学位論文の要旨

学位論文題目： Differential Evolution Explores a Multiobjective Knowledge-based Energy Function for Protein Structure Prediction

（**和訳**：差分進化アルゴリズムが多目的な知識ベースのエネルギー関数を探索によるタンパク質構造予測研究）

専攻　数理・ヒューマンシステム科学

氏名　陳 星倩（チン　セイセン）
Xingqian Chen

Proteins are complex large organic macromolecules that are composed of one or more chains of 20 different amino acids in specific orders. Many fundamental biological functions in organisms are performed by proteins, such as structural support, material transport, and regulation functions. Since the structure of a protein determines its biological functions, knowledge of its native structures is essential for understanding its role in life activities. Three experimental methods, i.e., X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy, are commonly used to perform protein structure determination. However, all of these experimental methods are costly and waste much of time. On the other hand, since the three-dimensional structure of a protein is determined by its amino acid sequence, it is very meaningful for researchers to obtain the three-dimensional (3D) structure of the protein from its sequence by calculation methods.

How to predict the 3D structure of a given protein starting only from its amino acid sequence is called the protein structure prediction (PSP) problem. The high theoretical value and practical significance make the research of this problem necessary and promising. Despite the rapid development of computer techniques and the unremitting efforts of researchers, the PSP problem remains challenging in bioinformatics and computational biology. Numerous approaches have been proposed to solve the PSP problem. These approaches can be roughly grouped into two categories: template-based modeling and free modeling (FM). The two pivotal factors of a successful FM prediction approach are an efficient search strategy and an effective energy function.

Since the conformation search space is very large, an exhaustive search strategy is infeasible under normal circumstances. A successful FM must

employee efficient search strategies to find the global minimum of a given energy function. The most common conformation search method employed in FMs is the Monte Carlo algorithm or its variations. Recently, employing evolutionary computation techniques as the search strategy in FMs has attracted researchers' interest, and considerable success has been achieved. Protein energy functions are used to select more native-like conformations during the process of protein folding. The existing protein energy functions can be roughly classified into two groups: physics-based energy functions and knowledge-based energy functions.

In my research of defending PhD, I try to model the PSP problem as a multi-objective optimization problem and use an differential evolution search strategy to solve the problem. In details, the PSP problem is modeled as a multiobjective optimization problem, and a FM approach called MODE-K is proposed to solve this problem. my efforts center on two aspects. First, a knowledge-based energy function called RWplus is used as the evaluation criterion. This function is decomposed into two terms: an orientation-dependent energy term and a distance-dependent energy term. Second, a multiobjective differential evolution coupled with an external archive employed to perform conformation space searching. After conformation space searching, we introduce a cluster method to select the final predicted structure from series of decoy structures. The performance of the method was verified with eighteen test proteins. The experimental results demonstrate the effectiveness of the proposed method and indicate that incorporating knowledge-based energy functions into multiobjective approaches to solve the PSP problem is promising.

The contribution of this thesis is fourfold: first, the PSP problem is modeled as a multiobjective optimization problem and two knowledge-based energy terms are used to construct the energy function. Second, a new MODE algorithm that interacts with an external archive is proposed. Third, an integral work flow is provided. The clustering method which called MUFOLD-CL is used to identify the final predicted structure from a set of decoy structures that are stored in the archive. Fourth, eighteen test proteins categorized into three structural classes are used to evaluate the proposed method. More investigation of the experimental results provides evidence of the superior performance of the proposed approach.

【審査結果の要旨】

当博士学位論文審査委員会は，標記の博士学位申請論文を詳細に査読し，投稿された論文の査読プロセスを確認した。本博士論文と従来の論文との類似性指標は 18%であり，剽窃等の問題がないことを確認した。また論文公聴会を令和 4 年 2 月 3 日(木曜日)に公開で開催し，詳細な質疑応答を行って論文の審査を行った。以下に審査結果の要旨を記す。

タンパク質構造予測（PSP）問題は、アミノ酸配列をもとにタンパク質の三次元構造を予測する問題である。PSP の手法は、ab initio モデリングと比較モデリングの二つに大きく分類することができる。本研究では、多目的最適化に基づく、ab initio モデリング手法が提案された。その概要は、以下の通りである。

本研究では、PSP 問題を多目的最適化問題としてモデル化し、差分進化アルゴリズムを使用して問題の解決を試みた。具体的には、MODE-K と呼ばれる自由モデル化を行っている。評価基準として RWplus と呼ばれる知識ベースのエネルギー関数を使用する。この関数は、方向に依存するエネルギー項と距離に依存するエネルギー項の 2 つの項に分解される。そして、コンフォメーション空間検索を実行するために採用された外部アーカイブと組み合わせた多目的差分進化を提案している。空間探索の後、多くの構造から最終的な予測構造を選択するクラスター法を導入している。提案した手法の有効性を 18 種類のタンパク質で検証した。実験結果は、提案された方法の有効性を示し、知識ベースのエネルギー関数を多目的アルゴリズムに組み込む方法が PSP 問題に有効であることを示している。

本研究の学術的貢献は 4 つある。1） PSP 問題を多目的最適化問題としてモデル化し、エネルギー関数を構造するために 2 つの知識ベースのエネルギー項を提案した。2）外部アーカイブと相互作用する新しい MODE アルゴリズムを提案した。3）統合されたワークフローを提案さした。 MUFOLD-CL と呼ばれるクラスタリング手法を、アーカイブに保存されているおとり構造のセットから最終的な予測構造を識別するために使用した。4） 3 つの構造クラスに分類された 18 のテストタンパク質を使用して、提案された方法を評価し，良好な成績を得た。以上の実験結果から、提案された方法の優れた性能を実証した。

当博士論文審査委員会は、研究内容及び研究成果を慎重に吟味した結果、本博士学位申請論文が博士の学位を授与することに十分値するものと認め，合格と判断した。