INTISARI

PREDIKSI STRUKTUR SEKUNDER PROTEIN MENGGUNAKAN *EMBEDDING* DARI ANKH DAN BITCN-TRANSFORMER

Oleh

Venus Angela Kurniawan

21/473226/PA/20373

Prediksi Struktur Sekunder Protein (PSSP) merupakan salah satu permasalahan dalam bioinformatika yang dapat diselesaikan melalui pendekatan komputasi. Tugas PSSP termasuk dalam kategori *sequence labelling*, yaitu setiap residu pada sekuens protein diprediksi kelas struktur sekundernya. Pendekatan konvensional biasanya memanfaatkan fitur berbasis homolog, seperti *Position-Specific Scoring Matrix* (PSSM) dan *Hidden Markov Model* (HMM). Namun, proses untuk menghasilkan fitur tersebut membutuhkan waktu yang lama dan tidak dapat diterapkan pada protein tanpa homolog. Oleh karena itu, penelitian ini berfokus pada pengembangan model PSSP *single-sequence* yang hanya menggunakan sekuens protein sebagai *input*.

Penelitian ini menggunakan Ankh sebagai *Protein Language Model* (PLM) untuk memperoleh representasi kontekstual dari sekuens protein. Model yang dikembangkan penelitian ini menggabungkan *Bidirectional Temporal Convolutional Network* dan Transformer Encoder (BiTCN-Transformer). Arsitektur BiTCN digunakan untuk memproses informasi sekuensial dan menangkap pola lokal, sementara Transformer Encoder digunakan untuk mempelajari dependensi global.

Model pada penelitian ini dievaluasi menggunakan data uji CB513. Hasil pengujian menunjukkan bahwa model mencapai akurasi sebesar 78,767%. Akurasi tersebut mengungguli model acuan Ankh ConvBERT dengan selisih 0,317%.

Kata kunci: Prediksi Struktur Sekunder Protein, Ankh, *Bidirectional Temporal Convolutional Network*, Transformer *Encoder*

ABSTRACT

PROTEIN SECONDARY STRUCTURE PREDICTION USING EMBEDDING FROM ANKH AND BITCN-TRANSFORMER

By

Venus Angela Kurniawan

21/473226/PA/20373

Protein Secondary Structure Prediction (PSSP) is one of the problems in bioinformatics that can be approached through computational methods. PSSP belongs to the sequence labeling task, where each residue in a protein sequence is predicted according to its secondary structure class. Conventional approaches typically rely on homolog-based features such as the Position-Specific Scoring Matrix (PSSM) and Hidden Markov Model (HMM). However, generating these features takes a long time and cannot be applied to proteins without homologs. Therefore, this study focuses on developing a single-sequence PSSP model that uses only the protein sequence as input.

This study employs Ankh as a Protein Language Model (PLM) to obtain contextual representations of protein sequences. The proposed model integrates a Bidirectional Temporal Convolutional Network and a Transformer Encoder (BiTCN-Transformer). The BiTCN architecture is used to process sequential information and capture local patterns, while the Transformer Encoder is used to learn global dependencies.

The model in this study is evaluated using the CB513 test dataset. The results show that the model achieves a Q8 accuracy of 78,767%. This accuracy surpasses the Ankh ConvBERT baseline model by 0,317%.

Keywords: Protein Secondary Structure Prediction, Ankh, Bidirectional Temporal Convolutional Network, Transformer Encoder