

BAB I

PENDAHULUAN

1.1 Latar Belakang dan Permasalahan

Protein adalah unit fungsional utama dalam organisme hidup dan terlibat dalam banyak proses biologis dalam sel. Karena struktur tiga dimensi protein menentukan fungsi dari sebuah protein, maka penting untuk memprediksi struktur protein.

Struktur protein adalah susunan tiga dimensi dari atom-atom pada molekul protein yang diakibatkan oleh pelipatan protein. Pelipatan protein menjadi struktur tertentu ditentukan oleh sejumlah interaksi nonkovalen seperti ikatan hidrogen, interaksi ionik, dan gaya Van der Waals yang bekerja. Struktur tiga dimensi suatu protein perlu dipahami karena fungsi sebuah protein ditentukan oleh struktur protein tersebut.

Struktur sekunder (SS) mengacu pada bentuk “tulang punggung” polipeptida dari protein. Secara garis besar, SS terbagi dalam tiga kelompok, yaitu *helix* (H), *strand* (E), dan satu jenis SS tidak teratur, yaitu *coil* (C).

Prediksi struktur sekunder protein dapat membantu penemuan dalam bidang farmasi dan bidang bioteknologi, yaitu desain obat dan penemuan enzim baru.

Li & Yu (2016) dan Johansen dkk. (2016) melakukan penelitian prediksi struktur sekunder protein menggunakan Convolutional Neural Network (CNN) dan Recurrent Neural Network (RNN). CNN dan RNN digabungkan untuk memprediksi struktur sekunder protein dan memperoleh rekor akurasi tertinggi saat ini yaitu 68,9% dan 69,7%. CNN dan RNN digunakan secara bersamaan karena CNN hanya dapat menangkap informasi antar asam amino yang

dipisahkan oleh jarak tidak lebih besar dari panjang filter terpanjang. Sehingga, untuk menangkap informasi pada jarak yang lebih besar, digunakan RNN.

Penggunaan Recurrent Neural Network (RNN) juga memungkinkan penggunaan pendekatan berbeda yaitu *sequence labelling*, seperti pada kasus POS-tagging. Menurut Graves (2008), *sequence labelling* adalah semua tugas di mana data *sequence* ditranskripsi dengan *sequence* label diskrit. Lebih jauh lagi, prediksi struktur sekunder protein dapat dipandang sebagai masalah *sequence labelling*, karena kita melabeli *sequence* asam amino dengan label diskrit (kelas), yaitu struktur sekundernya sendiri.

Namun, model dengan lapisan *recurrent* sulit untuk dilatih, karena butuh strategi untuk mencegah *exploding gradient* atau *vanishing gradient*, yaitu kondisi di mana *norm* dari gradien menjadi sangat besar atau sangat kecil (Pascanu, 2013). *Exploding gradient* dan *vanishing gradient* ditandai saat *loss* dari gradien menjadi tak terhingga, sehingga model memiliki akurasi buruk dan tak bisa dilatih lagi.

1.2 Rumusan Masalah

Rumusan masalah yang terdapat pada penelitian ini adalah:

1. Bagaimana menerapkan model Deep Learning dengan lapisan konvolusi dan Bidirectional Gated Recurrent Unit untuk memprediksi struktur sekunder protein dari struktur primernya?
2. Bagaimana akurasi Deep Learning model Deep Learning dengan lapisan lapisan konvolusi dan Bidirectional Gated Recurrent Unit dalam prediksi struktur sekunder protein?

1.3 Batasan Masalah

Batasan-batasan masalah di dalam penelitian ini antara lain adalah:

1. Dataset yang digunakan adalah CullPDB, CullPDB-Filtered, dan CB513
2. Data diklasifikasikan ke dalam 8 label.
3. Arsitektur Deep Learning yang akan diteliti memiliki lapisan konvolusi dan Bidirectional Gated Recurrent Unit.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk menggunakan Deep Learning untuk melakukan prediksi struktur sekunder protein ke dalam delapan kelas dan lebih akurat dari metode-metode pada penelitian sebelumnya.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat membantu peneliti biomedis dalam menentukan struktur sekunder protein secara otomatis dengan akurasi yang tinggi dan waktu yang cepat.