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

Segala puji syukur kehadirat Allah *Subhanallahu Wata'ala* atas segala hidayah, karunia, dan rahmat-Nya, sehingga penulis dapat menyelesaikan penelitian dan penyusunan skripsi yang berjudul "**PENERAPAN MODEL *SEQUENCE-TO-SEQUENCE* BERBASIS *TRANSFORMER* UNTUK *MULTIPLE SEQUENCE ALIGNMENT***" ini dengan baik dan *In Syaa* Allah penuh dengan keberkahan. Shalawat dan salam senantiasa tercurah kepada Baginda Nabi Muhammad *Shallahu 'Alaihi Wasallam*, beserta keluarga, sahabat, dan para pengikutnya hingga akhir zaman.

Penyelesaian skripsi ini tidak akan terwujud tanpa adanya bantuan, bimbingan, dukungan, serta doa dari berbagai pihak. Oleh karena itu, dengan segala kerendahan hati, penulis ingin menyampaikan terima kasih yang sebesar-besarnya kepada :

1. Kedua orang tua tercinta, Bapak Slamet dan Ibu Iin Sarinah, yang senantiasa memberikan doa yang tak terputus, motivasi, dan dukungan moril terbesar bagi penulis.
2. Bapak Husni Teja Sukmana, S.T., M.Sc, Ph.D, Selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Syarif Hidayatullah Jakarta.
3. Ibu Dewi Khairani, M.Sc, selaku Ketua Program Studi dan Bapak Saepul Aripiyanto M.Kom., Selaku Sekretaris Program Studi Teknik Informatika.
4. Ibu Fenty Eka Muzayyana Agustin, M.Kom., selaku Dosen Pembimbing 1, yang dengan penuh kesabaran telah memberikan bimbingan, arahan, dan masukan yang sangat berharga.
5. Bapak Dr. Achmad Fatchuttamam Abka, M.Kom., selaku Dosen Pembimbing 2, yang wawasan dan perspektifnya yang mendalam telah secara fundamental membentuk arah dan mempertajam fokus penelitian ini. Diskusi-diskusi berharga dengan beliau menjadi sumber inspirasi yang tak ternilai.
6. Ibu Maulida Mazaya, Ph.D., dan Bapak I Wayan Aditya Swardiana, M.Kom., yang telah berperan penting sebagai narasumber ahli. Bimbingan

teknis dan masukan-masukan krusial dari keduanya sangat membantu penulis dalam mengatasi berbagai tantangan praktis selama penelitian.

7. M Oskhar Mubarok dan M Fatihul Choir, yang telah menjadi teman diskusi, banyak membantu dalam pengerjaan teknis, serta senantiasa memberikan dukungan dselama proses pengerjaan skripsi.
8. Serta seluruh pihak yang tidak dapat penulis sebutkan satu per satu, yang telah memberikan dukungan dan bantuan dalam berbagai bentuk selama proses penelitian ini.

Penulis menyadari sepenuhnya bahwa skripsi ini masih jauh dari kesempurnaan. Oleh karena itu, segala bentuk kritik dan saran yang membangun akan diterima dengan lapang dada demi perbaikan di masa mendatang. Akhir kata, semoga skripsi ini dapat memberikan manfaat dan kontribusi positif bagi pengembangan ilmu pengetahuan.

Jakarta, 12 Agustus 2025

Tanjung Arswendo Yudha

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I
PENDAHULUAN

1.1 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-residu 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 et al., 2014), unggul karena kemampuannya untuk mempelajari pola dan tata bahasa secara langsung dari data teks. 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 (Vaswani et al., 2023). 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).

1.2 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?

1.3 Batasan Masalah

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

1.3.1 Batasan Model dan Arsitektur

1. Penelitian ini secara spesifik hanya akan merancang dan menguji model dengan arsitektur *Transformer*, dan tidak akan melakukan perbandingan dengan arsitektur *deep learning* lain seperti *Convolutional Neural Networks* (CNN) atau *Recurrent Neural Networks* (RNN).
2. *Output* akhir dari penelitian ini adalah sebuah model *aligner* fungsional untuk tujuan riset, bukan sebuah perangkat lunak (*software*) yang telah dioptimalkan untuk pengguna akhir.

1.3.2 Batasan *Dataset*

1. Model yang akan dibangun hanya akan dilatih dan dievaluasi menggunakan *dataset* sintetis yang dihasilkan dari pipeline simulasi.
2. Penelitian ini tidak akan menggunakan *dataset* biologis empiris (data nyata dari organisme) untuk tahap pelatihan model.

3. Data biologis hanya menjadi rujukan untuk pembuatan *dataset* sintetis.

1.3.3 Batasan Evaluasi

1. Evaluasi kinerja model akan difokuskan pada metrik akurasi utama, yaitu *Column Score* (CS).
2. Perbandingan kinerja akan dilakukan terhadap satu atau dua metode konvensional yang representatif (misalnya MAFFT) sebagai *baseline*, bukan terhadap semua *tools* MSA yang ada.

1.4 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.

1.5 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 penelitian 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.

1.6 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.
2. 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*.

2. Tahap Perancangan

   Mencakup desain arsitektur model dan skenario eksperimen

3. Tahap Implementasi

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

1.7 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.

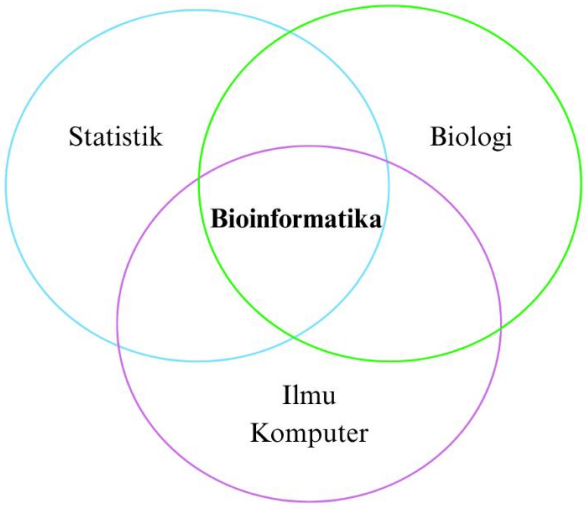
BAB VI PENUTUP

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

2.1 Bioinformatika

Bioinformatika adalah bidang ilmu interdisipliner yang mengintegrasikan biologi dengan ilmu komputer dan statistika (Gambar 2.1) 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).

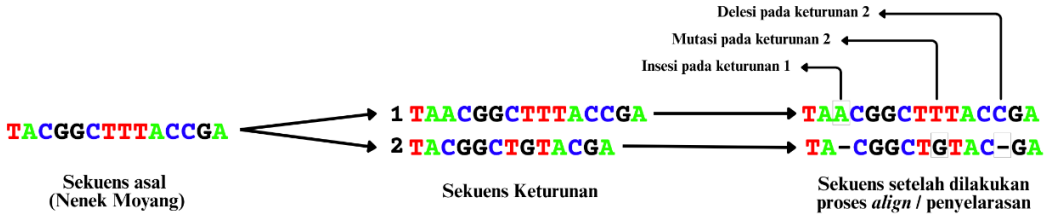


Gambar 2. 1 Ilustrasi bioinformatika dalam diagram venn.

2.2 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 tercermin dari perubahan seperti mutasi (perubahan), insersi (penyisipan), dan delesi (penghapusan). Proses insersi dan delesi (indel) ini sendiri memiliki dinamika evolusi yang kompleks dan dapat dimodelkan secara probabilistik untuk menghasilkan data yang realistis

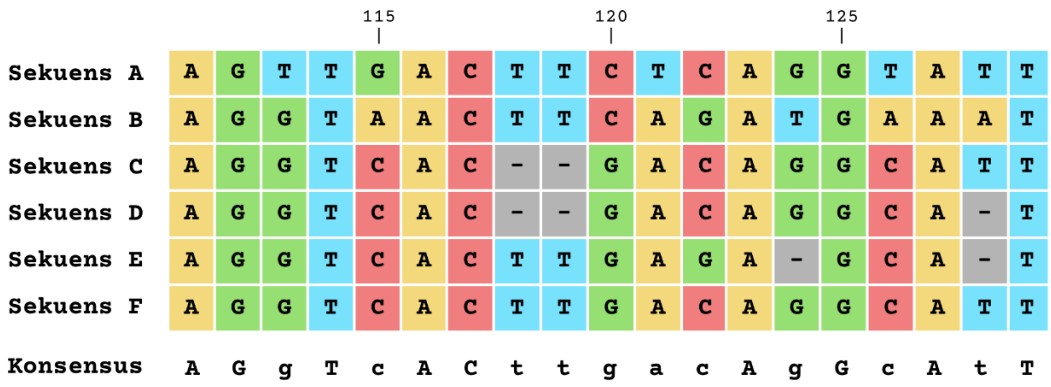
(Loewenthal et al., 2021). Untuk mendapatkan gambaran yang utuh, peneliti perlu membandingkan banyak sekuens secara bersamaan, sebuah tugas yang memerlukan metode *Multiple Sequence Alignment (MSA)*. Untuk lebih jelasnya, bisa dilihat pada contoh sekuens biologis pada Gambar 2.2. dibawah ini :



Gambar 2. 2 Contoh Sekuens Biologis

2.3 *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 (Reddy & Fields, 2022). 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 2.3 dibawah ini merupakan contoh dari *Multiple Sequence Alignment* (MSA).



Gambar 2. 3 Contoh Multiple Sequence Alignment

2.4 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 (Reddy & Fields, 2022). 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 (Xie et al., 2023).

Perbandingan utama antara kedua pendekatan tersebut dapat dilihat dalam tabel 2.1 berikut :

Tabel 2. 1 Perbandingan Metode MSA Konvensional

Fitur	Metode Progresif	Metode Iteratif
Prinsip Kerja	Alignment bertahap berdasarkan *guide tree*.	Penyempurnaan berulang dari MSA awal.
Contoh	ClustalW, T-Coffee	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).

2.5 Memandang Sekuens Biologis sebagai Sebuah Bahasa

Keterbatasan metode konvensional mendorong para peneliti untuk mencari paradigma baru yang lebih dinamis. Salah satu pergeseran cara pandang yang paling fundamental adalah dengan memandang sekuens biologis sebagai sebuah bahasa. Perspektif ini mengasumsikan bahwa untaian asam amino tidak hanya sekadar barisan karakter, tetapi juga memiliki struktur, aturan, dan konteks yang mirip dengan bahasa manusia, sehingga membuka pintu untuk mengaplikasikan teknik canggih yang awalnya dikembangkan untuk pemrosesan (Chandra et al., 2023). Keterkaitan konseptual yang kuat antara kedua domain tersebut dapat diringkas dalam perbandingan pada Tabel 2.2. berikut ini

Tabel 2. 2 Perbandingan Konseptual antara Bahasa dan Sekuens Biologis

Konsep	Dalam Bahasa & Liguistik	Dalam Sekuens Biologis
Unit Dasar	Huruf / Karakter	Nukleotida / Asam Amino
Struktur Hierarkis	Huruf → Kata → Kalimat	Nukleotida → Kodon → Gen
Ketergantungan Kontekstual	Makna sebuah kata ditentukan oleh kata – kata disekitarnya.	Fungsi sebuah kodon atau situs pengikat ditentukan oleh nukleotida di sekitarnya.
Dependensi Jarak Jauh	Hubungan gramatikal antara bagian awal dan akhir kalimat.	Intraksi fungsional antara elemen sekuens yang letaknya berjauhan (misal : promoter & enhancer).
Grammar	Aturan sintaksis dan semantik yang mengetur struktur kalimat.	Aturan evolusioner dan biofisik yang mengatur pola mutasi dan indel (Ferruz & Höcker, 2022).

Berdasarkan kesamaan konseptual ini, masalah *Multiple Sequence Alignment (MSA)* dapat diformulasikan ulang dari masalah optimasi menjadi sebuah masalah penerjemahan. Paradigma inilah yang menjadi landasan logis untuk menerapkan model-model canggih dari bidang Artificial Intelligence dan *Deep Learning*.

2.6 *Artificial Intelligence* (AI) dan *Deep learning*

Setelah menetapkan bahwa MSA dapat dipandang sebagai masalah penerjemahan, maka diperlukan pendekatan komputasi yang mampu "belajar" untuk melakukan tugas tersebut. Pendekatan ini datang dari bidang Artificial Intelligence (AI), yaitu cabang ilmu komputer yang bertujuan untuk menciptakan sistem yang dapat meniru kemampuan kognitif manusia. Di dalam AI, pendekatan yang paling dominan saat ini adalah *Deep Learning*, yang memanfaatkan jaringan saraf tiruan berlapis-lapis (*deep neural networks*) untuk belajar secara langsung dari data dalam jumlah besar (Islam et al., 2023). Keunggulan utama *Deep Learning* adalah kemampuannya untuk mengekstrak fitur dan pola-pola rumit dari data mentah secara otomatis. Contohnya, model deep learning berbasis Transformer telah terbukti mampu memprediksi fungsi enzim secara akurat hanya dari sekuens protein dengan cara mempelajari motif-motif fungsionalnya (Kim et al., 2023). Hal ini menjadikannya sangat cocok untuk memecahkan masalah bioinformatika yang sulit diatasi oleh metode konvensional (Chandra et al., 2023).

2.7 Sequence-to-sequence (S2S)

Salah satu paradigma paling kuat yang lahir dari pendekatan *Deep Learning* untuk memproses data sekuensial adalah *Sequence-to-Sequence (S2S)*. S2S adalah sebuah kerangka kerja yang dirancang untuk memecahkan masalah di mana sebuah sekuens *input* dengan panjang variabel perlu diubah menjadi sekuens *output* yang panjangnya juga bisa berbeda (Sutskever et al., 2014). Model S2S mulai menunjukkan kemajuan signifikan saat mekanisme atensi (*attention*) diperkenalkan, yang memungkinkan model untuk secara selektif fokus pada bagian *input* yang paling relevan saat menghasilkan *output* (Bahdanau et al., 2016). Aplikasi asli dan paling terkenal dari paradigma ini adalah pada tugas penerjemahan mesin (*machine translation*), misalnya menerjemahkan kalimat dari Bahasa Inggris ke Bahasa Indonesia.

Secara konseptual, model s2s terdiri dari 2 komponen utama :

1. *Encoder*

Sebuah jaringan saraf yang membaca keseluruhan sekuens *input* dan mengompres informasinya menjadi sebuah representasi vektor dengan dimensi yang tetap, yang sering disebut sebagai *context vector*.

2. *Decoder*

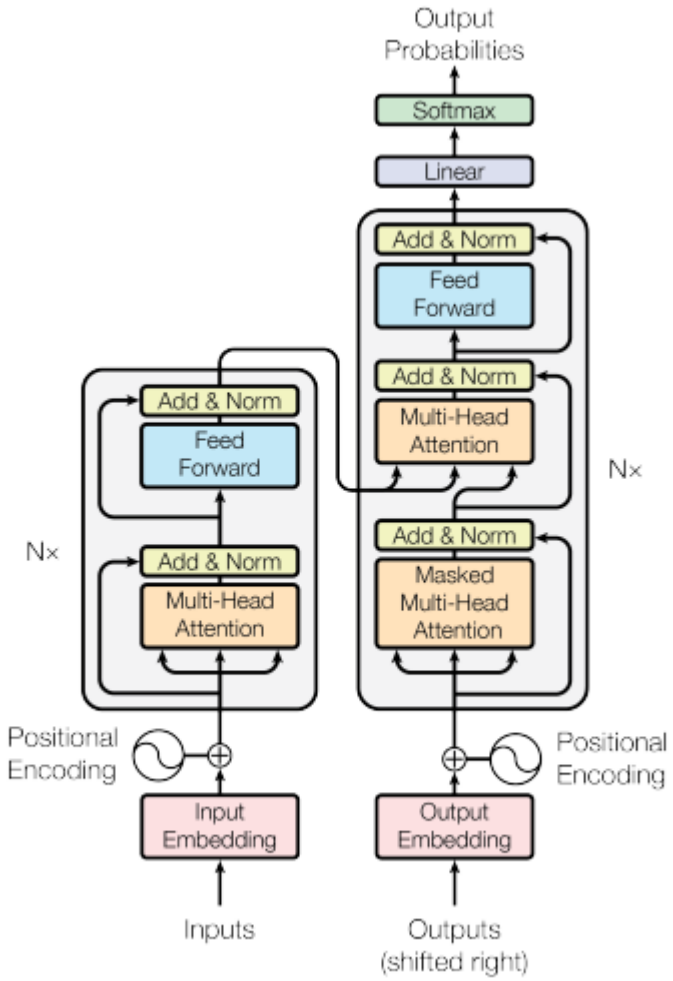
   Sebuah jaringan saraf kedua yang mengambil *context vector* tersebut sebagai titik awal, lalu menghasilkan sekuens *output* elemen per elemen hingga selesai.

Pada penelitian ini, paradigma S2S diadaptasi secara langsung untuk masalah MSA. Sekumpulan sekuens biologis yang belum selaras digabungkan menjadi satu sekuens *input* tunggal untuk dibaca oleh *Encoder*. Selanjutnya, *Decoder* dilatih untuk menghasilkan sekuens *output* yang merepresentasikan MSA yang utuh, lengkap dengan penyisipan karakter celah (gap) di posisi yang tepat (Dotan et al., 2023). Meskipun model S2S awalnya menggunakan arsitektur seperti LSTM, sebuah arsitektur yang lebih baru dan kuat telah terbukti jauh lebih efektif untuk tugas ini, yaitu arsitektur *Transformer*.

2.8 Arsitektur *Transformer*

Arsitektur *Transformer* adalah sebuah model deep learning yang diperkenalkan dalam paper "Attention Is All You Need" dan secara fundamental merevolusi cara mesin memproses data sekuensial (Vaswani et al., 2023). Berbeda dengan model-model sebelumnya yang mengandalkan pemrosesan data secara berurutan (rekuren), *Transformer* mampu memproses seluruh token dalam sebuah sekuens secara bersamaan. Kemampuan ini tidak hanya membuatnya jauh lebih efisien dalam hal waktu pelatihan, tetapi juga lebih unggul dalam menangkap hubungan-hubungan kompleks di dalam data (Zeyer et al., 2019).

Secara garis besar, arsitektur *Transformer* mempertahankan struktur *Encoder-Decoder* dari model S2S pendahulunya, namun dengan komponen internal yang sepenuhnya baru. Gambar 2.4 dibawah ini merupakan gambaran dari arsitektur *Transformer*



Gambar 2. 4 Arsitektur *Transformer* (Vaswani et al., 2023)

1. *Encoder*

   Terdiri dari tumpukan (stack) lapisan identik yang bertugas untuk membaca seluruh sekuens *input* dan membangun representasi kontekstual untuk setiap token. Representasi ini kaya akan informasi karena setiap token diproses dengan mempertimbangkan hubungannya dengan semua token lain dalam sekuens.

2. *Decoder:*

   Juga terdiri dari tumpukan lapisan yang bertugas untuk mengambil representasi dari *Encoder* dan menghasilkan sekuens *output* token per token secara *auto-regressive* (menghasilkan token berikutnya berdasarkan token yang sudah dihasilkan sebelumnya).

Kemampuan *Encoder* dan *Decoder* untuk memproses informasi secara holistik ini dimungkinkan oleh sebuah komponen fundamental yang menjadi inti dari setiap lapisannya, yaitu sebuah mekanisme yang disebut *self-attention (Borhani et al., 2022)*.

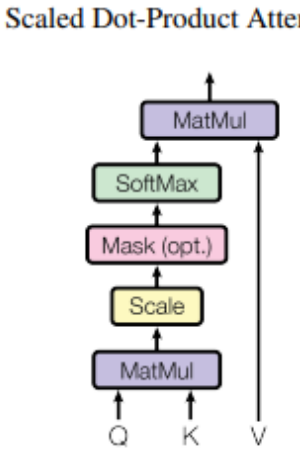
2.9 Mekanisme Self-Attention

Self-attention adalah mekanisme komputasi yang menjadi inti dari arsitektur *Transformer*. Mekanisme ini memungkinkan model untuk menimbang tingkat kepentingan dari setiap token dalam sebuah sekuens *input* ketika memproses token lainnya di sekuens yang sama. Untuk setiap token, model menghasilkan tiga vektor terpisah: *Query* (Q), *Key* (K), dan *Value* (V). Dan dapat dirumuskan secara matematis sebagai berikut :

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \mathit{softmax}\left(\frac{(\text{QK}^{\text{T}})}{\sqrt{\text{d}_{\text{k}}}}\right) \text{V}$$

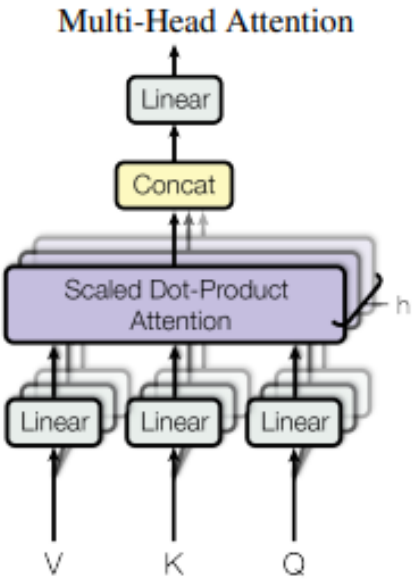
Dalam rumus tersebut, skor atensi dihitung dari perkalian titik antara *Query* (Q) dan *Key* (K), yang kemudian diskalakan dengan akar dari dimensi key (dk) dan dinormalisasi menggunakan fungsi *softmax* untuk menghasilkan bobot. Bobot ini kemudian digunakan untuk mengagregasi *Value* (V), menghasilkan *output* atensi yang kontekstual.

Proses dasar *self-attention*, yang dikenal sebagai *Scaled Dot-Product Attention*, yang diilustrasikan pada Gambar 2.5 berikut ini :



Gambar 2. 5 Mekanisme *Scaled Dot-Product Attention* sebagai unit dasar *self-attention*.(Vaswani et al., 2023)

Dalam implementasinya, *Transformer* menjalankan beberapa mekanisme atensi ini secara paralel, sebuah konsep yang disebut *Multi-Head Attention* (Gambar 2.6). Hal ini memungkinkan model untuk mempelajari berbagai jenis hubungan antar token secara bersamaan dari berbagai 'sudut pandang' representasi yang berbeda (Singh & Mahmood, 2021).



Gambar 2. 6 Mekanisme *Multi-Head Attention*(Vaswani et al., 2023)

2.10 *Transfer Learning* dan *Fine-Tuning*

Transfer learning adalah sebuah strategi dalam machine learning di mana sebuah model yang telah dilatih pada satu tugas besar dan umum, digunakan kembali sebagai titik awal untuk tugas lain yang lebih spesifik. Pendekatan ini sangat efektif karena model dasar (*pre-trained model*) telah mempelajari representasi fitur yang kaya dan berguna dari data dalam jumlah besar, sehingga tidak perlu lagi melatih model dari nol (Singh & Mahmood, 2021). Konsep ini menjadi dasar dari pengembangan *foundational models*, yaitu model-model berskala sangat besar yang dilatih pada data masif dan kemudian dapat diadaptasi untuk berbagai tugas turunan (Willemink et al., 2022).

Pada penelitian ini, strategi *transfer learning* akan diterapkan melalui proses *Fine-Tuning*. Proses ini melibatkan dua tahap utama:

1. Mengambil *Pre-trained Model*
   Menggunakan model *Transformer* yang telah dilatih sebelumnya pada korpus data sekuens biologis yang sangat besar. Model ini sudah memiliki "pemahaman" dasar mengenai "tata bahasa" sekuens biologis.
2. Melatih Ulang (*Fine-Tuning*)
   Model dasar tersebut kemudian dilatih kembali dalam skala yang lebih kecil menggunakan *dataset* sintetis spesifik yang telah kita siapkan. Pada tahap ini, model akan beradaptasi dan menyesuaikan "pengetahuannya" untuk tugas spesifik MSA.

Pendekatan ini tidak hanya menghemat waktu dan sumber daya komputasi secara signifikan, tetapi juga seringkali menghasilkan model dengan kinerja yang lebih baik dan kemampuan generalisasi yang lebih kuat.

2.11 *Dataset* Sintetis dan *Ground Truth*

Keberhasilan model deep learning dalam tugas terarah (*supervised learning*) sangat bergantung pada ketersediaan data pelatihan yang besar dan memiliki anotasi yang benar. Ini berarti diperlukan contoh-contoh penjajaran yang "benar" secara evolusioner. Terdapat dua sumber utama untuk mendapatkan data semacam

ini, yaitu *dataset* empiris dan *dataset* sintetis, yang masing-masing memiliki kelebihan dan kekurangan yang dirangkum pada Tabel 2.3.

Tabel 2. 3 Perbandingan *Dataset* Empiris dan Sintetis

Fitur	*Dataset* Empiris	*Dataset* Sintetis
Sumber	Eksperimen Biologis nyata, dikurasi secara manual.	Dihasilkan melalui simulasi komputasi berdasarkan model evolusi.
Skala / Jumlah	Terbatas, sulit dan mahal untuk diperbanyak.	Dapat diproduksi secara massal dengan jumlah sangat besar.
Kebenaran (*Ground Truth*)	Dibuat melalui inferensi, kebenaran tidak dijamin 100% (Gotoh et al., 2014).	Kebenaran absolut, karena seluruh proses evolusi diketahui secara pasti (Loewenthal et al., 2021).
Kelebihan Utama	Mempresentasikan realitas biologis yang sebenarnya.	Menyediakan "kunci jawaban" yang sempurna untuk *supervised learning*.
Kekurangan Utama	Seringkali tidak cukup besar dan banyak, serta kebenarannya diragukan	Berpotensi memiliki celah realisme edngan data biologis nyata (Trost et al., 2023).

Berdasarkan perbandingan tersebut, meskipun *dataset* empiris sangat penting untuk validasi akhir, keterbatasannya dalam hal skala dan ketiadaan *ground truth* yang absolut membuatnya kurang ideal untuk melatih model deep learning berskala besar dari awal. Oleh karena itu, penelitian ini memilih untuk menggunakan *dataset* sintetis. Pendekatan ini menyediakan pasangan *input* dan *output* dengan *ground truth* yang sempurna, yang esensial untuk melatih model secara efektif dan merupakan strategi yang paling layak untuk menghasilkan data dalam skala masif yang dibutuhkan oleh arsitektur *Transformer* (Korosteleva & Lee, 2021).

2.12 SpartaABC

Untuk menghasilkan dataset sintetis yang realistis, diperlukan sebuah metode untuk memperkirakan parameter evolusi yang sesuai. SpartaABC adalah program yang dirancang untuk tujuan ini, dengan mengimplementasikan algoritma *Approximate Bayesian Computation* (ABC) untuk menyimpulkan (*infer*) parameter-parameter indel dari data sekuens yang diberikan. Dalam prosesnya,

SpartaABC menjalankan *ribuan simulasi sekuens* menggunakan simulator yang terintegrasi. Dengan beberapa penyesuaian, output dari simulasi-simulasi ini dapat dibuat agar menjadi pasangan file sekuens aligned (MSA ground truth) dan unaligned yang sangat cocok untuk dijadikan dataset pelatihan deep learning (Ashkenazy et al., 2017). Lalu dan dengan krip tambahan untuk otomasi, SpartaABC dapat difungsikan sebagai *generator* dalam sebuah pipeline yang sangat efektif untuk menghasilkan dataset dalam skala besar.

2.13 Studi Literatur terkait

Untuk memposisikan penelitian ini secara kokoh dalam konteks ilmiah yang ada, serta untuk membangun fondasi teoretis yang kuat, dilakukan tinjauan mendalam terhadap studi-studi literatur yang relevan. Tinjauan ini tidak hanya mencakup penelitian-penelitian fundamental yang memperkenalkan teknologi inti seperti Sequence-to-Sequence dan *Transformer*, tetapi juga mencakup paper ulasan (survey) yang memetakan perkembangan terkini, serta studi-studi aplikasi yang menunjukkan keberhasilan penerapan deep learning di domain-domain terkait. Perbandingan dari studi-studi literatur utama tersebut disajikan pada Tabel 2.4 untuk memberikan gambaran yang komprehensif.

Tabel 2. 4 Literatur Terkait

No	Penulis, Tahun	Metode	*Dataset*	Evaluasi Performa	Hasil
1.	Kim et al., 2023	*Deep Learning*, *Transformer*	UniProtKB/TrEMBL	*Precision, Recall*	Model *Transformer* berhasil memprediksi fungsi enzim dengan mempelajari motif fungsional dari sekuens.
2.	Islam et al., 2023	Tinjauan (Survey)	>600 model *Transformer*.	*Analisis Kualitatif*	Menyediakan taksonomi model *Transformer* berdasarkan domain aplikasi dan tugasnya.
3.	Trost et al., 2023	Machine Learning	Empiris vs. Sintetis	*Balanced Accuracy*	Menunjukkan adanya "celah realisme" antara data simulasi dan data biologis nyata.
4.	Chandra et al., 2023	Tinjauan (*Review*)	Repositori sekuens protein	Analisis Kualitatif	Menjelaskan bagaimana *Transformer* menjanjikan untuk mengungkap informasi tersembunyi dalam sekuens protein.
5.	Ferruz & Höcker, 2022	*Deep Learning*	Kajian Konseptual	Analisis Kualitatif	Membahas potensi model bahasa generatif untuk mendesain protein baru dengan fungsi yang terkontrol.

6.	Borhani et al., 2022	*Deep Learning*, Vision *Transformer* (ViT)	*Dataset* gambar daun	Akurasi, *F1-Score*	Menunjukkan keberhasilan adaptasi *Transformer* untuk computer vision, menyoroti trade-off akurasi vs kecepatan.
7.	Willemink et al., 2022	Tinjauan (*Review*)	Kajian Konseptual	Analisis Kualitatif	Membahas tantangan dan potensi penggunaan *Transformer* skala besar untuk analisis citra medis.
8.	Singh & Mahmood, 2021	Tinjauan (Survey)	Analisis Model	Analisis Kualitatif	Memberikan taksonomi model NLP berbasis *Transformer*, menyoroti tren menuju efisiensi komputasi.
9.	Loewenthal et al., 2021	Simulasi & Approximate Bayesian Computation (ABC).	Empiris & Sintetis (SpartaABC).	Analisis Statistik	Mengembangkan model realistis untuk simulasi evolusi sekuens yang menjadi dasar tool SpartaABC.
10.	Korosteleva & Lee, 2021	*Pipeline* Generatif Otomatis dengan Simulasi Fisika.	Kajian Konseptual	Analisis Kualitatif	Menghasilkan *pipeline* dan *dataset* sintetis berskala besar untuk melatih model deep learning pada objek 3D.
11.	Zeyer et al., 2019	Perbandingan Kinerja (*Benchmarking*).	Audio (*LibriSpeech*).	*Word Error Rate (WER).*	Menemukan bahwa *Transformer* lebih cepat dan lebih stabil saat pelatihan dibandingkan LSTM untuk tugas ASR.

12.	Vaswani et al., 2023	*Deep Learning*, *Transformer*	Teks (WMT'14)	BLEU *Score.*	Memperkenalkan arsitektur *Transformer* yang sepenuhnya berbasis *self-attention*, mengungguli model RNN/CNN.
13.	Bahdanau et al., 2016	*Deep Learning*, RNN, Atensi	Teks (WMT'14)	BLEU *Score.*	Memperkenalkan mekanisme atensi yang memungkinkan model S2S mengatasi masalah sekuens panjang.
14.	Sutskever et al., 2014	*Deep Learning*, LSTM	Teks (WMT'14)	BLEU *Score*	Menunjukkan bahwa LSTM *encoder-decoder* dapat memetakan sekuens *input* ke vektor dan menghasilkan *output*.
15.	(Chao et al., 2022)	Tinjauan (*Review*)	Literatur & *Benchmark* MSA	Analisis Kualitatif	Merangkum berbagai metode heuristik yang dikembangkan untuk mengatasi tantangan komputasi dari MSA, namun tetap memiliki *trade-off* akurasi dan kecepatan.
16.	(Ju et al., 2021)	*Deep Learning*	*Benchmark* (CASP13)	Presisi Kontak	Mengidentifikasi adanya **kehilangan informasi fundamental** pada metode konvensional dan mengusulkan pendekatan *Deep Learning* sebagai solusi.

Dari tinjauan yang dilakukan sebelumnya, dapat ditarik beberapa kesimpulan sebagai berikut:

1. Studi menunjukkan bahwa metode MSA konvensional memiliki keterbatasan fundamental dalam hal akurasi dan kecepatan, terutama saat dihadapkan pada dataset yang sangat besar di mana kinerjanya cenderung menurun drastis.
2. Arsitektur *Transformer* telah terbukti sangat sukses untuk berbagai tugas *deep learning* di banyak domain, termasuk di dalam domain bioinformatika.
3. Meskipun penerapan *Transformer* untuk MSA sebagai tugas *sequence-to-sequence* sudah mulai dieksplorasi, area ini masih tergolong baru dan terbuka untuk pengembangan lebih lanjut.

Oleh karena itu, penelitian ini bertujuan untuk mengisi celah tersebut dengan mengintegrasikan kekuatan arsitektur *Transformer*, paradigma S2S, dan metodologi data sintetis untuk membangun dan memvalidasi sebuah model *aligner* MSA modern.

BAB III
METODOLOGI PENELITIAN

3.1 Objek Penelitian

Objek penelitian ini terdiri dari dua komponen utama. Objek utamanya adalah sebuah model aligner *Multiple Sequence Alignment* (MSA) yang dibangun menggunakan arsitektur *deep learning Transformer*. Selain itu, yang menjadi objek pendukung adalah *dataset* sintetis yang berfungsi sebagai data pelatihan dan pengujian, di mana *dataset* ini merupakan *output* dari *pipeline* otomatis yang telah dikembangkan pada penelitian sebelumnya.

3.2 Metode Pengumpulan data

Pengumpulan data dalam penelitian ini dilakukan melalui beberapa pendekatan untuk memperoleh data sekunder berupa landasan teori dan data primer berupa *dataset* untuk pelatihan model.

3.2.1 Studi Pustaka

Studi pustaka dilakukan untuk mengumpulkan landasan teori yang relevan dari berbagai sumber ilmiah seperti jurnal, buku, dan laporan penelitian sebelumnya. Fokus dari studi ini adalah untuk memahami konsep-konsep kunci yang berkaitan dengan *Multiple Sequence Alignment*, metode-metode konvensional, arsitektur Transformer, paradigma *Sequence-to-Sequence*, serta teknik-teknik implementasi *deep learning* yang relevan.

3.2.2 Pemanfaat *Pipeline Dataset*

Data primer untuk penelitian ini dihasilkan (*generated*) melalui pemanfaatan *pipeline* komputasi yang telah dikembangkan pada penelitian sebelumnya. *Pipeline* ini berfungsi untuk menjalankan simulasi evolusi sekuens menggunakan program SpartaABC. Proses ini secara otomatis menghasilkan pasangan data yang terdiri dari sekumpulan sekuens *unaligned* sebagai *input* dan sekuens *aligned* yang benar secara evolusioner sebagai

output (*ground truth*), yang kemudian menjadi *dataset* utama untuk pelatihan dan pengujian model.

3.2.3 Wawancara

Metode wawancara akan digunakan sebagai salah satu cara untuk validasi kualitatif dalam penelitian ini. Wawancara akan dilakukan secara semi-terstruktur dengan narasumber yang merupakan ahli di bidang bioinformatika. Tujuan dari wawancara ini adalah untuk memperoleh masukan dan tinjauan ahli terhadap dua aspek utama penelitian:

1. Validasi Metodologi

   Memastikan bahwa skenario dan parameter yang digunakan dalam pipeline untuk menghasilkan dataset sintetis sudah relevan dan dapat diterima secara keilmuan.
2. Validasi Hasil

   Meninjau sampel output dari model aligner untuk mendapatkan penilaian kualitatif mengenai kualitas penjajaran yang dihasilkan.

3.3 Instrumen Penelitian

Penelitian ini memanfaatkan beberapa perangkat keras (*hardware*), perangkat lunak (*software*), dan sumber data spesifik yang rinciannya sebagai berikut:

3.3.1 Perangkat Keras (Hardware)

1. Proses pengembangan awal, modifikasi skrip, dan pengujian skala kecil dilakukan pada laptop Dell dengan spesifikasi Prosesor Intel(R) Core(TM) i5-8350U dan RAM 8192MB.
2. Tahap pelatihan model skala besar yang membutuhkan sumber daya yang intensif akan memanfaatkan server *High-Performance-Computing* (HPC) BRIN

3.3.2 Perangkat Lunak (Software)

1. Pengembangan dilakukan di atas sistem operasi Windows 11 Pro 64-bit.

2. Docker digunakan untuk membangun dan mengelola *environment* tervirtualisasi yang konsisten, yang berisi semua dependensi *pipeline* SpartaABC.
3. SpartaABC berfungsi sebagai *engine* simulasi utama untuk menjalankan proses evolusi sekuens yang akan dijadikan dataset.
4. Visual Studio Code (VS Code) digunakan sebagai editor teks untuk pengembangan dan modifiaksi skrip *pipeline*.
5. Google Colaboratory (Colab) digunakan sebagai platform cloud untuk melakukan prototyping dan eksperimen awal model untuk pengujian skala kecil.
6. Implementasi arsitektur Transformer akan menggunakan *framework* PyTorch dengan *library* Fairseq.

3.3.3 Sumber Data

1. Data primer yang digunakan dalam penelitian ini adalah dataset sintetis yang dihasilkan dari sebuah prototipe *pipeline* fungsional.
2. *Pipeline* tersebut telah diuji dengan menjalankan 6 skenario eksperimental yang dirancang untuk mempresentasikan berbagai tingkat kesulitan evolusioner.
3. Output dari *pipeline* tersebut adalah file data yang telah diagregasi dan siap digunakan, berisi pasangan sekuens *unaligned* dan MSA *ground truth (aligned)*
4. Data sekunder berupa studi litertur dari jurnal ilmiah, buku, dan laporan penelitian yang dijadikan sebagai landasan teori, dan referensi penelitian.

3.4 Prosedur Penelitian

Prosedur penelitian ini dilaksanakan secara sistematis mengikuti pedekatan penelitian eksploratif dan iteratir. Alur kerja dibagi menjadi 4 tahapan utama, mulai dari investigasi awal hingga evaluasi akhir model, seperti yang dijelaskan di bawah ini:

3.4.1 Tahap Investigasi

Tahap investigasi merupakan fase eksplorasi awal untuk mendefinisikan masalah dan mengidentifikasi solusi potensial. Kegiatan pada tahap ini berfokus pada studi literatur untuk memahami keterbatasan fundamental dari metode MSA konvensional. Berdasarkan analisis masalah tersebut, investigasi dilanjutkan dengan mencari paradigma alternatif dari bidang *Artificial Intelligence*, yang hasilnya mengidentifikasi arsitektur *Transformer* dengan paradigma *Sequence-to-Sequence* sebagai pendekatan yang paling menjanjikan.

3.4.2 Tahap Perancangan Model

Setelah fondasi teoretis dan data dipahami, tahap selanjutnya adalah perancangan model. Pada tahap ini, arsitektur spesifik dari model *Transformer Encoder-Decoder* dirancang, dengan merujuk pada implementasi yang berhasil pada penelitian sebelumnya (Dotan et al., 2023). Perancangan juga mencakup penentuan format representasi input dan output, serta penyusunan skenario eksperimen untuk pelatihan.

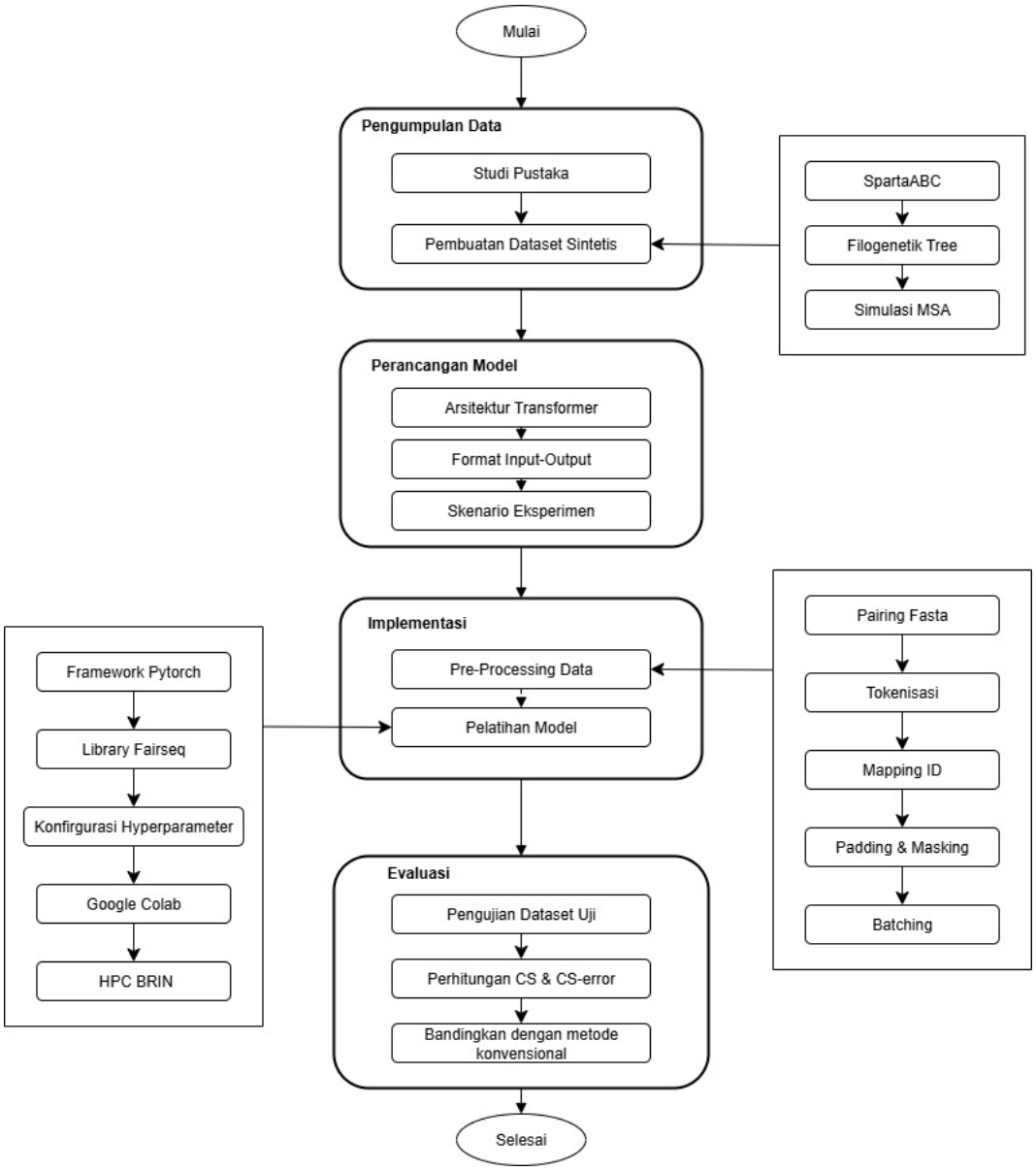
3.4.3 Tahap Implementasi dan Pelatihan

Tahap implementasi adalah proses teknis untuk mewujudkan model yang telah dirancang. Kegiatan pada tahap ini meliputi pengembangan kode untuk *pre-processing* data (*Pairing Fasta, Tokenisasi, Mapping, Padding & Masking, Batching*), membangun arsitektur model menggunakan framework PyTorch, dan menulis skrip untuk menjalankan siklus pelatihan pada platform yang telah ditentukan.

3.4.4 Tahap Evaluasi

Tahap terakhir adalah evaluasi untuk mengukur kinerja model yang telah dilatih. Model akan diuji menggunakan dataset uji yang terpisah, di mana kualitas *alignment* yang dihasilkan akan diukur menggunakan metrik *Column Score* (CS) dan hasilnya dibandingkan dengan *baseline* metode konvensional.

3.5 *Flowchart* Penelitian



Gambar 3. 1 *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

Ashkenazy, H., Levy Karin, E., Mertens, Z., Cartwright, R. A., & Pupko, T. (2017). Spartaabc: A Web Server To Simulate Sequences With Indel Parameters Inferred Using An Approximate Bayesian Computation Algorithm. *Nucleic Acids Research*, *45*(W1), W453–W457. Https://Doi.Org/10.1093/Nar/Gkx322

Bahdanau, D., Cho, K., & Bengio, Y. (2016). *Neural Machine Translation By Jointly Learning To Align And Translate*.

Borhani, Y., Khoramdel, J., & Najafi, E. (2022). A Deep Learning Based Approach For Automated Plant Disease Classification Using Vision Transformer. *Scientific Reports*, *12*(1), 11554. Https://Doi.Org/10.1038/S41598-022-15163-0

Chandra, A., Tünnermann, L., Löfstedt, T., & Gratz, R. (2023). Transformer-Based Deep Learning For Predicting Protein Properties In The Life Sciences. *Elife*, *12*. Https://Doi.Org/10.7554/Elife.82819

Chao, J., Tang, F., & Xu, L. (2022). Developments In Algorithms For Sequence Alignment: A Review. *Biomolecules*, *12*(4), 546. Https://Doi.Org/10.3390/Biom12040546

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

Ferruz, N., & Höcker, B. (2022). *Controllable Protein Design With Language Models*. Https://Doi.Org/10.1038/S42256-022-00499-Z

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

Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., & Pedrycz, W. (2023). *A Comprehensive Survey On Applications Of Transformers For Deep Learning Tasks*.

Ju, F., Zhu, J., Shao, B., Kong, L., Liu, T.-Y., Zheng, W.-M., & Bu, D. (2021). Copulanet: Learning Residue Co-Evolution Directly From Multiple Sequence Alignment For Protein Structure Prediction. *Nature Communications*, *12*(1), 2535. Https://Doi.Org/10.1038/S41467-021-22869-8

Kim, G. B., Kim, J. Y., Lee, J. A., Norsigian, C. J., Palsson, B. O., & Lee, S. Y. (2023). Functional Annotation Of Enzyme-Encoding Genes Using Deep Learning With Transformer Layers. *Nature Communications*, *14*(1), 7370. Https://Doi.Org/10.1038/S41467-023-43216-Z

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

Singh, S., & Mahmood, A. (2021). The NLP Cookbook: Modern Recipes For Transformer Based Deep Learning Architectures. *IEEE Access*, *9*, 68675–68702. Https://Doi.Org/10.1109/ACCESS.2021.3077350

Smirnov, V., & Warnow, T. (2021). MAGUS: Multiple Sequence Alignment Using Graph Clustering. *Bioinformatics*, *37*(12), 1666–1672. Https://Doi.Org/10.1093/Bioinformatics/Btaa992

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). *Sequence To Sequence Learning With Neural Networks*.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need*.

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

Willemink, M. J., Roth, H. R., & Sandfort, V. (2022). Toward Foundational Deep Learning Models For Medical Imaging In The New Era Of Transformer Networks. *Radiology: Artificial Intelligence*, *4*(6). Https://Doi.Org/10.1148/Ryai.210284

Xie, R., Zan, X., Chu, L., Su, Y., Xu, P., & Liu, W. (2023). Study Of The Error Correction Capability Of Multiple Sequence Alignment Algorithm (MAFFT) In DNA Storage. *BMC Bioinformatics*, *24*(1), 111. Https://Doi.Org/10.1186/S12859-023-05237-9

Zeyer, A., Bahar, P., Irie, K., Schluter, R., & Ney, H. (2019). A Comparison Of Transformer And LSTM Encoder Decoder Models For ASR. *2019 IEEE Automatic Speech Recognition And Understanding Workshop (ASRU)*, 8–15. Https://Doi.Org/10.1109/ASRU46091.2019.9004025