StructUnet

Deep Learning Based Approach for RNA Secondary Structure Prediction

Maria Eskerod Sørensen
Student ID: 201708588
Supervisor: Christian Storm Pedersen

Bioinformatic Research Center
Aarhus University
Aarhus, Spring 2024

Abstract

The molecular processes within cells are crucial for the life of all living organisms. Many of these processes are orchestrated by RNA molecules, which come in various types and fulfill a wide range of functions in cellular processes. Understanding the structure of RNA molecules is essential for comprehending their functions. Due to the challenges in experimentally determining RNA structure, developing accurate RNA secondary structure prediction methods has been an ongoing research focus since the 1970s.

This thesis explores the use of neural networks for RNA secondary structure prediction by developing a novel deep neural network, called StructUnet, and comparing it to existing models. The study includes a thorough examination of both algorithmic approaches and other neural network-based models, emphasizing their mechanisms and results. The fundamentals of RNA secondary structures and neural networks are discussed in depth, providing a foundation for understanding the context of this research. Traditional methods, including classical algorithmic approaches, Stochastic Context-Free Grammars (SCFG), and various machine learning and deep learning techniques, are reviewed. The development process of StructUnet involved numerous experiments, iteratively refining the architecture to optimize performance. The final model was extensively tested and compared with existing models. The proposed neural network demonstrates superior results, particularly in predicting pseudoknots, with significant improvements in F1 score metrics. The results show that StructUnet outperforms classical algorithmic approaches and other neural network-based models. Additionally, the model exhibits robust performance across various RNA types and lengths, showcasing its versatility and reliability. Despite these advancements, the model has certain limitations. These include stringent assumptions about RNA secondary structures and a limited dataset size, which may affect the generalizability of the findings. These constraints are critically examined, and potential avenues for future research are suggested. Future work should focus on relaxing the assumptions about RNA structures and expanding the dataset to include more diverse RNA sequences, which would likely enhance the model's accuracy and applicability.

Overall, this research contributes to the field of RNA secondary structure prediction by providing a powerful neural network-based approach that surpasses existing methods. It opens new avenues for leveraging deep learning in understanding and predicting RNA structures, which could have significant implications for biological research and applications.

Contents