PROTEIN REPRESENTATION

SEQUENCE EMBEDDINGS

MEREDITA SUSANTY

Penerbit Universitas Pertamina

Protein Representation Sequence Embeddings

Penulis: Meredita Susanty

ISBN: 978-623-88235-3-6

Editor: Eka Budhi Septiawan

Penerbit: Universitas Pertamina

Cetakan Pertama, 2022
Edisi Pertama, 2022

Jl Teuku Nyak Arief Simprug
Kebayoran Lama, Jakarta Selatan 12120
Telepon : 021-29044308
Website : https://universitaspertamina.ac.id/
Email : info@universitaspertamina.ac.id

Cetakan Pertama, 2022
Edisi Pertama, 2022

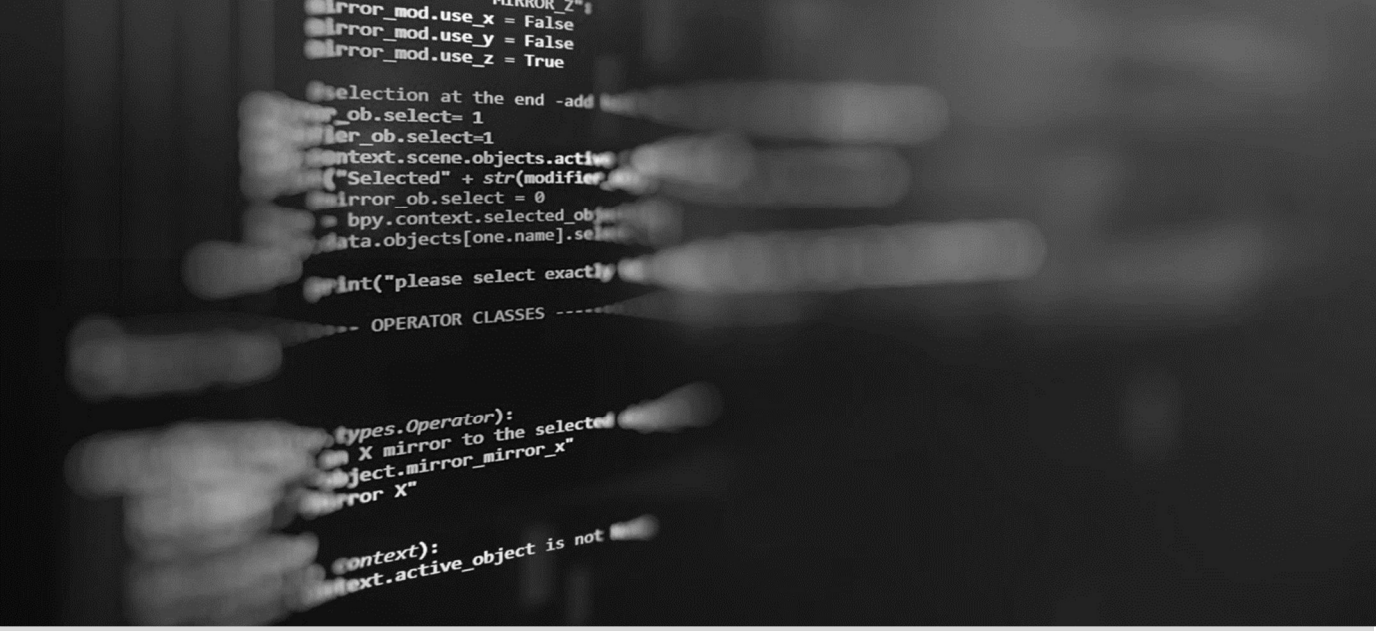
Susanty, Meredita

Proten Representation Sequence Embeddings

—Jakarta: Penerbit Universitas Pertamina,2022
1 jil., 155 hlm., 17,6 x 25 cm

ISBN. 978-623-88235-3-6

1. Ilmu Komputer
2. Kecerdasan Buatan
I. Judul
II. Susanty, Meredita



Tentang Penulis

Meredita Susanty adalah dosen tetap program studi Ilmu Komputer Universitas Pertamina sejak tahun 2016. Meredita meraih gelar sarjana komputer dari Universitas Gadjah Mada dan gelar master dari University of Nottingham. Di program studi Ilmu Komputer Meredita mengajar Pengantar Teknologi Informasi dan Algoritma, Dasar Pemrograman dan Rekayasa Perangkat Lunak.

Prakata

Pemrosesan bahasa alami (*natural language processing*, NLP) merupakan bagian dari bidang ilmu linguistik, ilmu komputer dan kecerdasan buatan. Beberapa tahun belakangan, seiring dengan berbagai terobosan dalam pembelajaran mesin, metode-metode dalam bidang NLP juga menunjukkan kemajuan yang luar biasa. Karena kemiripan struktur antara protein dan bahasa, metode-metode dalam bidang NLP banyak diadopsi dan diimplementasikan untuk mempelajari protein. Buku ini membahas konsep dasar berbagai metode NLP yang diadopsi dalam protein analisis dan kemudian penerapannya dalam mempelajari protein. Buku ini juga membahas keberhasilan, potensi, serta keterbatasan algoritma NLP dalam mempelajari protein.

Bagian awal dari buku ini membahas persamaan dan perbedaan antara protein dan bahasa serta beberapa *task* dalam mempelajari protein yang serupa dengan *task* dalam NLP. Pada bagian selanjutnya dibahas beragam cara merepresentasikan protein dengan teknik-teknik yang diadopsi dari NLP, dimulai dari konsep klasik seperti *bag-of-words*, *n-grams* hingga teknik yang terkini seperti *word embeddings*.

Buku ini diharapkan dapat membantu pembaca yang ingin mendalami bidang bioinformatika namun tidak memiliki pengetahuan mendalam di bidang pemrosesan bahasa alami. Karena urutan kemunculan kata dalam bahasa sangat penting, pemrosesan bahasa alami banyak menggunakan model sekuensial. Bagi pembaca yang ingin memiliki pemahaman yang lebih baik atau ingin mendalami konsep dasar model sekuensial yang mendasari beberapa teknik dan metode yang dibahas dalam buku ini, dapat membaca buku *deep learning - sequential model* dan referensi yang diacu di akhir setiap bab.

Daftar Isi

BAB 1

Protein dan Bahasa

Protein adalah komponen dasar pembangun kehidupan karena menggerakkan seluruh jaringan mekanisme fungsional yang saling terkoneksi dan terkorelasi pada suatu organisme. Protein terlibat dalam hampir setiap fungsi seluler, termasuk jalur pensinyalan, melakukan perbaikan DNA, mentransportasikan glukosa transmembran, aktivitas katalitik, serta aktivitas pengangkut. Struktur tersier protein dan mungkin juga interaksi protein ditentukan oleh urutan asam amino yang menyusun protein. Karena alas an inilah, sekuens protein sering disebut sebagai bahasa kehidupan.

Dibidang bioinformatika, analisis sekuens protein dan menggali berbagai informasi terkait struktur dan fungsi protein menjadi salah tujuan utama dan tema jangka panjang. Teknologi *sequencing* protein yang ada saat ini telah menyebabkan peningkatan ukuran basis data protein secara eksponensial. Rata-rata hampir dua kali lipat peningkatan ukuran basis data protein setiap dua tahun. Disisi lain, proses pemberian label yang valid dan bermakna untuk protein yang sudah di-*sequencing* membutuhkan banyak upaya, keahlian, eksperimen, dan juga biaya. Akibatnya, ada banyak data protein yang tidak memiliki anotasi atau label dibandingkan yang sudah dianotasi secara manual. Perbedaan ini dapat dilihat dari jumlah protein yang ada di pada basis data TrEMBL yang berisi 219 juta sekuens protein dengan basis data SwissProt yang dianotasi secara manual yang hanya berisi 565 ribu protein. Melihat angka ini, kesenjangan antara sekuens dan anotasi akan semakin meningkat jika proses anotasi terus dilakukan secara manual.

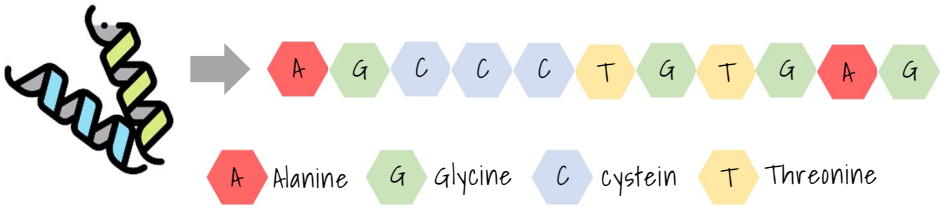
Selama beberapa decade, para peneliti di bidang bioinformatika berupaya mengembangkan berbagai metode komputasional untuk memprediksi struktur, fungsi, dan anotasi berdasarkan informasi urutan asam amino. Pada metode prediksi konvensional, informasi ukuran protein yang menjadi target disajikan dalam bentuk urutan protein tunggal, *position-specific scoring matrix* (PSSM), *Hidden-Markov Model*, dan k-grams. Terkadang, untuk melakukan prediksi dibutuhkan juga properti-properti selain sekuens asam amnio, misalnya sifat psikokimia seperti hidrofobisitas, muatan, dan ukuran molekul , baik untuk melengkapi informasi urutan asam amino pada protein atau justru menggantikan informasi ini.

Beberapa tahun belakangan, bidang pemrosesan bahasa alami atau natural language processing (NLP) mengalami perubahan paradigma dengan munculnya *pre-trained language model*. Tren di bidang ini saat ini adalah melatih sebuah *language model* terhadap sebuah *corpus* yang berukuran besar berisi data teks tanpa label menggunakan pendekatan *unsupervised* atau *semi-supervised*. Dengan pendekatan ini, model mampu mempelajari pola dan struktur dari suatu bahasa. Model *pre-training* ini memberikan pengetahuan umum dari suatu bahasa dalam bentuk *embeddings*. Biasanya untuk digunakan pada *downstream task*, perlu dilakukan *fine tuning* terlebih dahulu terhadap *embeddings* hasil *pre-trained* ini. Pendekatan pre-trained ini secara signifikan meningkatkan performa model metode supervised yang ada sebelumnya yang dilatih tehadap dataset untuk *task* tertentu yang datasetnya berukuran kecil.

Melihat keberhasilan *word embeddings* dan *pre-trained language model* di bidang NLP, pendekatan ini perlahan-lahan mulai populer digunakan dalam analisis sekuens protein. Pada bab ini kita akan membahas terlebih dahulu kesamaan dan perbedaan antara protein dan bahasa untuk mendapatkan gambaran mengapa pendekatan-pendekatan di bidang NLP dianggap sesuai untuk digunakan dalam mempelajari protein.

1.1 Struktur dan Substansi Protein dan Bahasa

Dalam bahasa, kalimat tersusun dari kata-kata dan kata tersusun dari huruf-huruf dengan urutan tertentu. Protein yang tersusun dari untaian asam amino dengan urutan tertentu secara alami dapat direpresentasikan sebagai untaian huruf atau kata. Ada 20 asam amino yang menyusun protein. Asam amino ini dinotasikan menggunakan satu huruf alfabet seperti yang ditunjukkan pada Gambar 1.1.



Gambar 1.1 Ilustrasi untaian protein dalam notasi alfabet

Bahasa disusun dari elemen-elemen modular yang dapat digunakan kembali. Elemen-elemen modular ini dapat diatur ulang dan disusun secara hierarkis sehingga membuat sedikit variasi pada bahasa. Jika pada bahasa elemen-elemen ini adalah kata, frasa, dan kalimat, pada protein hal ini dapat dianalogikan dengan motif dan domain yang merupakan *functional building block* dari protein.

Fitur utama lainnya yang dimiliki oleh protein dan bahasa adalah kelengkapan informasi. Hal ini bisa analogikan dengan kalimat majemuk dalam bahasa dimana sebuah kalimat terdiri dari beberapa klausa. Klausa ini bisa saja berdiri sebagai kalimat sendiri seperti dalam kalimat majemuk setara atau bergantung pada klausa lain seperti dalam kalimat majemuk bertingkat. Pada kalimat majemuk bertingkat, sebuah kata dalam suatu klausa memiliki keterkaitan dengan klausa lainnya. Jadi dalam bahasa sebuah kata tidak hanya memiliki keterkaitan dengan kata yang persis berada disampingnya tapi juga dengan kata atau klausa lain yang berdekatan. Untuk protein, selain sekuens asam amino yang menyusun protein yang dikenal dengan bentuk satu dimensi atau struktur primer, protein juga memiliki bentuk dua dan tiga dimensi yang disebut juga struktur sekunder dan tersier. Bentuk satu dimensi protein

berbentuk linear terdiri dari ikatan peptida antara asam amino yang berdampingan. Bentuk dua dimensi protein bisa berupa alpha-helix atau beta-sheet terbentuk karena ikatan hidrogen antara asam amino yang berdekatan (bukan berdampingan). Bentuk tiga dimensi protein adalah globular terdiri dari ikatan hidrogen, jembatan disulfida, dan jembatan garam. Meskipun protein lebih dari sekadar urutan asam amino – protein juga memiliki bentuk tiga dimensi dengan struktur dan fungsi tertentu– aspek-aspek ini semuanya ditentukan oleh urutan asam aminonya. Meskipun struktur dan fungsi protein bersifat dinamis dan bergantung pada konteks (misalnya pada keadaan seluler, molekul lain, atau *post-translation modification*), struktur dan fungsi ini masih ditentukan oleh urutan asam amino yang menyusunnya. Hal ini menunjukkan bahwa dari perspektif teori informasi, informasi protein (misalnya strukturnya) terkandung dalam urutannya.

Karena berbagai kesamaan bentuk dan substansi antara protein dan bahasa ini, wajar saja untuk menerapkan metode NLP terhadap untaian protein. Pada beberapa dekade terakhir pembelajaran mesin dari bidang NLP secara berkesinambungan digunakan di dalam bioinformatika.

Kesamaan antara bahasa dan protein hanya sebatas hal-hal yang sudah dibahas sebelumnya. Selanjutnya kita akan membahas perbedaan antara protein dan bahasa yang harus diperhatikan jika ingin menggunakan metode-metode NLP dalam mempelajari protein. Pertama, kita bisa membaca dan memahami bahasa namun tidak dengan protein. Tidak seperti protein, dalam bahasa kita mengenal tanda baca yang dengan jelas memisahkan struktur seperti kata, kalimat, dan paragraf. Pada protein kita tidak dapat mengetahui dengan jelas apakah serangkaian sekuens asam amino merupakan bagian dari sebuah unit fungsional atau bukan. Selain itu, tidak ada analogi yang jelas antara komponen yang membangun protein dan bahasa. Misalnya, menganggap bahwa protein domain setara dengan kata-kata dalam bahasa sering kali menyesatkan. Unit fungsional protein mungkin saja tumpang tindih (*overlap*). Akibatnya, protein tidak memiliki kosakata yang jelas. Tidak seperti bahasa yang memiliki kosakata yang terdefinisi dengan baik.

Dari sudut pandang teori informasi, entropi dari sekuens pada protein domain lebih rendah dibandingkan entropi dari sekuens pada bahasa Inggris. Namun demikian, nilai entropi sekuens protein berbeda dari distribusi acak. Protein juga menunjukkan variabilitas panjang urutan yang berbeda dengan bahasa. Panjang asam amino pada protein manusia berukuran kurang dari 20 asam amino untuk peptida hormon hingga 10.000 asam amino pada beberapa protein struktural. Angka ini bisa lebih dari tiga kali lipat dibandingkan panjang kalimat dalam bahasa. Rentang panjang untaian protein yang begitu luas merupakan hal lazim di semua domain kehidupan, dari virus hingga manusia. Karena perbedaan panjang yang cukup jauh, bahasa umumnya memiliki jumlah interaksi jauh (interaksi dengan kata yang jaraknya lebih dari satu kata) yang lebih sedikit dibandingkan protein. Pada protein, karena pada struktur tiga dimensi protein residu bisa membentuk interaksi antara residu dengan residu lain yang posisinya jauh pada urutan linier jumlah interaksi jauhnya akan lebih banyak.

Pada NLP, kata-kata tertentu mungkin memiliki pengaruh penting. Contohnya dalam bahasa Inggris ”I love you” dan ”I loved you” memiliki makna yang berbeda. Pada protein, yang mungkin memiliki pengaruh penting adalah sesuatu yang lebih agregat, misalnya rantai hidrofilik dalam urutan transmembran.

1.2 Prediksi pada Protein dan Bahasa

Metode NLP digunakan untuk berbagai *sequence-based prediction* tasks baik terhadap teks maupun protein. Pada level paling mendasar, *sequence-based prediction* terbagi menjadi *global* dan *local*. Tasks yang umum dilakukan pada kedua bidang ini adalah prediksi berdasarkan urutan atau sekues (*sequence based prediction*). Pada *global task*, *output* yang dihasilkan adalah prediksi terhadap keseluruhan urutan. Misalnya, dalam bahasa memprediksi sentimen terhadap ulasan sebuah film merupakan sebuah properti global dari teks.

Disisi lain, *local tasks* mencoba memprediksi setiap elemen dari sekuens yang diberikan. *Part of speech tagging*, dimana model mencoba mengkategorikan peran gramatikal setiap kata dalam teks (misalnya

apakah sebuah kata merupakan kata benda, kata kerja, atau kata sifat) merupakan contoh *task local* pada bidang NLP.

Gambar 1.2 Prediksi local dan global pada protein dan bahasa.

Dalam mempelajari protein, global tasks mencoba menarik kesimpulan atau prediksi terhadap sekuens protein secara keseluruhan. Misalnya, kita ingin mengetahui tipe protein, apakah protein merupakan enzim, reseptor, atau protein struktural. Contoh lainnya adalah jika kita ingin mengetahui dimana protein diekspresikan dalam sel, apakah di nukleus, cytoplasma, atau *extracellular space*. Properti global lainnya pada protein adalah stabilitas termal, asal organisme, dan anotasi protein fungsional (misalnya *gene ontologi* atau GO) seperti aktivitas antiviral.

Local task pada protein bertujuan melakukan prediksi tentang residue tertentu pada protein sekuens atau sering kali melakukan prediksi untuk setiap residue pada protein sekuens. Contoh yang umum adalah prediksi struktur 2D dan 3D dari sebuah untaian asam amino. Output dari prediksi ini adalah struktur 2D dari setiap residue pada protein misalnya *helix*, *turn*, atau *beta strand*. Pada contoh struktur 3D, output bisa berupa koordinat dari setiap residue yang menunjukkan posisi suatu residue atau lokasi relatif suatu residue terhadap residue lain (*contact-map prediction*). Contoh lain untuk *locak task* adalah prediksi modifikasi pasca-translasi (*post-translation modification*, PTM) seperti situs fosforilasi atau pembelahan.

Dalam melakukan *sequence-based prediction task*, dapat juga digunakan input tambahan selain sekuens protein baik yang dikumpulkan secara