INTISARI

PREDIKSI STRUKTUR SEKUNDER PROTEIN MENGGUNAKAN N-GRAMS DAN CONVOLUTIONAL NEURAL NETWORK

Oleh:

Annisa Rizqiana

21/475974/PPA/06145

Struktur sekunder protein memiliki peran penting dalam memahami fungsi dan interaksi protein, namun prediksinya tetap menjadi tantangan yang kompleks dalam bioinformatika. Prediksi struktur sekunder protein dilakukan dengan mengklasifikasi setiap urutan struktur primer ke dalam bentuk sekundernya. Fitur-fitur protein juga menjadi hal yang penting untuk merepresentasikan bagaimana pola setiap asam amino dalam suatu urutan.

Dalam penelitian ini, pemodelan *n*-grams akan digunakan untuk mengkodekan urutan asam amino protein menjadi *n*-gram, yaitu *n* urutan asam amino yang berdekatan. Penggunaan *n-grams* diharapkan dapat menangkap pola dan hubungan yang tersembunyi dalam urutan asam amino protein. Nilai *n* pada *n*-grams yang dianalisis adalah 2 (atau bisa disebut dengan bigrams) dan 3 (atau biasa disebut dengan trigrams). Kemudian, data urutan protein yang telah dimodelkan dengan *n*-grams akan dimasukkan ke dalam Convolutional Neural Network 1 Dimensi dengan mengoptimalkan parameternya untuk tugas prediksi struktur sekunder protein.

Pertama-tama, *n*-grams dibandingkan dengan set data one-hot encoding. Hasilnya, *n*-grams lebih baik daripada one-hot encoding, yakni bigrams 55.23%, lebih baik 0.22% daripada one-hot encoding, dan trigrams 55.07% lebih baik 0.06% daripada one hot encoding. Sementara one-hot encoding menghasilkan akurasi terbaik sebesar 55.01%. Bigrams menunjukkan performa terbaik, sehingga set data ini digabungkan dengan fitur profil PSSM untuk meningkatkan akurasi, dan menghasilkan akurasi Q3 sebesar 82.75% dan Q8 sebesar 68.1%. Secara keseluruhan, performa *bigrams* lebih baik daripada *trigrams*, serta penambahan fitur lain berupa PSSM juga dapat mendukung performa model untuk belajar lebih baik.

Kata Kunci: *N-Grams*, Prediksi Struktur Sekunder Protein, *Convolutional Neural Network*, *Sequence Labeling*, Bioinformatika

ABSTRACT

PROTEIN SECONDARY STRUCTURE PREDICTION USING N-GRAMS AND CONVOLUTIONAL NEURAL NETWORK

Annisa Rizqiana

21/475974/PPA/06145

Protein secondary structures have an important role in understanding protein function and its interaction, but their prediction remains a complex challenge in bioinformatics. Protein secondary structure prediction done by classifying each sequence of primary structures into their secondary forms. Protein features are also important to represent the pattern of each amino acid in a sequence.

In this study, *n*-grams modeling will be used to encode the amino acid sequence of a protein into n-grams, that is *n* contiguous amino acid sequences. The use of *n*-grams is supposed to capture hidden patterns and relationships in the amino acid sequences of proteins. The value of *n* in the *n*-grams being analyzed is 2 (or know as bigrams) and 3 (or commonly called trigrams). Then, the protein sequence data that has been modeled with *n*-grams will be use as input for a 1-dimensional Convolutional Neural Network by optimizing its parameters for the protein secondary structure prediction task.

First, the n-grams are compared with the one-hot coding dataset. As a result, n-grams are better than one-hot coding, namely bigrams 55.23%, 0.22% better than one-hot coding, and trigrams 55.07%, 0.06% better than one hot coding. Meanwhile one-hot coding produces the best accuracy of 55.01%. Bigrams showed the best performance, so this dataset was combined with PSSM profile features to improve accuracy, and resulted in Q3 accuracy of 82.75% and Q8 of 68.1%. Overall, the performance of bigrams is better than trigrams, and the addition of other features in the form of PSSM can also support model performance for better learning.

Keywords: *N*-Grams, Protein Secondary Structure Prediction, Convolutional Neural Network, Sequence Labeling, Bioinformatics