

BAB I

PENDAHULUAN

1.1. Latar Belakang

Perkembangan ilmu bioinformatika dalam dua dekade terakhir telah merevolusi cara ilmuwan memahami kehidupan pada tingkat molekuler, khususnya dalam kajian terhadap sekuens DNA, RNA, dan protein. Kemajuan teknologi sekuensing dan kemampuan komputasi modern telah memungkinkan pengolahan data biologis dalam skala besar, sehingga analisis kompleks yang sebelumnya memerlukan waktu bertahun-tahun kini dapat diselesaikan dalam hitungan jam atau bahkan menit. Salah satu teknik kunci dalam bidang ini adalah *Multiple sequence alignment* (MSA), yang berperan penting dalam berbagai aplikasi, termasuk analisis filogenetik untuk menelusuri hubungan evolusi antar organisme, prediksi struktur protein untuk memahami fungsi molekul biologis, serta identifikasi domain konservatif yang menjadi indikator fungsi biologis kritis dan mekanisme evolusi. Dengan demikian, MSA tidak hanya menjadi alat analisis, tetapi juga fondasi bagi berbagai penelitian bioinformatika yang mendalam dan aplikatif (Katoh & Standley, 2024).

Multiple sequence alignment (MSA) memungkinkan penyelarasan beberapa sekuens biologis secara bersamaan, sehingga pola evolusioner, hubungan filogenetik, dan motif fungsional dapat dikenali dengan lebih akurat dibandingkan analisis sekuens tunggal. Dengan menyelaraskan sejumlah sekuens secara paralel, MSA membantu mengidentifikasi residu atau domain konservatif yang penting secara biologis, memetakan variabilitas sekuens antar spesies, dan menyoroti region yang mengalami seleksi evolusioner. Teknik ini juga menjadi dasar bagi berbagai aplikasi bioinformatika, seperti prediksi struktur tiga dimensi protein, anotasi fungsi gen, serta analisis mutasi dalam studi penyakit genetik, karena pola konservasi yang terdeteksi melalui *alignment* memberikan indikasi lokasi yang kritis bagi fungsi atau stabilitas molekul. Dengan demikian, MSA bukan hanya alat untuk menyusun sekuens secara sistematis, tetapi juga sarana penting untuk memahami mekanisme biologis dan evolusi molekuler secara mendalam.

Seiring dengan melimpahnya data sekuens protein yang tersimpan dalam basis data global seperti UniProtKB/Swiss-Prot dan Pfam, kebutuhan akan algoritma penyelarasan yang efisien, akurat, dan mampu menangani volume data besar menjadi sangat krusial. Sebagai contoh, UniProt mencatat lebih dari 250 juta entri protein pada tahun 2024, dan jumlah ini terus bertambah seiring dengan kemajuan teknologi sekuensing dan proyek genomik global. Di sisi lain, database Pfam mencatat lebih dari 20.000 keluarga protein, yang mencakup domain konservatif yang hanya dapat diidentifikasi melalui *alignment* yang presisi. Volume data yang sangat besar dan keragaman sekuens yang tinggi ini menuntut algoritma *Multiple sequence alignment* (MSA) yang tidak hanya tepat secara biologis, tetapi juga efisien secara komputasi, sehingga memungkinkan analisis skala besar, pengenalan pola evolusioner, dan deteksi domain fungsional kritis secara cepat dan andal. Dengan kata lain, pertumbuhan eksponensial data protein menegaskan pentingnya pengembangan dan evaluasi algoritma MSA yang handal dan fleksibel untuk mendukung penelitian bioinformatika modern (Mistry, 2024).

Di antara berbagai tool MSA yang tersedia saat ini, MAFFT dan Clustal Omega menonjol sebagai dua metode yang paling sering digunakan dalam studi bioinformatika modern. MAFFT, yang diperkenalkan oleh Katoh dan Standley, memanfaatkan pendekatan *Fast Fourier Transform* (FFT) untuk mempercepat proses *alignment*, sambil tetap mempertahankan tingkat akurasi yang tinggi. Pendekatan FFT memungkinkan MAFFT untuk secara efisien menangani sejumlah besar sekuens dengan panjang bervariasi, sehingga algoritma ini sangat berguna dalam analisis dataset besar atau kompleks. Selain itu, MAFFT menawarkan berbagai mode *alignment* dan opsi optimisasi, termasuk *iterative refinement* dan strategi *progressive alignment*, yang membuatnya fleksibel untuk penelitian eksploratif maupun studi mendalam yang menuntut presisi biologis tinggi. Keunggulan-keunggulan ini menjadikan MAFFT sebagai salah satu algoritma pilihan dalam analisis bioinformatika yang membutuhkan kombinasi efisiensi dan akurasi (Katoh & Standley, 2024). Versi terbaru MAFFT (v7) bahkan mengklaim mengalami peningkatan performa yang signifikan dalam menangani dataset protein berskala besar, baik dari sisi kecepatan eksekusi maupun ketepatan *alignment*.

Perbaikan ini mencakup optimisasi algoritma FFT, peningkatan efisiensi memori, serta penyempurnaan strategi *iterative refinement* yang memungkinkan MAFFT menyelaraskan ribuan hingga puluhan ribu sekuens tanpa mengorbankan akurasi biologis. Dengan kemampuan ini, MAFFT v7 sangat cocok digunakan dalam analisis dataset kompleks, studi evolusi protein, dan aplikasi bioinformatika yang memerlukan penyelarasan presisi tinggi pada skala besar. Selain itu, fleksibilitas pengaturan parameter pada versi terbaru ini memungkinkan peneliti menyesuaikan algoritma sesuai dengan karakteristik dataset, seperti variasi panjang sekuens atau tingkat konservasi, sehingga MAFFT tetap menjadi salah satu tool MSA yang handal dan banyak digunakan dalam penelitian modern (Katoh & Rozewicki, 2023).

Sementara itu, Clustal Omega, yang dikembangkan oleh Sievers dan Higgins, memanfaatkan model *Hidden Markov Models* (HMM) dalam proses penyelarasan sekuens, sehingga algoritma ini lebih tangguh dalam menangani sekuens yang sangat divergen atau memiliki variasi panjang yang signifikan. Pendekatan berbasis HMM memungkinkan Clustal Omega membangun profil probabilitas dari sekelompok sekuens, yang kemudian digunakan untuk menghasilkan *alignment* yang konsisten dan biologis relevan, bahkan ketika sekuens memiliki kesamaan rendah. Keunggulan ini membuat Clustal Omega sangat efektif untuk dataset dengan homogenitas rendah atau studi evolusi yang melibatkan organisme berbeda secara jauh. Selain itu, algoritma ini dirancang untuk skalabilitas tinggi, sehingga dapat menyelaraskan ribuan sekuens dengan cepat, menjadikannya pilihan populer untuk analisis skala besar, integrasi *pipeline* otomatis, dan aplikasi berbasis web server. Dengan kombinasi kecepatan, stabilitas, dan kemampuan menangani divergensi tinggi, Clustal Omega melengkapi MAFFT sebagai salah satu tool MSA yang paling handal dalam penelitian bioinformatika modern (Sievers & Higgins, 2023).

Clustal Omega secara luas diakui dalam literatur sebagai pilihan utama untuk *large-scale alignment*, terutama dalam konteks studi evolusi protein yang melibatkan spesies yang sangat berjauhan. Algoritma ini dirancang untuk menyelaraskan ribuan hingga puluhan ribu sekuens secara efisien, sambil tetap

mempertahankan akurasi biologis tinggi, sehingga sangat cocok untuk analisis filogenetik lintas spesies dan identifikasi domain konservatif yang tersebar di berbagai organisme. Pendekatan berbasis *Hidden Markov Models* (HMM) memungkinkan Clustal Omega menangani divergensi tinggi antar sekuens, sehingga menghasilkan *alignment* yang stabil dan konsisten meskipun sekuens memiliki kesamaan rendah. Keunggulan ini menjadikan Clustal Omega tidak hanya efisien secara komputasi, tetapi juga andal dalam memberikan hasil yang biologis relevan, sehingga algoritma ini banyak digunakan dalam penelitian evolusi protein, anotasi fungsi gen, dan proyek bioinformatika berskala besar di berbagai laboratorium dan basis data publik (Seemann, 2024).

Namun demikian, hasil studi perbandingan terkini menunjukkan bahwa kedua algoritma, MAFFT dan Clustal Omega, memiliki keunggulan dan kelemahan relatif, yang sangat tergantung pada karakteristik dataset yang digunakan. Misalnya, MAFFT cenderung lebih unggul pada dataset yang kompleks dengan panjang sekuens yang bervariasi atau dataset eksploratif yang memerlukan penyesuaian parameter dan mode *iterative refinement*, sehingga memungkinkan *alignment* yang lebih presisi. Di sisi lain, Clustal Omega menunjukkan keunggulan pada dataset skala besar dan homogen, serta dalam analisis evolusi antar spesies yang sangat divergen, berkat arsitektur berbasis *Hidden Markov Models* (HMM) yang stabil dan efisien secara komputasi. Temuan ini menegaskan bahwa pemilihan algoritma MSA tidak bersifat universal, melainkan harus disesuaikan dengan tujuan analisis, skala dataset, dan prioritas antara efisiensi komputasi dan akurasi biologis, sehingga peneliti dapat memaksimalkan kinerja *alignment* sesuai kebutuhan aplikasi bioinformatika yang spesifik.

Penelitian (Sharma *et al.*, 2024) menemukan bahwa MAFFT lebih unggul dalam hal kecepatan ketika diaplikasikan pada dataset protein yang besar, sementara Clustal Omega menunjukkan akurasi lebih baik dalam mendeteksi domain konservatif pada keluarga protein yang sangat bervariasi. Temuan ini menyoroti bahwa kedua algoritma memiliki keunggulan yang bersifat kontekstual, tergantung pada karakteristik dataset dan tujuan analisis. Kondisi ini juga membuka *research gap* penting, karena hingga saat ini belum ada studi komprehensif yang

secara sistematis membandingkan performa kedua algoritma menggunakan dataset protein terbaru yang benar-benar representatif dari kondisi nyata di lapangan. Kekosongan penelitian ini memberikan dasar kuat bagi penelitian lanjutan untuk mengevaluasi akurasi, efisiensi komputasi, dan kehandalan biologis MAFFT dan Clustal Omega secara menyeluruh, sehingga dapat memberikan panduan yang lebih jelas bagi pengguna dalam memilih algoritma MSA sesuai kebutuhan bioinformatika modern.

Selain faktor akurasi dan kecepatan, dimensi lain yang kini menjadi sorotan dalam studi MSA adalah relevansi biologis dari *alignment* yang dihasilkan. Seperti ditegaskan oleh (Steinegger & Soding, 2023), kualitas *alignment* protein seharusnya tidak hanya dinilai dari skor kesamaan sekuens, tetapi juga dari sejauh mana *alignment* tersebut mencerminkan struktur tiga dimensi (3D) protein yang sebenarnya serta kesesuaian dengan domain fungsional yang diketahui. Dengan semakin meningkatnya jumlah data struktur protein hasil pemodelan berbasis AI, seperti AlphaFold, menjadi sangat penting untuk mengevaluasi apakah algoritma MSA tradisional, termasuk MAFFT dan Clustal Omega, masih mampu menyelaraskan motif-motif penting secara biologis dengan akurat. Pendekatan ini tidak hanya menekankan keandalan algoritma dari sisi statistik, tetapi juga menekankan validitas biologis dari hasil *alignment*, sehingga dapat mendukung aplikasi lanjutan seperti homology modeling, prediksi fungsi protein, dan studi evolusi molekuler secara lebih komprehensif.

Sejalan dengan itu, tantangan dalam *Multiple sequence alignment* (MSA) untuk protein tidak hanya terbatas pada aspek teknis seperti kecepatan eksekusi dan efisiensi algoritma, tetapi juga mencakup tantangan biologis yang kompleks. Protein sebagai molekul biologis memiliki struktur tiga dimensi (3D) yang rumit, di mana hubungan evolusioner sering kali tersembunyi dalam pola konservasi yang tersebar di seluruh sekuens. Penyelarasan yang akurat harus mampu mengidentifikasi residu kritis yang berperan dalam fungsi atau stabilitas protein, meskipun residu tersebut mungkin berada jauh dalam urutan linear namun berdekatan secara struktural. Hal ini menuntut algoritma MSA untuk tidak hanya menyelaraskan sekuens berdasarkan kesamaan linear, tetapi juga

mempertimbangkan konservasi fungsional dan struktural, sehingga hasil *alignment* dapat digunakan secara andal untuk prediksi struktur protein, analisis motif fungsional, dan studi evolusi molekuler. Dengan meningkatnya jumlah data protein hasil pemodelan AI, tantangan ini semakin relevan untuk memastikan bahwa metode MSA tetap menghasilkan *alignment* biologis yang valid dan aplikatif (Steinegger & Soding, 2023). Hal ini menjadikan proses *alignment* bukan sekadar penyusunan urutan sekuens, melainkan juga sebuah upaya untuk merekonstruksi sejarah evolusi molekuler secara mendalam.

Setiap posisi residu dalam sekuens dapat menyimpan informasi tentang konservasi evolusioner, adaptasi fungsional, dan hubungan filogenetik antar spesies. Dengan demikian, penyelarasan yang akurat tidak hanya menekankan kesamaan linear antar sekuens, tetapi juga mampu menyingkap pola-pola evolusi tersembunyi yang mencerminkan perubahan fungsi, tekanan seleksi, dan inovasi struktural protein sepanjang waktu. Perspektif ini menekankan bahwa *Multiple sequence alignment* (MSA) merupakan alat penting dalam biologi komputasional, yang tidak hanya menghasilkan data statistik, tetapi juga memberikan wawasan biologis yang dapat diterapkan untuk homology modeling, prediksi struktur protein, analisis domain fungsional, dan studi evolusi molekuler.

Ketergantungan penelitian biologi modern terhadap hasil *alignment* sangat besar, karena kualitas dan validitas analisis lanjutan sering kali bergantung pada presisi MSA. Misalnya, dalam penelitian struktur protein berbasis machine learning seperti AlphaFold, akurasi prediksi struktur tiga dimensi protein sangat dipengaruhi oleh kualitas *alignment* input. *Alignment* yang tepat memungkinkan model untuk mengenali konservasi residu, motif fungsional, dan domain struktural, sehingga prediksi yang dihasilkan lebih akurat dan biologis relevan. Sebaliknya, *alignment* yang kurang akurat dapat menimbulkan kesalahan prediksi struktur, mengaburkan hubungan evolusi, dan menurunkan keandalan analisis fungsional. Oleh karena itu, pemilihan algoritma MSA yang handal, seperti MAFFT atau Clustal Omega, menjadi langkah krusial dalam *pipeline* penelitian modern, memastikan bahwa data input untuk pemodelan, anotasi gen, atau analisis evolusi memiliki kualitas biologis

yang tinggi dan dapat diandalkan untuk pengambilan keputusan ilmiah (Mistry, 2024).

Jika *alignment* yang digunakan mengandung kesalahan, maka prediksi struktur protein pun berpotensi bias atau keliru, sehingga dapat berdampak signifikan pada penelitian lanjutan, seperti desain vaksin, pengembangan obat, dan rekayasa protein. Kesalahan sekecil apa pun dalam penyelarasan residu dapat mengubah interpretasi domain fungsional, motif konservatif, atau interaksi protein, yang pada akhirnya memengaruhi keputusan ilmiah dan aplikasi terapeutik. Oleh karena itu, evaluasi menyeluruh terhadap algoritma MSA menjadi kebutuhan yang mendesak, terutama di tengah tren eksplorasi besar-besaran dataset protein dari berbagai spesies, yang terus bertambah baik dari basis data publik maupun hasil pemodelan berbasis AI. Penilaian ini tidak hanya mencakup akurasi *alignment*, tetapi juga efisiensi komputasi, relevansi biologis, dan kemampuan algoritma untuk menangani dataset heterogen, sehingga pemilihan metode yang tepat dapat menjamin validitas dan keandalan hasil penelitian bioinformatika modern.

Di sisi lain, semakin beragamnya jenis protein yang dianalisis oleh para peneliti menuntut algoritma MSA yang lebih adaptif dan fleksibel. Dataset protein modern tidak lagi terbatas pada organisme model seperti manusia atau tikus, melainkan telah meluas hingga mencakup protein dari virus, bakteri, hingga organisme laut dalam yang hidup di kondisi ekstrem. Keanekaragaman ini menghadirkan tantangan baru, karena sekuens protein dari spesies yang sangat berbeda sering memiliki kesamaan rendah, panjang sekuens yang bervariasi, dan motif fungsional yang tersembunyi. Oleh karena itu, algoritma MSA harus mampu menangani divergensi tinggi, mempertahankan kesetiaan biologis, dan tetap efisien dalam hal penggunaan memori dan waktu eksekusi. Kemampuan adaptif ini menjadi sangat penting untuk mendukung analisis evolusi molekuler, identifikasi domain konservatif, dan prediksi fungsi protein pada berbagai jenis organisme, sehingga hasil *alignment* tetap relevan dan dapat diandalkan untuk penelitian lanjutan (Seemann, 2024).

Keragaman ini memunculkan tantangan baru bagi algoritma MSA seperti MAFFT dan Clustal Omega, karena keduanya harus mampu menjaga akurasi

alignment tanpa mengorbankan efisiensi waktu maupun penggunaan daya komputasi. Protein dari organisme yang sangat berbeda seringkali memiliki sekuens yang sangat divergen, panjang sekuens yang bervariasi, dan motif fungsional yang tersembunyi, sehingga algoritma harus menyeimbangkan antara ketelitian biologis dan kecepatan eksekusi. MAFFT, dengan pendekatan *Fast Fourier Transform* (FFT) dan berbagai mode *iterative refinement*, menawarkan fleksibilitas dalam menangani kompleksitas dataset, sedangkan Clustal Omega, dengan *Hidden Markov Models* (HMM), unggul dalam menyelaraskan sekuens yang sangat berbeda secara evolusioner. Tantangan ini menegaskan bahwa pemilihan algoritma harus mempertimbangkan karakteristik dataset, tujuan penelitian, dan skala analisis, sehingga hasil *alignment* tetap akurat, biologis relevan, dan efisien secara komputasi.

Penelitian (Sharma *et al.*, 2024) menegaskan bahwa metode *alignment* yang efektif pada satu kelompok protein belum tentu menunjukkan performa serupa pada kelompok protein lainnya. Hal ini menyoroti variabilitas kinerja algoritma MSA tergantung pada karakteristik spesifik dataset, seperti tingkat konservasi, panjang sekuens, dan keragaman evolusioner. Temuan ini menekankan pentingnya studi komparatif lintas dataset yang lebih luas, di mana evaluasi algoritma dilakukan pada berbagai jenis protein dan kondisi biologis yang berbeda. Pendekatan ini memungkinkan peneliti untuk memahami kekuatan dan kelemahan relatif setiap algoritma, sekaligus memberikan panduan yang lebih tepat dalam memilih metode *alignment* yang sesuai dengan tujuan penelitian, ukuran dataset, dan kebutuhan efisiensi komputasi. Dengan demikian, penelitian lintas dataset menjadi langkah krusial untuk meningkatkan relevansi biologis dan keandalan hasil MSA dalam aplikasi bioinformatika modern.

Mengacu pada data terbaru dari UniProt pada tahun 2024, tercatat adanya lonjakan signifikan dalam jumlah entri protein yang dihasilkan dari berbagai proyek metagenomik global. Peningkatan jumlah data protein ini menunjukkan bahwa kompleksitas dan keragaman informasi biologis yang harus dianalisis semakin besar, sehingga menuntut ketersediaan metode analisis bioinformatika yang lebih efisien dan adaptif. Salah satu kebutuhan utama dalam konteks ini adalah

pengembangan dan pemanfaatan algoritma *Multiple sequence alignment* (MSA) yang mampu bekerja pada skala besar dengan tetap menjaga akurasi hasil. Seiring dengan itu, penelitian yang membandingkan performa dua algoritma populer, yaitu MAFFT dan Clustal Omega, terhadap dataset protein terkini menjadi semakin relevan. Analisis komparatif tersebut tidak hanya penting dalam ranah akademik sebagai kontribusi terhadap pengembangan ilmu bioinformatika, tetapi juga memiliki dampak praktis yang signifikan bagi berbagai sektor. Misalnya, di bidang biofarmasi, kemampuan menyelaraskan protein secara tepat dapat mempercepat proses penemuan obat baru berbasis protein; dalam pertanian molekuler, teknologi ini dapat dimanfaatkan untuk meningkatkan ketahanan tanaman melalui identifikasi protein fungsional tertentu; sedangkan dalam teknologi diagnostik modern, hasil MSA yang akurat dapat mendukung pengembangan biomarker baru untuk deteksi penyakit. Dengan demikian, penelitian terkait efektivitas dan efisiensi MAFFT serta Clustal Omega bukan hanya bersifat teknis, tetapi juga strategis dalam mendukung kemajuan ilmu pengetahuan dan penerapannya dalam berbagai industri vital.

Dengan latar belakang tersebut, penelitian ini secara khusus diarahkan untuk mengevaluasi dan menganalisis perbandingan kinerja algoritma MAFFT dan Clustal Omega dalam konteks penyelarasan sekuens protein. Fokus analisis ditekankan pada tiga dimensi utama yang saling melengkapi, yaitu akurasi *alignment* terhadap domain konservatif, efisiensi komputasi dalam memproses data yang semakin kompleks dan berskala besar, serta relevansi biologis dari hasil *alignment* terhadap struktur dan fungsi protein yang telah diketahui melalui basis data eksperimental maupun prediksi struktur modern. Ketiga aspek ini dipilih karena memiliki peran penting dalam memastikan bahwa hasil penyelarasan tidak hanya bernilai secara teknis, tetapi juga bermakna secara biologis dan aplikatif. Diharapkan temuan dari penelitian ini mampu memberikan rekomendasi praktis bagi komunitas bioinformatika, baik peneliti akademik maupun praktisi industri, dalam menentukan algoritma yang paling sesuai dengan kebutuhan spesifik mereka.

Penelitian ini juga berpotensi membuka wawasan baru bagi pengembangan algoritma MSA generasi berikutnya yang lebih adaptif, presisi, dan mampu menjawab tantangan analisis data protein yang terus berkembang pesat seiring dengan akselerasi proyek-proyek metagenomik global. Oleh karena itu peneliti akan mengangkat judul tentang “**Analisa Komperatif Algoritma untuk Masalah *Multiple sequence alignment* (MSA) dalam Bidang Bioinformatika**”.

1.2. Rumusan Permasalahan

Berdasarkan uraian dari latar belakang, maka rumusan masalah pada penelitian ini adalah:

- a. Bagaimana perbandingan performa algoritma MAFFT dan Clustal Omega dalam melakukan *multiple sequence alignment* (MSA) pada dataset protein?
- b. Apakah terdapat perbedaan signifikan dalam hal akurasi *alignment* melalui SP dan TC *Score* yang dihasilkan oleh MAFFT dan Clustal Omega berdasarkan standar *benchmark* dataset?
- c. Bagaimana efisiensi komputasi kedua algoritma (dalam hal waktu proses dan penggunaan sumber daya) saat diaplikasikan pada dataset protein berskala besar?
- d. Sejauh mana hasil *alignment* dari MAFFT dan Clustal Omega relevan secara biologis, khususnya dalam kesesuaiannya dengan domain fungsional protein yang terdaftar dalam database NCBI?
- e. Algoritma mana yang lebih optimal untuk digunakan dalam aplikasi bioinformatika modern, seperti analisis filogenetik, berdasarkan karakteristik dataset protein yang digunakan?

1.3. Batasan Masalah

Batasan masalah pada penelitian ini adalah:

- a. Algoritma yang dibandingkan dalam penelitian ini hanya mencakup MAFFT dan Clustal Omega.
- b. Dataset yang digunakan dalam penelitian ini terbatas pada sekuens protein.

- c. Evaluasi akurasi *alignment* dengan membandingkan antara MAFFT dan Clustal Omega dengan *ground truth* dibatasi pada pengukuran terhadap *benchmark* standar, yaitu skor *Sum-of-Pairs (SP)* dan *Total Column (TC)*, serta kesesuaian *alignment* dengan domain fungsional yang terdaftar dalam database NCBI.
- d. Pengujian efisiensi komputasi algoritma dibatasi pada parameter waktu proses (*runtime*) dan penggunaan memori pada platform komputer standar.
- e. Aplikasi hasil *alignment* dalam penelitian ini dibatasi untuk tujuan evaluasi metode MSA saja, dan tidak meluas pada aplikasi lanjutan seperti prediksi struktur protein 3D atau desain vaksin.

1.4. Tujuan Penelitian

Tujuan penelitian ini adalah:

- a. Menganalisis dan membandingkan performa algoritma MAFFT dan Clustal Omega dalam melakukan *multiple sequence alignment (MSA)* pada dataset protein.
- b. Mengukur tingkat akurasi *alignment* dengan membandingkan terhadap *ground truth* yang dihasilkan oleh MAFFT dan Clustal Omega dengan menggunakan *benchmark* dataset standar.
- c. Mengevaluasi efisiensi komputasi kedua algoritma, khususnya dalam aspek waktu proses dan penggunaan sumber daya saat menangani dataset protein berukuran besar.
- d. Menilai relevansi biologis hasil *alignment* dari kedua algoritma dengan membandingkannya terhadap domain fungsional protein yang terdaftar dalam database NCBI.
- e. Memberikan rekomendasi praktis bagi peneliti bioinformatika mengenai algoritma MSA yang lebih optimal digunakan, baik untuk keperluan prediksi struktur protein, analisis filogenetik, maupun aplikasi bioinformatika modern lainnya.

1.5. Manfaat Penelitian

Manfaat yang diperoleh dari penelitian ini adalah:

a. Manfaat Akademis

1. Menambah khasanah ilmu pengetahuan dalam bidang bioinformatika, khususnya terkait analisa algoritma *multiple sequence alignment* (MSA).
2. Menyediakan data empiris terbaru mengenai performa algoritma MAFFT dan Clustal Omega pada dataset protein yang lebih luas.
3. Mendorong pengembangan standar baru dalam evaluasi algoritma MSA yang menggabungkan aspek akurasi matematis dengan relevansi biologis.

b. Manfaat Praktis

1. Memberikan panduan praktis bagi peneliti bioinformatika dalam memilih algoritma MSA yang paling sesuai dengan karakteristik dataset protein yang mereka gunakan.
2. Mendukung aplikasi-aplikasi bioinformatika modern seperti prediksi struktur protein, analisis filogenetik, hingga desain vaksin.
3. Membantu laboratorium riset dan industri bioteknologi dalam menangani dataset protein berskala besar dengan lebih efisien, tanpa mengorbankan kualitas *alignment*.

c. Manfaat bagi Pengembangan Teknologi

Mendorong inovasi lebih lanjut dalam pengembangan algoritma MSA yang lebih cepat, akurat, dan relevan secara biologis.

1.6. Ruang Lingkup Penelitian

Ruang lingkup penelitian ini secara spesifik difokuskan pada kajian komparatif dua algoritma *Multiple sequence alignment* (MSA), yaitu MAFFT dan Clustal Omega, dalam konteks penyelarasan sekuens protein. Batasan ruang lingkup ini mencakup beberapa aspek berikut:

- a. Objek Penelitian
 1. Penelitian hanya difokuskan pada dataset sekuens protein yang diambil dari database global seperti UniProtKB/Swiss-Prot dan Pfam.
 2. Tidak melibatkan jenis sekuens biologis lain seperti DNA atau RNA.
- b. Aspek Algoritma
 1. Algoritma yang dibandingkan terbatas pada MAFFT dan Clustal Omega, sebagai dua algoritma MSA yang representatif dan banyak digunakan di komunitas bioinformatika.
 2. Perbandingan performa algoritma mencakup kecepatan (*runtime*), penggunaan memori, dan efisiensi komputasi secara umum pada platform komputer standar.
- c. Dimensi Evaluasi
 1. Evaluasi hasil *alignment* dilakukan dengan standar benchmark yang diakui, yaitu menggunakan metrik *Sum-of-Pairs* (SP) *Score* dan *Total Column* (TC) *Score*.
 2. Penilaian relevansi biologis *alignment* dilakukan dengan membandingkan kesesuaian domain konservatif hasil *alignment* terhadap data domain fungsional yang tercatat di database NCBI.
 3. Penelitian tidak meluas pada validasi lanjutan seperti uji prediksi struktur 3D atau validasi eksperimental di laboratorium.
- d. Ruang Aplikasi
 1. Fokus analisis terletak pada kebutuhan bioinformatika dasar seperti analisis filogenetik, prediksi struktur, dan identifikasi domain konservatif.
 2. Penelitian tidak membahas aplikasi *downstream* secara detail, misalnya integrasi langsung dengan *pipeline* AI seperti AlphaFold, desain vaksin, atau rekayasa protein.
- e. Skala Dataset
 1. Dataset yang digunakan disusun agar mencerminkan kondisi nyata di lapangan, dengan variasi tingkat homologi sekuens rendah hingga tinggi.

2. Penelitian ini juga membahas bagaimana skala dan keragaman dataset dapat memengaruhi performa kedua algoritma.

f. Sumber Daya Komputasi

Pengujian dilakukan pada platform komputer standar, bukan pada superkomputer atau cluster komputasi awan, sehingga hasil dapat diaplikasikan secara praktis oleh laboratorium bioinformatika dengan fasilitas komputasi menengah.

Dengan ruang lingkup tersebut, penelitian ini diharapkan dapat menghadirkan gambaran yang komprehensif mengenai keunggulan sekaligus keterbatasan algoritma MAFFT dan Clustal Omega melalui pendekatan analisis yang sistematis dan terukur. Penelitian ini tidak hanya bertujuan untuk memetakan performa kedua algoritma dalam konteks penyelarasan sekuens protein skala besar, tetapi juga untuk mengidentifikasi aspek-aspek kritis yang dapat memengaruhi kualitas hasil, baik dari sisi akurasi biologis, efisiensi komputasi, maupun relevansinya terhadap kebutuhan praktis di berbagai bidang. Dengan demikian, hasil penelitian ini berpotensi menjadi rujukan penting bagi peneliti dan akademisi yang bergerak di bidang bioinformatika, sekaligus memberikan panduan aplikatif bagi industri yang mengandalkan teknologi *Multiple sequence alignment*, seperti biofarmasi, bioteknologi pertanian, dan pengembangan diagnostik modern. Pada akhirnya, kontribusi penelitian ini diharapkan tidak hanya terbatas pada tataran evaluatif, tetapi juga mampu memberikan dasar pemikiran baru untuk mendukung inovasi dalam pengembangan algoritma MSA generasi berikutnya yang lebih adaptif dan efisien.

1.7. Sistematika Penulisan

Berikut ini merupakan susunan dan struktur Tesis:

BAB I PENDAHULUAN

Pada bab I berisi latar belakang, rumusan permasalahan, batasan masalah, tujuan penelitian, manfaat penelitian, ruang lingkup penelitian, serta susunan dan struktur Tesis.

BAB II KAJIAN PUSTAKA

Pada bab II berisi kajian pustaka yang memuat penelitian-penelitian terdahulu dan landasan teori yang memuat teori-teori yang berkaitan dengan penelitian.

BAB III METODOLOGI PENELITIAN

Bab III berisi waktu dan tempat penelitian, prosedur penelitian, metode penelitian, analisis data, serta jadwal penelitian.

BAB IV HASIL PENELITIAN DAN PEMBAHASAN

Pada bab hasil penelitian dijabarkan hasil MSA dari metode MAFFT dan clustal omega meliputi deskripsi dataset yang digunakan, hasil evaluasi akurasi *alignment*, dan hasil evaluasi efisiensi komputasi. Pada pembahasan dilakukan analisis perbandingan dari MAFFT dan clustal omega dari masing-masing poin hasil penelitian.

BAB V PENUTUP

Bagian ini merupakan BAB terakhir tesis, yang memuat kesimpulan hasil penelitian secara sistematis yang berkaitan dengan upaya menjawab hipotesa dan/atau tujuan penelitian. Pada akhir bab ini dikemukakan saran-saran yang berkaitan dengan penelitian yang telah dilakukan maupun model/prototipe yang dihasilkan. Saran-saran tersebut dapat berupa bentuk kebijakan dan upaya praktis pemecahan masalah yang dihadapi dan bahan atau aspek yang dapat diteliti lebih lanjut. Saran harus dibuat se-operasional mungkin sehingga bermanfaat bagi mereka yang menerima saran tersebut. Isi bab ini adalah: a) kesimpulan, dan b) saran

DAFTAR PUSTAKA

Daftar pustaka berisi referensi-referensi yang digunakan sebagai landasan penelitian.

DAFTAR LAMPIRAN

Bagian ini diawali halaman kosong yang ditulis dengan kata lampiran di tengah bidang pengetikan. Halaman ini tidak diberi nomor, tetapi ikut dihitung. Dalam lampiran disajikan keterangan-keterangan yang dianggap penting untuk tesis, tetapi yang akan mengganggu kelancaran membaca bila dicantumkan di bagian utama tesis.