Context Convolutional Neural Network dan Protein Language Model untuk Prediksi Struktur Sekunder
Protein
Ririth Coshy Yah, Afiahayati, S.Kom., M.Cs., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

CONTEXT CONVOLUTIONAL NEURAL NETWORK DAN *PROTEIN LANGUAGE MODEL* UNTUK PREDIKSI STRUKTUR SEKUNDER PROTEIN

Oleh

Ririth Coshy Yah
22/493350/PA/21181

Prediksi struktur sekunder protein merupakan salah satu permasalahan penting dalam bioinformatika untuk memahami fungsi biologis protein. Pendekatan berbasis *Multiple Sequence Alignment* (MSA) membutuhkan sumber daya komputasi yang besar. Sementara itu, mekanisme *Attention* dan model Transformer mampu menangkap dependensi jarak jauh pada sekuens, namun memiliki kompleksitas kuadratik yang menjadi tantangan dalam pemrosesan sekuens protein yang panjang. Penelitian ini mengusulkan pendekatan alternatif dengan memanfaatkan Protein *Language Model* (PLM) Ankh sebagai ekstraktor *embedding* sekuens protein tanpa menggunakan MSA, serta *Context Convolutional Neural Network* (Contextnet) sebagai *downstream* model. Pendekatan ini menawarkan alternatif dengan kompleksitas teoritis yang lebih rendah dibandingkan model berbasis *attention* penuh. Contextnet memanfaatkan *dilated convolution* dan mekanisme *concatenation* multiskala untuk memodelkan pola lokal serta konteks sekuens protein secara efektif.

Penelitian ini mengevaluasi beberapa konfigurasi pelatihan melalui penggunaan *optimizer* SGD dan AdamW, serta penerapan mekanisme regularisasi *white noise* dan *label masking*. Model dievaluasi menggunakan dataset *benchmark* CB513. Hasil penelitian menunjukkan bahwa konfigurasi berbasis *optimizer* AdamW memberikan performa terbaik dengan akurasi sebesar 79.34% pada Q8 dan 88.87% pada Q3. Hasil tersebut mengungguli beberapa penelitian sebelumnya. Sementara itu, konfigurasi berbasis *optimizer* SGD juga menghasilkan performa yang kompetitif dengan akurasi sebesar 78.39% pada Q8 dan 88.75% pada Q3. Selain itu, penerapan *white noise* dan *label masking* membantu menjaga stabilitas pelatihan dan mengurangi kecenderungan *overfitting*, meskipun tidak secara langsung meningkatkan akurasi akhir model.

Kata Kunci: Prediksi Struktur Sekunder Protein, *Protein Language Model*, Ankh, *Context Convolutional Neural Network, Dilated Convolution*.

Context Convolutional Neural Network dan Protein Language Model untuk Prediksi Struktur Sekunder
Protein
Ririth Coshy Yah, Afiahayati, S.Kom., M.Cs., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

CONTEXT CONVOLUTIONAL NEURAL NETWORK WITH PROTEIN LANGUAGE MODEL FOR PROTEIN SECONDARY STRUCTURE PREDICTION

By

Ririth Coshy Yah
22/493350/PA/21181

Protein secondary structure prediction is one of the important problems in bioinformatics for understanding the biological functions of proteins. Multiple Sequence Alignment (MSA)-based approaches require substantial computational resources. Meanwhile, Attention mechanisms and Transformer models can capture long-range sequence dependencies, but their quadratic complexity becomes a challenge when processing long protein sequences. This study proposes an alternative approach by utilizing the Ankh Protein Language Model (PLM) as a protein sequence embedding extractor without using MSA and employing a Context Convolutional Neural Network (Contextnet) as the downstream model. This approach offers a lower theoretical complexity alternative compared to full attention-based models. Contextnet utilizes dilated convolution and a multiscale concatenation mechanism to effectively model local patterns and protein sequence context.

This study evaluated several training configurations through the use of the SGD and AdamW optimizers, as well as the implementation of white noise and label masking regularization mechanisms. The model was evaluated using the CB513 benchmark dataset. The results show that the configuration based on the AdamW optimizer achieved the best performance, with accuracies of 79.34% on the Q8 classification task and 88.87% on the Q3 classification task. These results outperform those reported in several previous studies. Meanwhile, the configuration based on the SGD optimizer also achieved competitive performance, with accuracies of 78.39% on Q8 and 88.75% on Q3. In addition, the application of white noise and label masking helped maintain training stability and reduce the tendency toward overfitting, although they did not directly improve the model's final accuracy.

Keywords: Protein Secondary Structure Prediction, Protein Language Model, Ankh, Context Convolutional Neural Network, Dilated Convolution.