INTISARI

SELF-ATTENTION DAN BIDIRECTIONAL GATED RECURRENT UNIT UNTUK PREDIKSI STRUKTUR SEKUNDER PROTEIN

Oleh

DEWI PRAMUDI ISMI
20/468169/SPA/00735

Prediksi struktur sekunder protein (PSSP) adalah salah satu *task* utama dan dikerjakan secara luas oleh peneliti di bidang Bioinformatika. *Deep neural networks* telah menjadi metode utama untuk membangun model PSSP dalam dekade terakhir karena potensinya untuk meningkatkan kinerja PSSP. Penggunaan fitur berbasis *evolutionary information*, seperti *position specific scoring matrix* dan *Hidden Markov Model profiles*, sebagai fitur input PSSP juga meningkatkan kinerja PSSP. Terlepas dari kemajuan yang telah dicapai saat ini, studi PSSP masih menghadapi beberapa masalah, termasuk: (1) Keakuratan model-model PSSP dengan fitur berbasis *evolutionary information* masih di bawah batas teoritis yang seharusnya bisa dicapai, (2) Protein dengan homolog yang tidak diketahui/terbatas tidak dapat memanfaatkan fitur berbasis *evolutionary information*, sedangkan jika hanya menggunakan sekuens asam amino sebagai input untuk model PSSP (*single-sequence* PSSP) hasil akurasi yang diraih jauh di bawah batas teoritis.

Dalam penelitian ini, kami mengusulkan dua model PSSP yang disebut SADGRU-SS (*Self-attention, Dense, Gated Recurrent Units for Secondary Structure Prediction*) dan GAMAPSS-Single (*Gadjah Mada Protein Secondary Structure - Single Sequence Prediction*). SADGRU-SS merupakan model PSSP yang menggunakan fitur berbasis *evolutionary information* sebagai input. GAMAPSS-Single adalah model PSSP *single-sequence*, yaitu model PSSP yang hanya menggunakan sekuens asam amino sebagai input.SADGRU-SS dibangun dengan arsitektur *deep learning* yang baru dan unik yang memanfaatkan *self-attention*, *asymmetric multi-layer perceptron (MLP)-gated recurrent unit (GRU) blocks*, dan *dense block* untuk menyelesaikan masalah PSSP. GAMAPSS-Single memanfaatkan *self-attention* untuk menghasilkan *attention enriched amino acid encoding* dan menggunakannya untuk meningkatkan kinerja PSSP *single-sequence*. Arsitektur GAMAPSS-Single juga menggunakan *gated recurrent units* dan *fully connected layers*.

Hasil eksperimen menunjukkan bahwa penggunaan *self-attention* dalam arsitektur SADGRU-SS berhasil meningkatkan performa SADGRU-SS. Selain itu, mele-

takkan *self-attention* pada posisi paling depan pada SADGRU-SS *networks* menghasilkan performa yang lebih baik dibandingkan menempatkannya pada posisi lain. Penggunaan konfigurasi asimetris di blok MLP-GRU menghasilkan performa lebih baik daripada konfigurasi yang simetris. Model SADGRU-SS diuji menggunakan dataset standar CB513. Hasil eksperimen menunjukkan bahwa performa SADGRU-SS pada *8-state* PSSP lebih unggul dari model PSSP lainnya. Model ini mencapai akurasi prediksi 70,74% dan 82,78% dalam *8-state* PSSP dan *3-state* PSSP.

Hasil eksperimen juga menunjukkan bahwa memperkaya pengkodean standar asam amino dengan fitur yang berasal dari *self-attention* meningkatkan performa PSSP *single-sequence* di *3-state* PSSP dan *8-state* PSSP. Selain itu, model GAMAPSS-Single mengungguli SPIDER3-Single dan ProteinUnet dengan memiliki akurasi *8-state* PSSP sebesar 62,30% pada data ujii TS1197. Model GAMAPSS-Single juga mencapai akurasi *3-state* PSSP sebesar 71,30% pada data uji TS1197 dan akurasi *3-state* PSSP sebesar 71,23% pada data uji TS2018. Performa ini cukup bersaing dengan performa model PSSP *single-sequence* sebelumnya. Di luar akurasi, GAMAPSS-Single memerlukan waktu inferensi yang lebih cepat 22x lipat dibandingkan model ProteinUnet dan lebih cepat 10x lipat dibandingkan SPOT1D-Single saat memprediksi protein pada dataset TS2018 menggunakan perangkat keras yang sama, yaitu GPU Nvidia GTX1080 Ti.

Kata kunci: protein, prediksi struktur sekunder protein, *self-attention*, *gated recurrent unit*, *multi layer perceptron*, akurasi.

ABSTRACT

SELF-ATTENTION AND BIDIRECTIONAL GATED RECURRENT UNIT FOR PROTEIN SECONDARY STRUCTURE PREDICTION

By

DEWI PRAMUDI ISMI
20/468169/SPA/00735

Protein secondary structure prediction (PSSP) is one of the prominent and widely conducted tasks in Bioinformatics. Deep neural networks have become the primary methods for building PSSP models in the last decade due to their potential to enhance PSSP performance. Taking the evolutionary information-based features, such as position-specific score matrices and hidden Markov model profiles, as PSSP input features also enhance PSSP performance. Despite the current progress that it has made, PSSP studies still face several issues, including: (1) The state of the art accuracy of PSSP with evolutionary information based features is still below the theoretical limit, (2) Proteins with unknown/limited homologous sequences cannot take advantage of the evolutionary information based features, whereas using only amino acid sequence as input for PSSP (single-sequence PSSP) results in accuracy that is far below the theoretical limit.

In this study, we propose two PSSP models, called SADGRU-SS (*Self-attention, Dense, Gated Recurrent Units for Secondary Structure Prediction*) and GAMAPSS-Single (*Gadjah Mada Protein Secondary Structure - Single Sequence Prediction*). SADGRU-SS is a PSSP model using evolutionary-information based features as inputs. GAMAPSS-Single is a single-sequence PSSP model taking solely amino acid sequence as input. SADGRU-SS is built with a novel and unique deep learning architecture that utilizes self-attention, asymmetric multi-layer perceptron (MLP)-gated recurrent unit (GRU) blocks, and a dense block for solving the PSSP problem. GAMAPSS-Single utilizes self-attention to generate attention-enriched amino acid encoding and uses it to improve the performance of single-sequence PSSP. GAMAPSS-Single's architecture also utilizes gated recurrent units and fully connected layers.

Our experiment results show that using self-attention in the SADGRU-SS architecture has successfully enhanced SADGRU-SS performance. Moreover, installing self-attention in the frontmost position of the networks produces better performance than locating it in other positions. Using the asymmetric configuration in the MLP-GRU blocks results in more excellent performance than the symmetric

ones. We evaluate the performance of our SADGRU-SS model using the standard CB513 test dataset. Our experiment shows that the performance of our model on 8-state PSSP outstands other PSSP models. The model achieves 70.74% and 82.78% prediction accuracy in the 8-state and 3-state PSSP, respectively.

Our experiment results also show that enriching the standard amino acid encoding with the self-attention-derived feature enhances the performance of single-sequence PSSP in both the 3-state and the 8-state PSSP. Furthermore, GAMAPSS-Single model outperforms SPIDER3-Single and ProteinUnet by having 62.30% Q8 accuracy in TS1197 test dataset. GAMAPSS-Single model also achieves 71.30% Q3 accuracy in TS1197 and 71.23% Q3 accuracy in TS2018, a comparable performance to previous single-sequence PSSP models. Moreover, GAMAPSS-Single requires a 22 times faster inference time than ProteinUnet and a 10 times faster inference time than SPOT1D-Single when predicting proteins in the TS2018 dataset using GPU Nvidia GTX1080 Ti.

Keywords: protein, protein secondary structure prediction, *self-attention*, *gated recurrent unit*, multi layer perceptron, accuracy.