simulation studies confirm that the central regions of six samples from Ig domains and FN3 domains were predicted to be an initial folding unit that forms the compact structure in the early event. This common feature is in line with the available experimental ϕ-value and protection factor derived from H-D exchange experiment of 1TIT and 1TEN, which also detected the stability of the central unit and the fluctuation of both terminal ends. Interestingly, the results underscore the importance

of the common structure of these proteins, in particular, the key strands for folding and the Greek-key motif. Moreover, the difference in the folding pathways and the whole story of the protein folding processes can be described by the present Gō-model simulations.

Chapter 4: Study of folding mechanisms for beta-trefoil proteins

The beta-trefoil proteins, including symmetric and irregular structures, were selected as the target proteins. Information on the location of conserved hydrophobic residues in combination with the results of ADM and an F-value plot reveals the significant residues for hydrophobic packing during folding. Almost every β-strand contains one or two conserved hydrophobic residues and these equally distributed hydrophobic residues seem to be significant to form the symmetrical beta-trefoil fold. The conserved hydrophobic residues in trefoil unit-2 may be more significant, according to the conserved residues at position β5N, β5C and β6 detected near the highest or second highest peak of the F-value plot in a protein from almost every superfamily. The compact regions of beta-trefoil unit-1 and unit-3 derived from ADM analyses tend to conserve among different superfamilies. The Gō-model simulations results of irregular beta-trefoil proteins in an initial state of folding coincide well with the predictions made by the ADMs and F-value analyses. Furthermore, these results also correlated to the experimental results of symmetric beta-trefoil proteins, that is, the irregular structures are not affected to the main folding mechanisms.

Conclusion

The initial folding processes of target proteins have been investigated by means of inter-residue average distance statistic methods, ADM and F-value analyses, as well as by 3D-based coarse-grained Cα Gō-model simulation. It is interesting that the sequence-based methods can capture changes in the folding route of a protein from that of another structurally similar protein and corresponded well to the available experimental information. Moreover, the whole story of protein folding can be descripted by the present Gō-model simulation. Finally, according to these results, it can confirm an ability of the present sequence-based techniques to decode the folding mechanisms, especially an initial state of protein folding.

References

1. Kikuchi T, Némethy G, Scheraga HA (1988) Prediction of the location of structural domains in globular proteins. *Journal of protein chemistry.* 7(4):427-471.

2. Sugita M, Kikuchi T (2013) Incorporating into a Cα Go model the effects of geometrical restriction on Cα atoms caused by side chain orientations. *Proteins: Structure, Function, and Bioinformatics.* 81(8):1434-1445.