論 文 の 内 容 の 要 旨

論文題目　Improving *de novo* model quality and its application in *ab initio* phasing

（de novo構造予測の改善とそのab initio位相決定への応用）

氏　　名　　ロジャン　シュレスタ

De novo models are computationally predicted three-dimensional models of the given proteins using only amino acids sequence information. The key components of *de novo* modeling are the methods responsible for conformational space searching and the evaluation of each conformation accurately using energy function. The conformational space is astronomically large due to the degrees of freedom associated with each residue, which creates the challenge to develop the efficient method for searching the conformational space. Another challenge in *de novo* modeling is to devise an accurate energy function to evaluate the conformers. Despite these challenges, the *de novo* modeling has succeeded to generate accurate models for small and single domain proteins in the blind prediction. Fragment assembly is an effective and efficient method for *de novo* modeling. This method assembles the fragments from known structures under the guidance of energy function. This concept was practically implemented in Rosetta, which achieved a number of break-through successes. Rosetta has two major stages, which are termed as coarse-grained sampling and all-atom refinement, to generate the final structure from the input amino acids sequence. At the initial stage, three-residue and nine-residue fragments obtained from known structures are assembled to generate full-length coarse-grained models. These models contain only backbone atoms and the centroid of side-chain atoms. Subsequently, side-chain atoms were packed to construct all-atom models followed by energy minimization in all-atom refinement. However, there exist many challenges in the prediction of accurate models needed for practical use such as solving the crystallographic phase problem. To address these issues, I have focused on developing the methods – biased conformation sampling and fragment quality improvement to enhance the quality of predicted models. Furthermore, when full-length *de novo* models are difficult to improve, I have developed the method to use the fragments to solve the phase problem.

First, I developed a method to improve the conformational space search for accuracy improvement. This method first generated coarse-grained models using Rosetta. Second, an ensemble of lowest energy coarse-grained models was selected and deviation for each model from other models of the ensemble was calculated at the residue level. This score was called as average pairwise residue distant score. The score correlated with the accuracy of predicted residues in the model. When the predicted residues had larger scores, the residues were considered as less accurate and vice versa. Lastly, conformational search was biased using the score as residues with larger scores were given higher frequency for sampling. This procedure rebuilt selected coarse-grained models and then packed the side

chain atoms followed by energy minimization. Molecular replacement was run on these all-atom models and the entire simulation was terminated after a few correct solutions were obtained. Our method was tested on 10 difficult targets, which were failed to achieve the success in previous studies using other methods - Rosetta and RosettaX. The rebuilding procedure improved the accuracy of coarse-grained models from 4.93 Å to 4.06 Å on average. Seven out of ten protein targets showed successful molecular replacement solution using rebuilt models.

The second method focused on improving the fragment quality for the model accuracy improvement. This study developed a method to generate new fragment libraries using a resampling process. Therefore, it has multiple steps. This study first selected the lowest energy all-atom models generated using Rosetta at initial run. These models were broken into overlapping fragments of three-residue and nine-residue. Average pairwise residue deviation score was computed for both fragments to remove most distant fragments. The resultant fragments were clustered and then twenty-five fragments were randomly selected from the top five clusters. These new fragments were used for the new prediction in second run. The performance of the method was tested on a benchmark set of 30 different proteins. The accuracy of new fragments and predicted models was evaluated. This result showed that the new fragment library contained better fragments and enriched with many high-quality fragments. In order to evaluate the performance, the lowest energy models and one of best from top five models were taken as the best prediction and computed their root mean square deviation, template modeling score, and global distance test – total score to the native structures. In all these assessment criteria, this method performed significantly better than Rosetta for lowest energy models and best in top five models. On average, this method improved from 5.99 Å to 5.03 Å when lowest energy models were selected as the best predicted models. Similarly, it improved both, the template modeling score and global distance test - total score, by 7%.

Lastly, a new method was developed to tackle the phase problem using fragment assembly approach when the full-length models were inaccurate as the template models for molecular replacement. In this method, *de novo* models were fragmented, independently phased, and assembled together in asymmetric unit. Rosetta generated lowest energy all-atom models were chosen to generate the fragments. For each residue position, constant-length overlapping fragments were generated. These fragments were clustered and two hundred candidate fragments were randomly selected for each residue position from top twenty clusters. The selected fragments were independently used as search model in molecular replacement. The fragments were assembled together after molecular replacement. In fragment assembly process, I selected one fragment as a seed fragment and one low-energy *de novo* model as a reference model. The reference model was superposed to the seed fragment. Using the seed fragment and the reference model, position and orientation of other fragments were determined in the crystallographic unit cell. The combinations of permissible origins of space group, their symmetry operators, and unit cell translation were computed to identify the location of other fragments in the asymmetric unit. The combination that gave the smallest distant between the reference model and the candidate fragment was taken as the correct location. In this way, all the fragments were assembled in the asymmetric unit. This method was tested in ten difficult proteins. Indeed, the full models of these targets were unable to solve the phase problem in molecular replacement trials. Using the method, the crystal structures of eight protein targets were solved from a total of ten.