ABSTRACT

PROTEIN SECONDARY STRUCTURE PREDICTION USING STATE SPACE MODEL-BASED DEEP LEARNING ARCHITECTURE

By
R Ferdian Dita Nugraha
21/476963/PA/20622

The use of Transformer in the case of protein secondary structure prediction is limited by its complexity, which grows quadratically with the length of amino acid sequences. This computational complexity is because the Attention mechanism works quadratically. In this study, a state space model, Mamba (S6), was used instead of Transformer with Attention mechanism since its complexity growth is linear according to the input size. By utilizing amino acid and protein secondary structure evolutionary information and ordinal coding method, a deep learning model using Mamba as the main component was built. Through this experiment, we found that the Mamba-based model is able to maintain high accuracy over a wide range of protein sequence lengths as evidenced by Q_3 and Q_8 accuracies of 92.13% and 79.74%, respectively. The model was also able to outperform several CNN, LSTM, and Attention-based models on the same dataset.

Keywords: Protein Secondary Structure, Amino Acids, PSSM, Sequence-to-sequence, State Space Models

INTISARI

PREDIKSI STRUKTUR SEKUNDER PROTEIN MENGGUNAKAN ARSITEKTUR *DEEP LEARNING* BERBASIS MODEL *STATE SPACE*

Oleh
R Ferdian Dita Nugraha
21/476963/PA/20622

Penggunaan Transformer pada kasus prediksi struktur sekunder protein memiliki keterbatasan pada kompleksitasnya yang bertumbuh secara kuadratik seiring panjang sekuens asam amino. Kompleksitas komputasi ini disebabkan karena mekanisme Attention bekerja secara kuadratik. Pada penelitian ini, model *state space*, yaitu Mamba (S6), digunakan sebagai pengganti Transformer dengan mekanisme Attention mengingat pertumbuhan kompleksitasnya yang linier sesuai dengan ukuran input. Dengan memanfaatkan asam amino dan informasi evolusioner struktur sekunder protein serta metode pengkodean ordinal, dibangun sebuah model *deep learning* menggunakan Mamba sebagai komponen utamanya. Melalui eksperimen ini, didapatkan model berbasis Mamba yang mampu mempertahankan akurasi tinggi pada berbagai panjang sekuens protein yang dibuktikan dengan akurasi Q_3 dan Q_8 sebesar 92.13% dan 79.74%. Model tersebut juga mampu mengungguli beberapa model lain berbasis CNN, LSTM, serta Attention pada dataset yang sama.

Kata Kunci: Struktur Sekunder Protein, Asam Amino, PSSM, *Sequence-to-Sequence*, State Space Model