**PENERAPAN MODEL *SEQUENCE-TO-SEQUENCE* BERBASIS *TRANSFORMER* UNTUK *MULTIPLE SEQUENCE ALIGNMENT***

**SKRIPSI**

Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana Komputer



Disusun Oleh :

TANJUNG ARSWENDO YUDHA

11220910000043

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS SAINS DAN TEKNOLOGI**

**UNIVERSITAS ISLAM NEGERI SYARIF HIDAYATULLAH JAKARTA**

**2025 M/1446 H**

# **KATA PENGANTAR**

# **DAFTAR ISI**

[KATA PENGANTAR i](#_Toc205361931)

[DAFTAR ISI ii](#_Toc205361932)

[BAB I PENDAHULUAN 1](#_Toc205361933)

[1.1 Latar Belakang 1](#_Toc205361934)

[1.2 Rumusan Masalah 4](#_Toc205361935)

[1.3 Batasan Masalah 4](#_Toc205361936)

[1.3.1 Metode 5](#_Toc205361937)

[1.3.2 *Tools* 5](#_Toc205361938)

[1.3.3 *Dataset* 5](#_Toc205361939)

[1.3.4 Proses 5](#_Toc205361940)

[1.4 Tujuan Penelitian 6](#_Toc205361941)

[1.5 Manfaat Penelitian 6](#_Toc205361942)

[1.5.1 Bagi Penulis 6](#_Toc205361943)

[1.5.2 Bagi Universitas 6](#_Toc205361944)

[1.5.3 Bagi Pembaca 7](#_Toc205361945)

[1.6 Metodologi Penelitian 7](#_Toc205361946)

[1.6.1 Metode Pengumpulan Data 7](#_Toc205361947)

[1.6.2 Metode Implementasi 7](#_Toc205361948)

[1.7 Sistematika Penulisan 8](#_Toc205361949)

[BAB II LANDASAN TEORI 10](#_Toc205361950)

[2.1 Bioinformatika 10](#_Toc205361951)

[2.2 Analisis Sekens Biologi 10](#_Toc205361952)

[2.3 Multiple Sequence Alignment 10](#_Toc205361953)

[2.4 Tinjauan metode MSA Konvensional 10](#_Toc205361954)

[2.5 Artificial Intelligence (AI) 10](#_Toc205361955)

[2.6 Deep Learning 10](#_Toc205361956)

[2.7 Sequence-to-sequence 10](#_Toc205361957)

[2.8 Arsitektur Transformer 10](#_Toc205361958)

[2.9 Mekanisme self-attention 10](#_Toc205361959)

[2.10 Dataset Sintetis dan Ground Truth 10](#_Toc205361960)

[2.11 Studi Literatur terkait 10](#_Toc205361961)

[BAB III METODOLOGI PENELITIAN 11](#_Toc205361962)

[3.1 Objek Penelitian 11](#_Toc205361963)

[3.2 Metode Pengumpulan data 11](#_Toc205361964)

[3.2.1 Studi Pustaka 11](#_Toc205361965)

[3.3 Metode Implementasi 11](#_Toc205361966)

[3.3.1 *Research* 11](#_Toc205361967)

[3.3.4 *Modelling* 12](#_Toc205361968)

[3.3.5 *Requirement* 13](#_Toc205361969)

[DAFTAR PUSTAKA 14](#_Toc205361970)

# **BAB I**

# **PENDAHULUAN**

# **Latar Belakang**

Bioinformatika adalah bidang ilmu yang menggunakan perangkat komputasi untuk memahami data biologis. Salah satu metode yang menjadi dasar penting dalam bidang ini adalah Multiple Sequence Alignment (MSA), yaitu sebuah proses komputasi untuk melakukan alignment (penyelarasan) tiga atau lebih sekuens biologis, seperti DNA, RNA, atau protein. Proses ini secara teknis dilakukan dengan menyisipkan celah (gap) ke dalam sekuens agar panjangnya menjadi sama, yang memungkinkan perbandingan kolom per kolom. Tujuan biologis utamanya adalah untuk memastikan bahwa setiap kolom berisi residu-residusu yang homolog, atau berasal dari posisi dan leluhur yang sama. Proses alignment yang berhasil dilakukan dengan tepat akan menampakkan daerah-daerah lestari (conserved regions), yaitu area – area dalam sekuens yang cenderung tidak banyak berubah selama proses evolusi yang memberikan petunjuk penting mengenai hubungan dari segi struktur, fungsi, dan evolusi antar sekuens tersebut (Almanza-Ruiz et al., 2023).

Pentingnya MSA terlihat jelas dari cakupan penggunaannya yang sangat luas. Dalam studi evolusi, MSA merupakan langkah pertama yang sangat diperlukan dalam studi evolusi untuk membuat pohon filogenetik, sebuah diagram percabangan yang menggambarkan hubungan evolusioner dan melacak bagaimana spesies atau gen saling berkerabat (Victor Aprilyanto & Langkah Sembiring, 2016). Di bidang biologi struktural, MSA juga digunakan untuk membantu memprediksi bentuk tiga dimensi sebuah protein. Lebih dari itu, metode ini juga sangat penting untuk menemukan pola-pola fungsional (motif) dan merekonstruksi sekuens dari masa lalu. Karena perannya yang sangat sentral ini, kualitas hasil sebuah MSA akan sangat menentukan kebenaran dan keandalan dari berbagai analisis selanjutnya (Reddy & Fields, 2022).

Pada dasarnya, untuk menilai apakah sebuah MSA itu "baik" atau tidak, digunakan sebuah sistem skor. Salah satu skema yang paling umum adalah Sum-of-Pairs (SP), sebuah metrik yang menilai kualitas alignment dengan menjumlahkan skor dari setiap kemungkinan pasangan sekuens di semua kolom, biasanya dengan bantuan matriks substitusi seperti BLOSUM62 dan aturan penalti jika ada celah (gap) (Almanza-Ruiz et al., 2023). Namun, menemukan MSA dengan skor terbaik merupakan tantangan komputasi yang sangat besar. Masalah ini secara teoretis tergolong NP-complete, yaitu kelas masalah komputasi yang solusinya sangat sulit ditemukan dalam waktu yang wajar saat ukuran data bertambah. Karena sifatnya yang eksponensial, akibatnya hampir semua perangkat lunak MSA yang ada saat ini mengandalkan pendekatan heuristik, yaitu sebuah cara untuk mendapatkan solusi yang "cukup baik" dalam waktu yang wajar, meskipun tidak dijamin sebagai yang terbaik. Pendekatan yang paling populer adalah progressive alignment, sebuah teknik yang membangun MSA secara bertahap dimulai dari sekuens yang paling mirip, yang mana digunakan oleh program-program terkenal seperti ClustalW, T-Coffee, dan MAFFT (Edgar & Batzoglou, 2006).

Meskipun sangat berguna, pendekatan heuristik ini memiliki kelemahan mendasar. Hasil akhirnya sangat bergantung pada alignment di tahap-tahap awal; jika terjadi kesalahan di sana, maka kesalahan tersebut akan terus terbawa sampai akhir. Selain itu, metode konvensional seringkali memakai aturan atau matriks skor yang sama untuk semua jenis data. Padahal, cara sekuens berevolusi, seperti laju penambahan (insersi) dan penghapusan (delesi) sekuens yang biasa dikenal dengan sebutan Indel, yang mana bisa sangat berbeda-beda di setiap organisme atau gen (Dotan et al., 2024). Keterbatasan inilah yang mendorong para peneliti untuk mencari pendekatan alternatif yang lebih fleksibel dan akurat.

Keterbatasan utama metode konvensional terletak pada ketergantungannya pada aturan-aturan statis, seperti matriks substitusi dan penalti celah (gap) yang nilainya sudah ditentukan sebelumnya. Sistem ini kesulitan untuk beradaptasi dengan pola-pola evolusi yang kompleks dan beragam. Di sisi lain, model sequence-to-sequence (S2S) yang awalnya dikembangkan untuk tugas penerjemahan bahasa (sutskever, 2014), unggul karena kemampuannya untuk mempelajari pola dan tata bahasa secara langsung dari data. Sama seperti model penerjemah yang belajar di mana harus menempatkan sebuah kata berdasarkan konteks kalimat, sebuah model S2S untuk MSA berpotensi untuk mempelajari pola-pola biologis yang menentukan di mana sebuah gap (indel) seharusnya ditempatkan secara efisien dan optimal, bukan hanya berdasarkan penalti yang kaku.

Seiring kemajuan pesat di bidang Artificial Intelligence (AI), kini muncul sebuah pendekatan baru yang mengubah cara pandang dalam menyelesaikan masalah MSA. Pendekatan ini datang dari bidang Deep Learning dan secara mendasar mengubah MSA dari masalah optimasi matematis menjadi masalah sequence-to-sequence (S2S), sebuah paradigma di mana model dilatih untuk mengubah satu urutan data (input) menjadi urutan data lainnya (output), yang idenya diadaptasi dari Natural Language Processing (NLP) (Dotan et al., 2023). Dalam cara pandang baru ini, sebuah model AI dilatih untuk "menerjemahkan" sekumpulan sekuens yang belum selaras menjadi sebuah MSA yang sudah selaras dan rapi.

Teknologi yang mendasari pendekatan S2S ini adalah arsitektur Transformer (Madan et al., 2024), sebuah arsitektur deep learning yang dirancang khusus untuk memproses data sekuensial. Model Transformer memiliki kemampuan luar biasa untuk menangkap hubungan antar bagian dalam data sekuens, bahkan yang letaknya berjauhan, melalui mekanisme yang disebut self-attention, yaitu kemampuan model untuk menimbang pentingnya setiap elemen dalam sekuens input saat memproses informasi. Kemampuan ini memungkinkan model untuk mempelajari pola-pola evolusi yang rumit, sesuatu yang seringkali luput dari metode konvensional yang berbasis skor.

Tentu saja, keberhasilan model deep learning sangat bergantung pada data yang digunakan untuk melatihnya. Model-model ini membutuhkan data pelatihan dalam jumlah yang sangat besar dan berkualitas tinggi untuk bisa belajar dengan baik. Untuk tugas supervised learning seperti ini, model memerlukan contoh-contoh MSA yang sudah "benar" atau memiliki ground truth. Di sinilah letak sebuah celah (gap) yang signifikan dalam penelitian MSA saat ini, yaitu masih langkanya dataset pelatihan yang besar dan dapat diandalkan. Dataset benchmark yang ada saat ini seringkali dibuat secara manual, sehingga jumlahnya tidak banyak, cakupannya terbatas, dan kebenarannya sebagai "standar emas" pun terkadang masih bisa diperdebatkan (Gotoh et al., 2014).

Salah satu solusi paling menjanjikan untuk mengatasi kelangkaan data ini adalah dengan membuat dataset sintetis berskala besar melalui proses simulasi evolusi. Dengan menggunakan program simulator, penulis dapat menciptakan jutaan contoh MSA di mana keseluruhan proses evolusinya diketahui secara pasti, termasuk setiap peristiwa insersi dan delesi (Loewenthal et al., 2021). Hal ini menghasilkan pasangan data input (unaligned) dan output (ground truth) yang sempurna untuk melatih model S2S. Meskipun mungkin ada sedikit perbedaan antara data hasil simulasi dan data biologis di dunia nyata (Trost et al., 2023), pendekatan ini merupakan cara yang paling mungkin dilakukan untuk menghasilkan data dalam skala masif yang dibutuhkan oleh model AI modern.

Penelitian yang dilakukan sebelumnya telah berhasil membangun sebuah pipeline otomatis untuk menghasilkan dataset sintetis MSA dalam skala besar (Korosteleva & Lee, 2021). Dengan landasan data tersebut, maka penelitian ini bertujuan untuk melanjutkan ke tahap berikutnya. Fokus dari skripsi ini adalah untuk menerapkan, melatih, dan menguji sebuah prototipe model berbasis Transformer, dengan tujuan utama untuk membuktikan bahwa paradigma sequence-to-sequence merupakan solusi modern yang layak untuk masalah Multiple Sequence Alignment (Dotan et al., 2024).

# **Rumusan Masalah**

Berdasarkan latar belakang tersebut, dapat diidentifikasi masalah utama yang melandasi penelitian ini. Terdapat keterbatasan komputasi dan skalabilitas pada metode Multiple Sequence Alignment (MSA) konvensional, yang mendorong perlunya solusi alternatif. Meskipun pendekatan modern berbasis Artificial Intelligence (AI) sangat menjanjikan, pengembangannya secara fundamental terhambat oleh kelangkaan dataset pelatihan yang besar dan berkualitas tinggi. Oleh karena itu, muncul kebutuhan untuk mengambil langkah selanjutnya, yaitu mengimplementasikan dan menguji sebuah model AI yang dilatih pada dataset yang andal untuk membuktikan kelayakan pendekatan baru ini.

Berdasarkan permasalahan tersebut, maka dapat dirumuskan pertanyaan penelitian sebagai berikut:

1. Bagaimana menerapkan arsitektur model Transformer untuk menyelesaikan masalah *Multiple Sequence Alignment* (MSA) dengan memformulasikannya sebagai tugas *sequence-to-sequence*?
2. Bagaimana kinerja akurasi dari prototipe model yang diusulkan setelah dilatih menggunakan dataset sintetis, jika dibandingkan dengan metode MSA konvensional yang berbasis heuristik?

# **Batasan Masalah**

Agar penelitian ini tetap terfokus dan mendalam, maka ditetapkan beberapa batasan masalah sebagai berikut:

### **1.3.1 Metode**

1. Penelitian ini menggunakan metode deep learning dengan arsitektur Transformer sebagai teknik utama.
2. Metode pembuatan dataset yang digunakan adalah teknik simulasi evolusi sekuens melalui pipeline SpartaABC.
3. Metodologi penelitian yang digunakan adalah pendekatan penelitian eksploratif dan iteratif.
4. Metode evaluasi kinerja model akan menggunakan **(catatan : dikonsultasikan dulu) (samakan dengan referensi jurnal Dotan)**

### **1.3.2 *Tools***

1. Visual Studio Code (VS Code) digunakan untuk pengembangan dan modifikasi skrip pada pipeline SpartaABC agar sesuai dengan kebutuhan penelitian.
2. SpartaABC digunakan sebagai program inti untuk menjalankan simulasi dan proses pembuatan dataset.
3. Docker digunakan sebagai lingkungan tervirtualisasi untuk menjamin konsistensi dan portabilitas dalam menjalankan pipeline SpartaABC.
4. Google Colaboratory (Colab) digunakan sebagai platform untuk pengembangan dan pengujian awal (prototyping) model dalam skala kecil.
5. High-Performance Computing (HPC) BRIN akan digunakan untuk tahap pelatihan model skala besar.

### **1.3.3 *Dataset***

**(Catatan : dikonsultasikan dulu) (Samakan dengan laporan KKL).**

### **1.3.4 Proses**

1. Proses penelitian akan diawali dengan melakukan studi literatur untuk mendalami arsitektur Transformer dan MSA.
2. Peneliti akan mengadaptasi dan melakukan eksekusi pada pipeline SpartaABC untuk menghasilkan dataset pelatihan dan pengujian.
3. Proses inti adalah merancang, mengimplementasikan, serta melatih model aligner berbasis Transformer hingga menghasilkan output berupa model yang siap digunakan serta diuji kinerjanya.

# **Tujuan Penelitian**

Berdasarkan rumusan masalah yang telah ditetapkan, maka tujuan dari penelitian ini adalah sebagai berikut :

1. Menerapkan dan mengimplementasikan model sequence-to-sequence berbasis arsitektur Transformer untuk membangun sebuah model aligner Multiple Sequence Alignment (MSA).
2. Menganalisis kinerja akurasi dari model aligner yang telah dibangun, terutama jika dibandingkan dengan metode konvensional berbasis heuristik.

# **Manfaat Penelitian**

Penelitian ini diharapkan dapat memberikan manfaat bagi berbagai pihak, antara lain :

### **1.5.1 Bagi Penulis**

1. Memenuhi salah satu syarat untuk memperoleh gelar Sarjana Komputer (S.Kom.) pada Program Studi Teknik Informatika.
2. Mendapat pengalaman riset praktis di bidang informatika yang mengintegrasikan deep learning dan arsitektur transformer.
3. Menerapkan pengetahuan teoritis yang diperoleh selama perkuliahan ke dalam sebuah proyek pengembangan model yang nyata.

### **1.5.2 Bagi Universitas**

1. Menambah literatur ilmiah dan koleksi penelitian di lingkungan universitas mengenai penerapan AI modern dalam bioinformatika.
2. Menghasilkan sebuah aset penilitian berupa model aligner fungsional yang dapat menjadi peluang untuk dikembangkan pada peneletian selanjutnya.
3. Menjadi kolaborasi antara lingkungan akademis universitas dengan lembaga riset nasional seperti BRIN.

### **1.5.3 Bagi Pembaca**

1. Memberikan informasi dan wawasan mengenai penerapan praktis arsitektur Transformer untuk menyelesaikan masalah Multiple Sequence Alignment.
2. Menyediakan studi kasus yang dapat menjadi referensi bagi mahasiswa atau peneliti lain yang tertarik pada topik yang serupa.

# **Metodologi Penelitian**

Untuk menyelesaikan penelitian ini penulis menggunakan metodologi, yaitu:

### **1.6.1 Metode Pengumpulan Data**

Metode pengumpulan data yang penulis lakukan ialah :

1. Studi Literatur

Penulis mengumpulkan landasan teori dari berbagai sumber ilmiah seperti jurnal, buku, dan laporan penelitian sebelumnya yang berhubungan untuk memperkuat dasar konseptual penelitian.

1. Pemanfaatan Pipeline

Menggunakan data primer yang dihasilkan oleh pipeline pembuat dataset sintetis dari penelitian sebelumnya, yang berfungsi sebagai data latih dan data uji untuk model.

### **1.6.2 Metode Implementasi**

Untuk mencapai tujuan penelitian, metode implementasi yang digunakan adalah pendekatan penelitian eksploratif dan iteratif. Pendekatan ini dibagi menjadi tiga tahapan utama, yaitu:

1. Tahap investigasi

Meliputi studi literatur dan analisis awal terhadap dataset.

1. Tahap Perancangan

Mencakup desain arsitektur model dan skenario eksperimen

1. Tahap Implementasi

Terdiri dari pengembangan kode, pelatihan model, dan evaluasi kinerja.

# **Sistematika Penulisan**

Penulis menyusun penelitian ini dengan sistematika penulisan yang akan melakukan pembahasan menjadi 6 bab, yaitu :

**BAB I PENDAHULUAN**

Bab ini berisi latar belakang masalah, perumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metodologi penelitian, dan sistematika penulisan.

**BAB II LANDASAN TEORI**

Bab ini berisi teori-teori yang berhubungan dengan konsep penulisan seperti teori dasar, dan beberapa konsep penting mengenai topik penelitian.

**BAB III METODOLOGI PENELITIAN**

Bab ini berisi langkah-langkah metode yang akan digunakan dalam metodologi penelitian.

**BAB IV IMPLEMENTASI PENELITIAN**

Bab ini menguraikan mengenai penyelesaian masalah berdasarkan metodologi yang telah dipilih serta berisi proses implementasi dari metode tersebut.

**BAB V HASIL DAN PEMBAHASAN**

Bab ini menyajikan hasil dari pengujian model yang telah diimplementasikan. Seluruh temuan akan dianalisis dan dibahas secara mendalam untuk menjawab rumusan masalah.

**BAB VI PENUTUP**

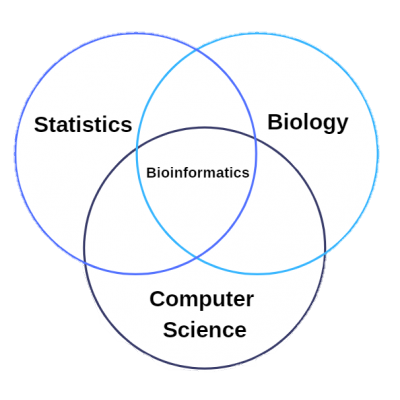
Bab ini berisi kesimpulan dan saran dari penelitian yang telah dilakukan yang dapat diperhatikan untuk penelitian selanjutnya yang lebih baik lagi.

# **BAB II**

# **LANDASAN TEORI**

# **Bioinformatika**

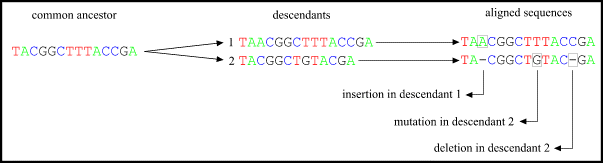
Bioinformatika adalah bidang ilmu interdisipliner yang mengintegrasikan biologi dengan ilmu komputer dan statistika untuk menganalisis data biologis dalam skala besar (Almanza-Ruiz et al., 2023). Di antara beragam jenis data yang diolah, informasi yang paling fundamental adalah data sekuens, sehingga kemampuan untuk memproses dan memahami untaian molekuler ini menjadi pilar utama dalam riset bioinformatika (Reddy & Fields, 2022).



Gambar 2. Ilustrasi bioinformatika dalam diagram venn.

# **Analisis Sekuens Biologis**

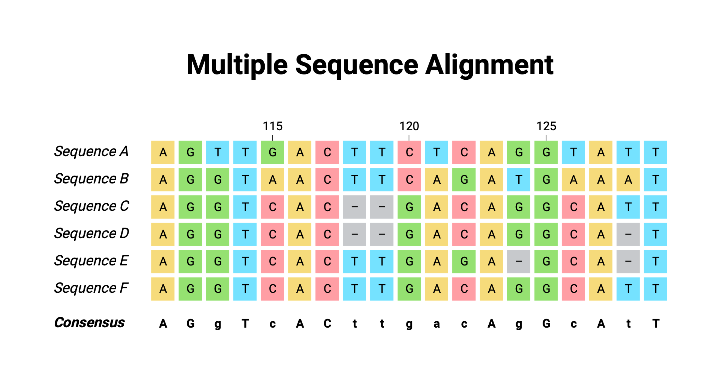
Analisis sekuens biologis berfokus pada studi mengenai untaian (sekuens) dari molekul biologis utama: DNA, RNA, dan protein, yang membawa instruksi untuk berbagai proses kehidupan (Victor Aprilyanto & Langkah Sembiring, 2016). Inti dari analisis ini adalah membandingkan sekuens untuk memahami hubungan evolusioner di antara keduanya, yang tecermin dari perubahan seperti mutasi, insersi, dan delesi. Untuk mendapatkan gambaran yang utuh, peneliti perlu membandingkan banyak sekuens secara bersamaan, sebuah tugas yang memerlukan metode Multiple Sequence Alignment (MSA)



Gambar 2. 2 Contoh Sekuens Biologis (diadaptasi dari https://users.ugent.be/~avierstr/principles*)*

# **Multiple Sequence Alignment (MSA)**

*Multiple Sequence Alignment* (MSA) adalah proses komputasi untuk menata ulang tiga atau lebih sekuens biologis ke dalam sebuah format matriks dengan menyisipkan karakter celah (*gap*) agar semua sekuens memiliki panjang yang sama (referensi 4, Reddy & Fields). Tujuan biologis utamanya adalah untuk memastikan bahwa setiap kolom berisi residu-residu yang homolog, atau berasal dari posisi yang sama pada satu leluhur bersama (Edgar & Batzoglou, 2006). Hasil dari MSA dapat menampakkan daerah-daerah lestari (*conserved regions*) yang penting untuk fungsi, serta pola-pola insersi dan delesi yang merepresentasikan sejarah evolusi.



Gambar 2. Contoh Multiple Sequence Alignment (Sumber : https://www.biorender.com/template/multiple-sequence-alignment-dna)

# **Tinjauan metode MSA Konvensional**

Karena menemukan penjajaran MSA yang optimal secara matematis adalah masalah yang sangat sulit (NP-complete), maka dikembangkanlah berbagai pendekatan heuristik untuk menghasilkan solusi yang "cukup baik" dalam waktu yang wajar (referensi 4, Reddy & Fields). Dua pendekatan yang paling dominan adalah metode progresif dan metode iteratif. Metode progresif, seperti pada ClustalW, membangun penjajaran secara bertahap berdasarkan pohon pemandu, namun kesalahan di tahap awal tidak dapat diperbaiki. Untuk mengatasinya, metode iteratif, seperti pada MAFFT, menyempurnakan penjajaran awal secara berulang-ulang untuk mendapatkan hasil yang lebih akurat (referensi 24, Xie et al.).

Perbandingan utama antara kedua pendekatan tersebut dapat dilihat dalam tabel berikut :

Tabel 2. 1 Perbandingan Metode MSA Konvensional

|  |  |  |
| --- | --- | --- |
| Fitur | Metode Progresif | Metode Iteratif |
| Prinsip Kerja | Alignment bertahap berdasarkan guide tree. | Penyempurnaan berilang dari MSA awal. |
| Contoh | ClustalW, T-Coffe | MAFFT, MUSCLE |
| Kelebihan | Cepat untuk dataset sederhana. | Umumnya lebih akurat. |
| Kekurangan | Kesalahan di tahap awal bersifat permanen. | Membutuhkan waktu komputasi lebih lama. |

Meskipun berbeda, semua metode konvensional ini bergantung pada skema penilaian yang statis dan tidak dapat beradaptasi dengan pola-pola evolusi yang unik dari data yang sedang dianalisis (Edgar, 2022). Keterbatasan untuk belajar dari data secara dinamis inilah yang menjadi celah utama, yang mendorong para peneliti untuk mencari paradigma baru.. Selain itu, performa metode-metode ini cenderung menurun drastis saat dihadapkan pada dataset yang sangat besar atau memiliki tingkat evolusi yang tinggi (Smirnov & Warnow, 2021). Keterbatasan untuk belajar dari data secara dinamis inilah yang menjadi celah utama, yang mendorong para peneliti untuk mencari paradigma baru.

# **Memandang Sekuens Biologis sebagai Sebuah Bahasa**

# **Artificial Intielligence (AI) dan Deep learning**

# **Sequence-to-sequence (S2S)**

# **Arsitektur Transformer**

# **Mekanisme Self-Attention**

# **Transfer Learning dan Fine-Tuning**

# **Dataset Sintetis dan Ground Truth**

# **Studi Literatur terkait**

# **BAB III METODOLOGI PENELITIAN**

## **3.1 Objek Penelitian**

Objek dalam penelitian ini melakukan

## **3.2 Metode Pengumpulan data**

### **3.2.1 Studi Pustaka**

### **3.2.2 Pemanfaat Pipeline Dataset**

### **3.3.3 Wawancara**

## **3.3 Alat dan Bahan Penelitian**

### **3.3.1 Perangkat Lunak**

### **3.3.2 Sumber Data**

## **3.4 Prosedur Penelitian**

### **3.4.1 Tahap Investigasi (menjelaskan tentang dataset dan pemodelan secara general)**

### **3.4.2 Tahap Perancangan Model (merujuk ke paper utama transformer)**

### **3.4.3 Tahap Implementasi dan Pelatihan**

### **3.4.4 Tahap Evaluasi**

## **3.5 Flowchart Penelitian**

# **DAFTAR PUSTAKA**

Almanza-Ruiz, S. H., Chavoya, A., & Duran-Limon, H. A. (2023). Parallel protein multiple sequence alignment approaches: a systematic literature review. *The Journal of Supercomputing*, *79*(2), 1201–1234. https://doi.org/10.1007/s11227-022-04697-9

Dotan, E., Belinkov, Y., Avram, O., Wygoda, E., Ecker, N., Alburquerque, M., Keren, O., Loewenthal, G., & Pupko, T. (2023). MULTIPLE SEQUENCE ALIGNMENT AS A SEQUENCE-TO-SEQUENCE LEARNING PROBLEM. *ICLR 2023*.

Dotan, E., Wygoda, E., Ecker, N., Alburquerque, M., Avram, O., Belinkov, Y., & Pupko, T. (2024). BetaAlign: a deep learning approach for multiple sequence alignment. *Bioinformatics*, *41*(1). https://doi.org/10.1093/bioinformatics/btaf009

Edgar, R. C. (2022). Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nature Communications*, *13*(1), 6968. https://doi.org/10.1038/s41467-022-34630-w

Edgar, R. C., & Batzoglou, S. (2006). Multiple sequence alignment. *Current Opinion in Structural Biology*, *16*(3), 368–373. https://doi.org/10.1016/j.sbi.2006.04.004

Gotoh, O., Morita, M., & Nelson, D. R. (2014). Assessment and refinement of eukaryotic gene structure prediction with gene-structure-aware multiple protein sequence alignment. *BMC Bioinformatics*, *15*(1), 189. https://doi.org/10.1186/1471-2105-15-189

Korosteleva, M., & Lee, S.-H. (2021). *Generating Datasets of 3D Garments with Sewing Patterns*.

Loewenthal, G., Rapoport, D., Avram, O., Moshe, A., Wygoda, E., Itzkovitch, A., Israeli, O., Azouri, D., Cartwright, R. A., Mayrose, I., & Pupko, T. (2021). A Probabilistic Model for Indel Evolution: Differentiating Insertions from Deletions. *Molecular Biology and Evolution*, *38*(12), 5769–5781. https://doi.org/10.1093/molbev/msab266

Madan, S., Lentzen, M., Brandt, J., Rueckert, D., Hofmann-Apitius, M., & Fröhlich, H. (2024). Transformer models in biomedicine. *BMC Medical Informatics and Decision Making*, *24*(1), 214. https://doi.org/10.1186/s12911-024-02600-5

Reddy, B., & Fields, R. (2022). *Multiple Sequence Alignment Algorithms in Bioinformatics* (pp. 89–98). https://doi.org/10.1007/978-981-16-4016-2\_9

Smirnov, V., & Warnow, T. (2021). MAGUS: Multiple sequence Alignment using Graph clUStering. *Bioinformatics*, *37*(12), 1666–1672. https://doi.org/10.1093/bioinformatics/btaa992

Trost, J., Haag, J., Höhler, D., Jacob, L., Stamatakis, A., & Boussau, B. (2023). *Simulations of sequence evolution: how (un)realistic they are and why*. https://doi.org/10.1101/2023.07.11.548509

Victor Aprilyanto, & Langkah Sembiring. (2016). *Filogenetika Molekuler: Teori dan Aplikasi*.

Trost, J., Haag, J., Höhler, D., Jacob, L., Stamatakis, A., & Boussau, B. (2023). Simulations of sequence evolution: how (un)realistic they are and why. https://doi.org/10.1101/2023.07.11.548509