学 位 論 文 内 容 の 要 旨/Dissertation Abstract

博士（環境科学）　　氏　名/XU SHIJIE

学 位 論 文 題 名/Title of Dissertation

Accurate and fast predictions of protein structures and properties via protein language models
（タンパク質言語モデルによるタンパク質の構造と特性の正確かつ迅速な予測）

The work presented in this thesis focuses on protein modeling using artificial intelligence. The field has become increasingly active in recent years due to the explosion of data produced by modern biological techniques. The huge amount of data undoubtedly stimulates the development of related fields and challenges the traditional methods used to process these data. Recent advances in deep learning have thus been introduced to reduce the gaps, but they require further exploration. The work presented in this thesis adopted supervised learning, unsupervised learning, transfer learning, and Bayesian optimization to understand the structures and properties of proteins, as well as their evolution. The primary aim of this thesis is to explore protein structures and properties through protein language models (PLMs). The specific problems addressed in the thesis are detailed below.

The introductory chapter discusses the foundational concepts of protein science and the importance of protein structure prediction, emphasizing challenges like protein complexity and data scarcity. Deep learning is highlighted as a promising but still evolving solution, with PLMs emerging as transformative tools that overcome traditional models' reliance on large, high-quality datasets. By leveraging both labeled and unlabeled data, PLMs offer a novel approach to protein modeling. The chapter also explores machine learning applications in phylogenetics, linking protein evolution to structure and function, and concludes with an outline of the thesis structure and its contributions.

Chapter 2 introduces an innovative, efficient, and accurate approach to predicting intrinsically disordered proteins (IDPs). IDPs, which lack stable structures, are crucial for understanding protein function and are associated with numerous human diseases. Traditional methods rely on multiple sequence alignments (MSAs) to predict IDPs, but these alignments are not universally available and necessitate exhaustive searches through ever-expanding protein sequence databases. The proposed PLM-based method replaces the laborious MSA reconstruction process with alternative information derived from PLMs. Furthermore, an ensemble learning strategy is employed to enhance prediction accuracy by combining multiple PLMs. A novel secondary structure predictor based on PLM is also introduced, further improving prediction accuracy. The functions of IDPs, including disordered linker regions and disordered protein-binding sites, are also predicted. These approaches not only achieve superior accuracy compared to state-of-the-art techniques but also operate at significantly faster speeds, making them ideal for genome-scale IDP predictions.

Chapter 3 presents a rapid and precise predictor for pK_a values. The pK_a values of ionizable groups in proteins are vital for understanding protein function and structure. However, numerous environmental factors, such as hydrogen bonding, ligand interactions, hydrophobicity, and electrostatics, can significantly alter pK_a values, complicating accurate prediction. Traditional methods often fail or produce inaccurate pK_a predictions, particularly for shifted pK_a values. This chapter introduces a novel method that uses only the PLM representation of a single residue as input. Two isoelectric point (pI) prediction methods based on bidirectional long short-term (BiLSTM) networks are also proposed to improve pK_a prediction, achieving state-of-the-art accuracy of 0.3393 RMSE (root mean square error). The method supports multiple ionizable groups, including the sidechains of Asp, Glu, His, Lys, Cys, Tyr, and N- and C-termini. Testing on challenging datasets demonstrates the method's superior performance, particularly for deeply buried residues with significantly shifted pK_a values. The proposed method also exhibits rapid prediction speeds, making it suitable for large-scale applications in structure-unknown proteins.

Chapter 4 introduces a novel method for identifying metal-binding sites in proteins. Metalloproteins, which play pivotal roles in cellular functions, have profound implications in fields such as industry, medicine, and environmental science. Existing sequence- or structure-based methods often fall short in terms of precision. This chapter presents a multi-step approach: a sequence model for initial binding residue identification, a probe generation algorithm for candidate site prediction, a structure model for site refinement, and a clustering algorithm to eliminate redundancy. Protein language models are used to quickly and accurately identify metal-binding residues from sequence data, which are then refined in subsequent steps. The proposed method is rigorously evaluated and compared with state-of-the-art techniques, demonstrating superior accuracy. It also achieves rapid prediction speeds, approximately twenty seconds per protein, faster than most existing structure-based methods. Additionally, the method's robustness is demonstrated on AlphaFold2-predicted structures, suggesting its potential for large-scale applications.

Chapter 5 introduces a new partitioning algorithm designed to improve the accuracy of phylogenetic inference through partitioned models. Biological sequence heterogeneity, encompassing proteins and nucleotides, is often poorly modeled. Traditional partitioning methods rely heavily on expert knowledge or heuristic approaches. This chapter introduces a machine learning-based algorithm that optimizes the partitioning process via parametrization and Bayesian optimization. The novel parameterized sorting indices (PSI) proposed by the author guides the partitioning process. The method reconstructs phylogenetic trees with significantly more accurate topologies on several simulated datasets. It also yields better topologies, and higher bootstrapping support on real datasets, aligning more closely with the morphological and geological evidence in the literature. The proposed method also demonstrates acceptable computational efficiency compared to traditional methods.

In conclusion, this thesis demonstrates the power of protein language models in predicting protein structures and properties, achieving superior accuracy and efficiency in tasks such as intrinsically disordered protein prediction, pK_a prediction, and metalloprotein prediction. It also highlights the role of machine learning in phylogenetics for understanding and engineering proteins. The findings underscore the potential of these approaches to advance protein science and inspire further research in related fields.