FUTURE SCIENCE

PENGANTAR BIOINFORMATIKA

Editor:
Didik Huswo Utomo, S.Si, M.Si., Ph.D

Yusril Ihza Farhan Wijaya, M.Sc. | Sonia Meta Angraini, S.Si., M.Biotech | Armansyah Maulana Harahap, S.Si, M.Biomed | Kiky Mey Putranty S.Si., M.Si | Rini Hafzari, S.Si., M.Si | M. Khoiron Ferdiansyah, STP, MSc., PhD | Rolina Anna Erica Sihombing, S.Si., M.Si. | dr. Lustyafa Inassani Alifia, M.Biomed. | apt. Lolita,M.Sc., Ph.D | Chindy Nur Rosmeita, M.Si. | Dr. Eko Kusumawati, S.Si, MP | apt. Rizki Rahmadi Pratama, M. Farm | Muhamad Firmanda, S.Si. | Pauline Elisabeth, S.Agr., M.Si.

ISBN 978-621-8438-20-0

ISBN:
978-621-8438-20-0 PDF (downloadable)

Published by:
FSH-PH Publications
Block 4 Lot 6, Lumina Homes,
Pamatawan, Subic, Zambales, Philippines

https://fsh-publication.com/

Pengantar Bioinformatika

Editor:
Didik Huswo Utomo, S.Si, M.Si., Ph.D

Penulis:
Yusril Ihza Farhan Wijaya, M.Sc.
Sonia Meta Angraini, S.Si., M.Biotech
Armansyah Maulana Harahap, S.Si, M.Biomed
Kiky Mey Putranty, S.Si., M.Si.
Rini Hafzari, S.Si., M.Si
M. Khoiron Ferdiansyah, STP, MSc, PhD
Rolina Anna Erica Sihombing, S.Si., M.Si.
dr. Lustyafa Inassani Alifia, M.Biomed
apt. Lolita, M.Sc., Ph.D
Chindy Nur Rosmeita, M.Si
Dr. Eko Kusumawati, S.Si, MP

apt. Rizki Rahmadi Pratama, M. Farm
Muhamad Firmanda, S.Si.
Pauline Elisabeth, S.Agr., M.Si

Technical Editor:
Naqiyah Afifah Mulachelah, S.Si., M.Si.

2026

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1
Dasar-Dasar Bioinformatika

Yusril Ihza Farhan Wijaya

Pengertian Bioinformatika

Bioinformatika didefinisikan sebagai implementasi alat serta analisis komputasi dengan tujuan untuk pengumpulan dan interpretasi data biologis. Bioinformatika merupakan rumpun ilmu interdisipliner yang memanfaatkan ilmu biologi, ilmu statistika, dan ilmu komputer. Pada era omics seperti sekarang, data biologis disajikan dalam berbagai bentuk yang merepresentasikan informasi pada berbagai tingkatan sistem biologi. Bahkan skala data biologis kini mampu melampaui *petabyte* (PB) hingga *exabyte* (EB).

Data biologis yang berskala besar membawa biologi memasuki era *big* data. *Big* data umumnya memiliki empat karakteristik, yang meliputi: 1) volume data yang sangat besar, 2) kecepatan pemrosesan yang tinggi, 3) keragaman sumber data, serta 4) keandalan dan kualitas data. Keempat ciri ini memerlukan teori dan teknologi khusus agar nantinya dapat dikelola secara optimal untuk berbagai kebutuhan.

Pertumbuhan eksponensial data biologis menghadirkan peluang besar sekaligus tantangan serius. Tantangan tersebut berupa bagaimana menangani kompleksitas informasi, mengintegrasikan data dari sumber yang sangat heterogen, serta

menentukan prinsip dan standar yang tepat dalam pengelolaan *big* data. Apabila tantangan ini dapat diatasi, data biologis akan berpotensi menghasilkan nilai ilmiah yang sangat besar.

Kondisi yang telah dipaparkan sebelumnya menciptakan urgensi untuk menjembatani kesenjangan antara perkembangan *high-throughput technology* dan kemampuan manusia dalam mengelola, menganalisis, serta mengintegrasikan data biologis berskala besar. Alat dan teknik komputasi merevolusi bidang bioinformatika dengan memungkinkan pengorganisasian, analisis, dan pemahaman sistematis terhadap data biologis yang kompleks. Oleh karena itu, Bioinformatika kini menjadi instrumen yang tak tergantikan dalam penelitian biologi modern.

Sejarah Perkembangan Bioinformatika

Ilmu bioinformatika yang terus berkembang menyelesaikan permintaan yang tinggi terhadap analisis dan interpretasi data biologis. Namun, sering dijumpai kesalahpahaman bahwa bioinformatika merupakan disiplin ilmu yang baru muncul seiring dengan lahirnya teknologi *Next-Generation Sequencing* (NGS). Fondasi bioinformatika telah ada bahkan sebelum sekuens DNA dapat dibaca. Perkembangan bioinformatika dimulai pada pertengahan abad ke-20.

Periode awal bioinformatika dimulai pada tahun 1950 hingga 1970, yang pada masa ini bioinformatika berfokus pada pendekatan analisis protein. Hal ini dikarenakan pada masa tersebut metode untuk menentukan sekuens DNA belum tersedia, sementara sekuens protein sudah mulai dapat ditentukan. Kemudian, dilakukan analisis protein awal yang ditandai dengan dilaksanakannya penentuan urutan insulin sapi oleh Frederick Sanger pada awal 1950-an, serta pengembangan

metode degradasi Edman. Peneliti pada era tersebut mulai mendapatkan akses ke data sekuens asam amino. Margaret Dayhoff dan Robert Ledley sebagai pionir penciptaan COMPROTEIN pada awal 1960-an, yang merupakan perangkat lunak pertama yang dirancang untuk menentukan urutan protein primer dari data peptida yang tumpang tindih (*de novo sequence assembly*). Lahirnya *database* biologis pertama terjadi pada tahun 1965, saat Dayhoff mempublikasikan "*Atlas of Protein Sequence and Structure*". Dalam rangka memfasilitasi penyimpanan data, ia menciptakan kode satu huruf untuk asam amino yang menjadi basis data sekuens biologis pertama di dunia. Zuckerkandl dan Pauling mencetuskan istilah "*paleogenetika*" pada tahun 1963. Mereka menunjukkan bahwa sejarah evolusi organisme dapat dilacak dengan membandingkan urutan molekuler (seperti hemoglobin) dari berbagai spesies, sebuah konsep yang menjadi cikal bakal analisis filogenetik modern.

Pergeseran paradigma dari protein ke DNA terjadi pada periode 1970 hingga 1980. Dekade ini merupakan transisi krusial dengan fokus analisis mulai beralih ke DNA, yang didorong oleh penemuan teknologi sekuensing. Needleman dan Wunsch mempublikasikan algoritme pemrograman dinamis untuk menjajarkan dua sekuens secara optimal pada tahun 1970. Algoritme ini tetap menjadi landasan bagi banyak alat bioinformatika hingga saat ini. Pada tahun 1976 dan 1977, muncul metode sekuensing Maxam-Gilbert dan Sanger yang memungkinkan pembacaan DNA secara langsung. Bakteriofage fX174 merupakan genom DNA pertama yang berhasil disekuens secara utuh. Pada tahun 1979, Roger Staden mengembangkan *"Staden Package",* yang merupakan perangkat lunak pertama yang secara khusus dirancang untuk menganalisis, mengedit,

dan menggabungkan data sekuensing DNA (metode *shotgun*), yang menjadi standar selama bertahun-tahun.

Kemajuan yang pesat dalam bidang biologi dan ilmu komputer terjadi pada periode 1980 hingga 1990. Hal ini dilandasi atas kemajuan biologi molekuler dengan penemuan PCR yang berjalan beriringan dengan revolusi komputer pribadi (PC), sehingga mendorong bioinformatika lebih mudah diakses oleh peneliti. Kebutuhan akan penyimpanan data terpusat melahirkan tiga bank data utama, yaitu: GenBank di Amerika Serikat (1982), EMBL Data Library di Eropa (1982), dan DDBJ di Jepang (1987). Tahun 1984 dan 1985, perangkat lunak GCG dan algoritme FASTA ditemukan oleh Lipman dan Pearson yang memungkinkan para peneliti melakukan pencarian kemiripan sekuens yang cepat. Kemudian, pada tahun 1988, Clustal diperkenalkan untuk melakukan penjajaran multi-sekuens (*Multiple Sequence Alignment*). Tahun 1987 menandai rilisnya bahasa pemrograman Perl oleh Larry Wall. Kemampuannya yang unggul dalam memproses teks menjadikan "Perl" sebagai bahasa *lingua franca* dalam bioinformatika selama dekade berikutnya. Selain itu, Felsenstein (1981) juga mengembangkan metode *Maximum Likelihood* (ML) pertama yang memungkinkan inferensi pohon filogenetik berbasis sekuens DNA dengan pendekatan statistik yang kuat.

Periode 1990 hingga 2000 didominasi oleh proyek genom skala besar dan dampak revolusioner internet terhadap distribusi data ilmiah. *Human Genome Project* (HGP) diluncurkan pada tahun 1990, proyek ini memicu perlombaan antara konsorsium publik dan perusahaan swasta (*Celera Genomics*). Kompetisi ini mempercepat pengembangan algoritme perakitan genom (*assembler*) baru seperti PHRAP, TIGR Assembler, dan Celera

Assembler. Pada tahun 1995, genom bakteri *Haemophilus influenzae* menjadi genom organisme hidup pertama yang dipublikasikan. NCBI meluncurkan situs webnya pada tahun 1994, yang menyediakan akses publik ke berbagai alat, termasuk BLAST (*Basic Local Alignment Search Tool*) yang dirilis tahun 1990 untuk pencarian sekuens yang jauh lebih cepat daripada FASTA. Prediksi struktur protein 3D mulai dikompetisikan secara formal melalui ajang CASP (*Critical Assessment of Structure Prediction*) yang dimulai pada tahun 1994.

Terjadi ledakan data pada periode 2000 hingga 2010 yang disebabkan oleh teknologi baru yang mengubah biologi menjadi ilmu yang "kaya data" (*data-rich*). Komersialisasi teknologi sekuensing generasi kedua (seperti 454 Pyrosequencing pada 2005 dan Solexa/Illumina) meningkatkan kapasitas data secara drastis dari 96 pembacaan per *run* menjadi jutaan. Algoritme lama tidak lagi mampu menangani volume data NGS. Hal ini mendorong pengembangan *assembler* berbasis *graf de Bruijn* seperti Velvet dan Abyss untuk menangani *short-reads*. Bioinformatika tidak lagi dipandang sebagai layanan teknis semata, melainkan menjadi komponen integral dari eksperimen biologi. Istilah *cloud computing* juga mulai diperkenalkan untuk mengatasi kebutuhan penyimpanan data yang masif.

Tantangan bioinformatika kian bergeser pada periode 2010 hingga sekarang. Dari analisis yang sekadar menghasilkan data, menjadi bagaimana mengintegrasikan dan memvalidasi data tersebut. Tersedianya ribuan alat bioinformatika menyebabkan luaran yang berbeda untuk data yang sama. Fokus penelitian beralih dari reduksionisme (mempelajari gen tunggal) ke pendekatan holistik (memodelkan interaksi kompleks dalam

sistem kehidupan, menggabungkan data genomik, transkriptomik, dan proteomik).

Hierarki Data Bioinformatika

Kemajuan teknologi sekuensing dan informatika memungkinkan para peneliti untuk menangkap gambaran utuh dari aktivitas seluler pada berbagai tingkatan biologis. Pendekatan ini bertujuan untuk memahami bagaimana interaksi antara berbagai lapisan molekuler berkontribusi pada fenotipe. *High-throughput technology* dalam biologi molekuler menghasilkan sejumlah besar data sekuensing DNA dan RNA, berupa tumpukan data omics. Genom mewakili seluruh materi genetik dalam suatu organisme. Genomika adalah studi tentang urutan DNA, yang merupakan instruksi cetak biru permanen bagi sel. Data genomik memiliki karakteristik yang cenderung statis, yang berarti urutan DNA dasar tidak berubah secara drastis sepanjang hidup individu. Analisis difokuskan pada identifikasi variasi genetik seperti *Single Nucleotide Polymorphisms* (SNPs), insersi, delesi, dan variasi jumlah salinan (*Copy Number Variations*). Aplikasi melalui *Genome-Wide Association Studies* (GWAS), peneliti menghubungkan varian genetik tertentu dengan risiko penyakit, yang menjadi fondasi utama bagi pengobatan presisi.

Transkriptom adalah kumpulan lengkap semua transkrip RNA (khususnya mRNA) yang diproduksi oleh genom pada satu waktu tertentu. Berbeda dengan genom, transkriptom memiliki karakteristik data yang bersifat dinamis dan sangat bergantung pada kondisi lingkungan, jenis sel, serta tahap perkembangan. Teknologi yang digunakan yaitu *microarray* dan kini RNA-*sequencing* (RNA-Seq), yang memungkinkan pengukuran tingkat

ekspresi ribuan gen secara bersamaan. Transkriptomika berfungsi sebagai jembatan antara kode genetik statis (DNA) dan produk fungsional (protein), yang memberikan gambaran tentang gen mana yang aktif atau "menyala" dalam kondisi tertentu.

Proteom mencakup seluruh set protein yang diekspresikan oleh sel atau jaringan. Protein adalah pelaksana fungsi biologis yang sebenarnya di dalam tubuh. Proteomika jauh lebih kompleks daripada genomika karena adanya modifikasi pasca-translasi (*post-translational modifications*), serta fakta bahwa satu gen dapat menghasilkan banyak protein berbeda melalui *alternative splicing*. Analisis utama dilakukan menggunakan Spektrometri Massa (*Mass Spectrometry*) untuk mengidentifikasi dan mengukur kelimpahan protein. Data proteomik memberikan informasi langsung tentang jalur sinyal seluler dan interaksi fisik antar molekul yang tidak dapat diprediksi hanya dari data DNA atau RNA.

Epigenomika mempelajari modifikasi kimia pada DNA dan protein histon yang tidak mengubah urutan basa DNA asli, tetapi secara drastis memengaruhi ekspresi gen. Fokus utama meliputi metilasi DNA dan modifikasi histon (seperti asetilasi). Data epigenomik sangat penting karena menunjukkan bagaimana faktor luar (seperti diet, stres, dan paparan kimia) dapat meninggalkan "jejak" pada genom yang mengatur aktivitas genetik.

Metabolomika adalah studi tentang metabolit, yaitu molekul kecil (seperti gula, asam amino, dan lipid) yang merupakan produk akhir dari proses seluler. Metabolomik dapat digunakan sebagai indikator fenotipe karena metabolit mencerminkan hasil akhir dari interaksi genom, transkriptom, dan proteom dengan

lingkungan. Metabolomika dianggap sebagai indikator paling dekat dengan fenotipe. Analisis metabolit memungkinkan identifikasi *biomarker* penyakit dan pemahaman mengenai perubahan metabolisme secara sistemik.

Meskipun data genomik, transkriptomik, dan proteomik sering dianalisis secara terpisah (pendekatan monotematik), konsep Integrasi multi-omics yang menggabungkan lapisan-lapisan informasi sangatlah bermanfaat untuk memodelkan sistem biologis secara holistik. Hal ini sangat penting guna memahami fenomena biologi sistem yang kompleks pada sistem hayati, yang tidak terjadi disebabkan oleh satu gen tunggal, melainkan pada integrasi di berbagai tingkat molekuler.

DAFTAR PUSTAKA

Bayat, A. (2002). Science, medicine, and the future: Bioinformatics. *BMJ (Clinical Research Ed.), 324*(7344), 1018–1022. https://doi.org/10.1136/bmj.324.7344.1018

Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2019). A brief history of bioinformatics. *Briefings in Bioinformatics, 20*(6), 1981–1996. https://doi.org/10.1093/bib/bby063

Li, Y., & Chen, L. (2014). Big biological data: Challenges and opportunities. *Genomics, Proteomics & Bioinformatics, 12*(5), 187–189. https://doi.org/10.1016/j.gbp.2014.10.001

Liu, X., & Zhang, W. (2024). Bioinformatics in the age of big data: Leveraging computational tools for biological discoveries. *Computational Molecular Biology, 14*. https://doi.org/10.5376/cmb.2024.14.0020

Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods of Information in Medicine, 40*(4), 346–358. https://doi.org/10.1055/s-0038-1634431

Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., & Ferrari, R. (2018). Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics, 19*(2), 286–302. https://doi.org/10.1093/bib/bbw/114

Pal, S., Mondal, S., Das, G., Khatua, S., & Ghosh, Z. (2020). Big data in biology: The hope and present-day challenges in IT. *Gene Reports, 21*, 100869. https://doi.org/10.1016/j.genrep.2020.100869

BAB 2
Data Biologis dan Format File

Sonia Meta Angraini

Perkembangan bioinformatika sebagai bidang interdisipliner tidak dapat dilepaskan dari kemajuan teknologi biologi molekuler yang mampu menghasilkan data biologis dalam jumlah besar dan kompleks. Transformasi biologi dari ilmu yang bersifat deskriptif menjadi ilmu berbasis data menempatkan data biologis sebagai pusat proses analisis dan pengambilan kesimpulan ilmiah. Dalam konteks ini, kemampuan untuk memahami, mengelola, dan menginterpretasikan data biologis menjadi kompetensi dasar bagi mahasiswa dan peneliti di bidang bioinformatika.

Salah satu aspek penting dalam pengelolaan data biologis adalah penggunaan format file yang sesuai. Format file berfungsi sebagai wadah yang menghubungkan data biologis dengan proses analisis komputasi. Kesalahan dalam memahami atau menggunakan format file dapat menyebabkan hilangnya informasi penting, kesalahan analisis, ataupun interpretasi data biologi yang tidak tepat. Oleh karena itu, pemahaman mengenai format file bukan hanya bersifat teknis, tetapi juga memiliki implikasi langsung terhadap validitas hasil analisis bioinformatika.

Bab ini membahas secara komprehensif konsep data biologis, karakteristik dan klasifikasinya, serta format file utama yang umum digunakan dalam bioinformatika. Pembahasan difokuskan pada format file sekuens seperti FASTA, FASTQ, dan GenBank, yang merupakan format dasar dalam berbagai analisis bioinformatika modern. Selain itu, bab ini juga menjelaskan alur pengolahan data biologis, mulai dari tahap data mentah hasil sekuensing hingga siap dianalisis. Dengan pemahaman yang baik terhadap materi dalam bab ini, pembaca diharapkan memiliki landasan yang kuat untuk mempelajari metode analisis bioinformatika pada bab-bab selanjutnya.

Pengertian Data Biologis

Data biologis dapat didefinisikan sebagai representasi digital dari informasi biologis yang diperoleh melalui pengamatan, pengukuran, atau eksperimen terhadap sistem kehidupan. Dalam bioinformatika, data biologis mencerminkan struktur, fungsi, dan dinamika sistem biologis pada berbagai tingkat organisasi, mulai dari molekul hingga tingkat populasi.

Data biologis yang digunakan dalam bioinformatika tidak bersifat tunggal dan sederhana. Data tersebut mencakup berbagai tingkat organisasi kehidupan, mulai dari molekul seperti DNA, RNA, dan protein, hingga data seluler, jaringan, organisme, dan populasi. Setiap jenis data memiliki karakteristik, struktur, maupun kebutuhan pengelolaan yang berbeda. Perbedaan tersebut menuntut adanya sistem representasi data yang terstandarisasi agar data dapat disimpan dan dianalisis secara konsisten menggunakan perangkat lunak bioinformatika.

Karakteristik Data Biologis

Berbeda dengan data pada bidang ilmu lain, data biologis memiliki karakteristik khusus. Salah satu karakteristik utamanya adalah tingkat kompleksitas yang tinggi. Sistem biologis bersifat dinamis dan terdiri atas berbagai komponen yang saling berinteraksi, sehingga perubahan pada satu komponen dapat memengaruhi komponen lainnya. Kompleksitas ini tercermin dalam data biologis yang sering kali bersifat multidimensional dan sulit dianalisis tanpa pendekatan komputasional.

Selain kompleks, data biologis juga bersifat heterogen. Data dapat berasal dari berbagai sumber, seperti eksperimen laboratorium, simulasi komputasi (*in silico*), maupun observasi lapangan. Setiap sumber data memiliki metode pengambilan dan format representasi yang berbeda, sehingga diperlukan standarisasi untuk memungkinkan integrasi data lintas studi.

Karakteristik lain yang menonjol adalah volume data yang sangat besar. Perkembangan teknologi sekuensing modern dan berkecepatan tinggi memungkinkan pengurutan genom dalam skala besar dengan biaya yang relatif rendah. Akibatnya, jumlah data biologis yang dihasilkan meningkat secara eksponensial dan menimbulkan tantangan tersendiri dalam pengelolaan dan analisis data.

Selain itu, data biologis sangat bergantung pada konteks biologis. Informasi mengenai spesies, jaringan, kondisi lingkungan, dan metode eksperimen sangat memengaruhi interpretasi data. Tanpa metadata yang memadai, data biologis berpotensi menghasilkan kesimpulan yang keliru.

Klasifikasi Data Biologis

Klasifikasi Berdasarkan Tingkat Organisasi Biologi

Data biologis dapat diklasifikasikan berdasarkan tingkat organisasi kehidupan, mulai dari tingkat molekuler hingga populasi. Data molekuler, seperti DNA, RNA, dan protein, merupakan data yang paling banyak digunakan dalam bioinformatika karena dapat direpresentasikan secara langsung dalam bentuk sekuens digital.

Data seluler mencakup informasi mengenai ekspresi gen, interaksi molekuler, dan dinamika sel. Data ini sering digunakan untuk memahami mekanisme regulasi gen dan respons sel terhadap kondisi tertentu. Sementara itu, data jaringan dan organ diperoleh melalui teknik pencitraan biomedis dan analisis histologi, yang menghasilkan data visual digital yang merepresentasikan karakteristik struktural jaringan biologis.

Pada tingkat organisme dan populasi, data biologis mencakup data fenotip, variasi genetik, dan distribusi populasi. Data ini banyak digunakan dalam studi genetika populasi, evolusi, dan ekologi molekuler.

Data Molekuler dalam Bioinformatika

Data DNA

Data DNA merupakan fondasi utama bioinformatika. Urutan nukleotida DNA menyimpan informasi genetik yang mengatur struktur dan fungsi organisme. Analisis data DNA memungkinkan identifikasi gen, prediksi fungsi gen, serta studi hubungan evolusioner antarorganisme.

Data RNA

Data RNA mencerminkan ekspresi gen dan aktivitas transkripsi dalam sel. Analisis data RNA, khususnya melalui

pendekatan RNA-Seq, memungkinkan peneliti memahami pola ekspresi gen pada berbagai kondisi biologis dan lingkungan.

Data Protein

Data protein mencakup sekuens asam amino, struktur tiga dimensi, dan interaksi protein. Analisis data protein sangat penting dalam memahami mekanisme molekuler dan pengembangan terapi berbasis target protein.

Format File Data Biologis Dalam Bioinformatika

Format FASTA

Format FASTA merupakan salah satu format file paling sederhana dan paling luas digunakan dalam bioinformatika. Format ini digunakan untuk menyimpan sekuens biologis, baik DNA, RNA, maupun protein, tanpa menyertakan informasi kualitas sekuens. Kesederhanaan struktur FASTA menjadikannya kompatibel dengan hampir seluruh perangkat lunak bioinformatika dan sering digunakan sebagai format dasar dalam berbagai analisis sekuens.

Secara umum, file FASTA terdiri atas dua komponen utama, yaitu baris *header* dan baris sekuens. Baris *header* diawali dengan simbol ">" dan berisi identitas sekuens, seperti nama gen, spesies asal, atau informasi tambahan lainnya. Baris sekuens berisi urutan nukleotida atau asam amino yang dapat dituliskan dalam satu atau beberapa baris. Tidak terdapat aturan baku mengenai panjang baris sekuens, selama urutan karakter tetap konsisten.

Contoh berikut menunjukkan struktur sederhana file FASTA:
DNA

>BRCA1_Homo_sapiens_partial
ATGAAAGGTTTGGAAAGCTTCTGCTCCTGTTGGTGAC
AGCCTTCTCAGGGAAGTTGAGTTCTCTGAGACCTTGA
GACAGTGGCTTGGAACTGTGAGTGGATTTGAGGAGGA

Contoh berikut menunjukkan struktur sederhana file FASTA: Protein

>sp|P01308|INS_HUMAN Insulin precursor OS=Homo sapiens OX=9606 GN=INS PE=1 SV=1
MALWMRLLPLLALLALWGPDPAAA
FVNQHLCGSHLVEALYLVCGERGFF
YTPKTRREAEDLQVGQVELGGGPGAG

Keterbatasan utama format FASTA adalah tidak adanya informasi mengenai kualitas pembacaan sekuens. Oleh karena itu, FASTA tidak digunakan sebagai format data mentah hasil sekuensing, melainkan sebagai format data hasil pemrosesan.

Format FASTQ

Format FASTQ dikembangkan untuk mengatasi keterbatasan format FASTA dalam menyimpan informasi kualitas sekuens. FASTQ merupakan format standar untuk menyimpan data mentah hasil mesin sekuensing generasi baru (*next-generation sequencing*). Format ini menyimpan dua jenis informasi utama, yaitu urutan nukleotida dan skor kualitas untuk setiap basa.

Setiap entri FASTQ terdiri atas empat baris. Baris pertama berisi identitas sekuens dan diawali dengan simbol "@". Baris kedua berisi urutan nukleotida. Baris ketiga diawali dengan

simbol "+" sebagai pemisah, dan baris keempat berisi skor kualitas yang direpresentasikan dalam bentuk karakter ASCII.

Contoh Fomat FASTQ:

@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAA
+
!''*((((***+))%%%++)(%%%%).1***-+*''

Skor kualitas pada format FASTQ mencerminkan tingkat kepercayaan mesin sekuensing terhadap pembacaan setiap basa. Informasi ini sangat penting dalam tahap kontrol kualitas dan pemrosesan awal data sekuensing. Data dengan kualitas rendah dapat dipangkas atau dihapus untuk mencegah kesalahan analisis pada tahap selanjutnya.

Dalam praktik bioinformatika, FASTQ digunakan sebagai format awal sebelum data melalui tahap kontrol kualitas (*quality control*) dan pemangkasan (*trimming*). Setelah tahap ini, sekuens yang telah bersih dapat dikonversi ke format FASTA untuk analisis lebih lanjut.

Format GenBank dan Anotasi Gen

Format GenBank merupakan format file yang lebih kompleks dibandingkan FASTA dan FASTQ karena tidak hanya menyimpan sekuens nukleotida, tetapi juga informasi anotasi biologis yang kaya. Format ini dikembangkan oleh National Center for Biotechnology Information (NCBI) dan digunakan secara luas sebagai standar penyimpanan data sekuens teranotasi.

File GenBank memuat berbagai informasi penting, seperti identitas sekuens, klasifikasi taksonomi organisme, lokasi gen, *coding sequence* (CDS), serta produk protein hasil translasi. Informasi ini memungkinkan peneliti untuk memahami hubungan antara sekuens nukleotida dan fungsi biologisnya.

Contoh Format GenBank:

```
LOCUS       NM_000546    3937 bp    mRNA    linear
PRI 18-JUL-2023
DEFINITION  Homo sapiens tumor protein p53 (TP53),
mRNA.
ACCESSION   NM_000546
VERSION NM_000546.6
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Mammalia;
            Primates;
            Hominidae; Homo.
REFERENCE   1  (bases 1 to 3937)
  AUTHORS   Vousden,K.H., Lane,D.P.
  TITLE     p53 in health and disease
  JOURNAL   Nat. Rev. Mol. Cell Biol. 8(4):275-283 (2007)
FEATURES    Location/Qualifiers
     source     1..3937
                /organism="Homo sapiens"
                /mol_type="mRNA"
                /db_xref="taxon:9606"
     gene       1..3937
```

```
                /gene="TP53"
     CDS          117..1500
                /gene="TP53"
                /product="cellular tumor antigen p53"
                /protein_id="NP_000537.3"
                 /translation="MEEPQSDPSVEPPLSQETFSDLW
                 KLLPENNVLSPLPSQAMDDL
                 MLSPDDIEQWFTEDPGPDEAPRMPEAAPPVAP
                 APAAPTPAAPAPAPSWPLSSSVPSQK"
ORIGIN
        1 atggaggagg cccagagcga cccaggtggt gctgctcccc
        agccaggtcc
        61 agggccaggg cccctgagct gctggcctgg gctccctgcc
        agctggagct
        121 gctgctgctg gctggagctg gctgctggag ctcctgctgg
        agctgctgct
//
```

Anotasi gen dalam format GenBank berperan penting dalam bioinformatika karena memberikan konteks biologis terhadap data sekuens. Tanpa anotasi, sekuens hanya berupa rangkaian karakter tanpa makna biologis yang jelas. Oleh karena itu, format GenBank banyak digunakan sebagai sumber data referensi dalam analisis genom dan fungsi gen.

Format EMBL

Format EMBL merupakan format penyimpanan data sekuens nukleotida yang dikembangkan dan dikelola oleh *European Molecular Biology Laboratory* (EMBL). Secara fungsional, format EMBL memiliki kemiripan dengan format GenBank karena sama-

sama menyimpan sekuens nukleotida yang dilengkapi dengan informasi anotasi biologis, seperti identitas gen, lokasi fitur genom, dan keterangan fungsional.

Format EMBL umumnya digunakan sebagai bagian dari basis data internasional yang terintegrasi dengan GenBank dan DNA Data Bank of Japan (DDBJ). Ketiga basis data tersebut saling bertukar data secara berkala, sehingga informasi sekuens yang tersimpan pada EMBL pada prinsipnya setara dengan data yang tersedia pada GenBank, meskipun struktur penulisan dan penamaan kolomnya sedikit berbeda.

Dalam konteks bioinformatika pengantar, pemahaman terhadap format EMBL dinilai penting untuk mengenali variasi representasi data sekuens teranotasi yang digunakan secara global. Namun, dalam praktik analisis sehari-hari, format EMBL jarang digunakan secara langsung oleh pengguna pemula, karena sebagian besar perangkat lunak bioinformatika lebih banyak menggunakan format FASTA atau GenBank sebagai format input utama.

UniProt

UniProt merupakan basis data terintegrasi yang berfokus pada informasi protein, mencakup sekuens asam amino, anotasi fungsi, struktur, serta keterkaitan dengan jalur biologis dan penyakit. Tidak seperti FASTA, FASTQ, atau GenBank yang fungsi utamanya sebagai format file, UniProt lebih difungsikan sebagai sistem basis data protein yang menyediakan berbagai format representasi data.

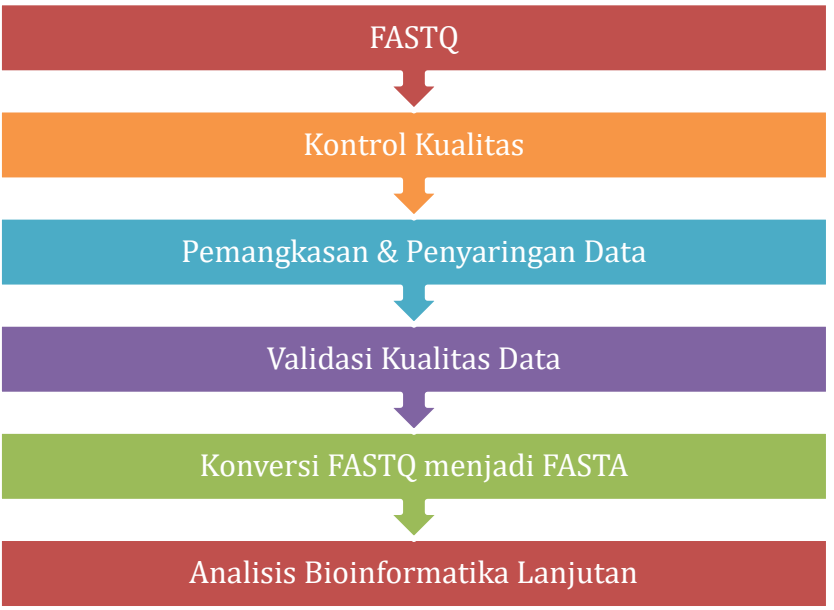
Data protein dalam UniProt dapat diunduh dalam format FASTA atau format teks teranotasi yang memuat informasi biologis tambahan, seperti fungsi protein, domain struktural,

modifikasi pascatranslasi, dan interaksi molekuler. Informasi ini menjadikan UniProt sebagai sumber referensi utama dalam analisis protein dan proteomik.

Dalam buku pengantar bioinformatika, UniProt diperkenalkan untuk menekankan bahwa data biologis tidak hanya mencakup sekuens nukleotida, tetapi juga data protein yang memiliki tingkat anotasi fungsional yang lebih kompleks. Pembahasan mendalam mengenai penggunaan UniProt dalam analisis protein umumnya ditempatkan pada bab lanjutan yang membahas basis data biologis dan aplikasi bioinformatika.

Alur Pengolahan Data FASTQ ke FASTA dalam Bioinformatika

Data sekuensing mentah dalam format FASTQ terlebih dahulu melalui tahap pengendalian kualitas dan pemangkasan untuk menghilangkan adaptor serta basa berkualitas rendah. Sekuens hasil penyaringan selanjutnya dikonversi ke format FASTA dengan menghilangkan informasi kualitas, sehingga menghasilkan data sekuens yang siap digunakan untuk berbagai analisis bioinformatika lanjutan (Gambar 2.1).



Sumber: Cock *et al.* (2010), Bolger *et al.* (2014), dan Conesa *et al.* (2016)

Gambar 2.1. Alur umum pengolahan data bioinformatika dari FASTQ ke FASTA

Data bioinformatika umumnya diproses melalui alur yang sistematis. Prosesnya dimulai dari sampel biologis yang dianalisis menggunakan mesin sekuensing dan menghasilkan data mentah dalam format FASTQ. Data ini kemudian melalui tahap kontrol kualitas untuk diidentifikasi dan dihilangkan sekuens yang kualitasnya rendah.

Setelah tahap kontrol kualitas dan pemangkasan selesai, data sekuens yang telah bersih dapat dikonversi ke format FASTA. Format FASTA kemudian digunakan dalam berbagai analisis bioinformatika, seperti penjajaran sekuens, pencarian kemiripan, dan anotasi gen. Transformasi format file ini

merupakan tahap krusial dalam memastikan bahwa data biologis dapat dianalisis secara akurat dan bermakna.

Integrasi Data dan Tantangan Format File

Integrasi data merupakan salah satu tantangan utama dalam bioinformatika modern, seiring dengan meningkatnya jumlah dan keragaman data biologis yang dihasilkan dari berbagai platform teknologi. Data biologis dapat berasal dari berbagai sumber, seperti hasil sekuensing genom, analisis transkriptom, proteomik, hingga data fenotip dan lingkungan. Setiap jenis data tersebut umumnya disimpan dalam format file yang berbeda dan memiliki struktur serta metadata yang tidak selalu seragam.

Perbedaan format file dan standar metadata sering kali menyulitkan proses integrasi data lintas studi maupun lintas platform. Sebagai contoh, data sekuens DNA dalam format FASTA atau FASTQ perlu dikaitkan dengan data anotasi gen dalam format GenBank atau GFF agar dapat diinterpretasikan secara biologis. Ketidaksesuaian format atau ketidaklengkapan metadata dapat menyebabkan hilangnya informasi penting dan menghambat proses analisis lanjutan.

Selain perbedaan format, tantangan lain dalam integrasi data biologis adalah volume data yang sangat besar. Data hasil *next-generation sequencing* dapat mencapai ukuran ratusan gigabita hingga terabita, sehingga memerlukan sistem penyimpanan dan pengelolaan data yang efisien. Format file yang tidak optimal dapat memperbesar kebutuhan ruang penyimpanan dan memperlambat proses analisis, terutama pada lingkungan komputasi berskala besar.

Tantangan integrasi data juga berkaitan erat dengan isu interoperabilitas antarperangkat lunak bioinformatika. Tidak

semua perangkat lunak mendukung format file yang sama, sehingga proses konversi format sering kali diperlukan. Proses konversi ini berpotensi menimbulkan kesalahan jika tidak dilakukan dengan hati-hati, terutama ketika melibatkan data yang kompleks dan teranotasi.

Oleh karena itu, pemahaman yang baik mengenai struktur, fungsi, dan keterbatasan setiap format file menjadi sangat penting dalam bioinformatika. Standarisasi format file dan metadata, serta penerapan prinsip pengelolaan data yang baik, merupakan langkah penting untuk memastikan bahwa data biologis dapat diintegrasikan dan dianalisis secara akurat serta dapat digunakan kembali dalam penelitian selanjutnya.

DAFTAR PUSTAKA

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores. *Nucleic Acids Research, 38*(6), 1767–1771. https://doi.org/10.1093/nar/gkp1137

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology, 17*, 13. https://doi.org/10.1186/s13059-016-0881-8

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics, 17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Lesk, A. M. (2019). *Introduction to bioinformatics* (5th ed.). Oxford University Press.

Mount, D. W. (2004). *Bioinformatics: Sequence and genome analysis* (2nd ed.). Cold Spring Harbor Laboratory Press.

National Center for Biotechnology Information. (2024). *GenBank overview*. National Institutes of Health. https://www.ncbi.nlm.nih.gov/genbank/

Pevsner, J. (2015). *Bioinformatics and functional genomics* (3rd ed.). Wiley-Blackwell.

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Robinson, G. E., & Schatz, M. C. (2015). Big data:

Astronomical or genomical? *PLOS Biology, 13*(7), e1002195. https://doi.org/10.1371/journal.pbio.1002195

BAB 3
Basis Data Biologi (*Biological Database*)

Armansyah Maulana Harahap

Pengantar

Studi Bioinformatika dan basis data biologis mengukuhkan posisinya sebagai entitas fundamental yang esensial dalam arsitektur penelitian, serta menyediakan landasan komprehensif bagi investigasi sistem biologis pada resolusi molekuler dan genomik. Struktur repositori terorganisir ini menampung spektrum data biologis yang ekstensif dan menyeluruh, namun tidak terbatas pada sekuensial nukleotida dan asam amino, karakterisasi struktural makromolekuler tiga dimensi, delineasi jejaring metabolik, profil ekspresi genetik, serta kompleksitas interaksi protein-protein dan ligan-protein.

Konsep Basis Data Biologis

Dalam bidang bioinformatika, data biologis merujuk pada kumpulan informasi yang diperoleh dari berbagai aspek kehidupan biologis, seperti sekuens DNA, data populasi, data genetik, serta data ekologi dan lain-lain (Agustini *et al.*, 2024). Namun demikian, fokus utama bioinformatika adalah pengelolaan dan analisis data yang berkaitan dengan biomolekul, yang diperoleh melalui metode eksperimental, kajian

literatur ilmiah, dan pendekatan komputasi. Dengan kemajuan teknologi komputasi dan pengembangan alat analisis bioinformatika, volume informasi biomolekul meningkat secara eksponensial, mencakup data sekuens DNA, RNA, dan protein. Data ini kemudian didigitalisasi dan diorganisasikan ke dalam basis data biologis yang memiliki struktur sistematis untuk mendukung penyimpanan, akses, dan pertukaran informasi secara global melalui jaringan internet (Jia *et al.*, 2025).

Basis data dapat diklasifikasikan berdasarkan berbagai kriteria, salah satunya adalah berdasarkan struktur penyimpanan datanya. Ada beberapa jenis struktur yang umum digunakan, yaitu file teks, file datar, basis data berorientasi objek, dan basis data relasional. Klasifikasi yang paling sering dipakai adalah berdasarkan jenis data yang disimpan, yakni basis data primer, sekunder, dan komposit (Jamroz *et al.*, 2014).

Klasifikasi Berdasarkan Struktur

- *Abstract Syntax Notation One* (ASN.1): Standar formal untuk mendefinisikan struktur data dan aturan *encoding* untuk representasi fisik data dalam aliran atau berkas berurutan (Sympli, 2021). Contohya seperti basis data CDD, Genbank, dan OMIM.
- Berkas Datar (*Flat File*): Implementasi basis data berbasis struktur tabel tunggal, dengan setiap baris mendefinisikan satu rekaman diskrit yang mengkompilasi seluruh atribut terkait suatu entitas (Souaifi *et al.*, 2025). Contohya seperti basis data EMBJ dan DDBJ.
- Basis Data Berorientasi Objek (*Object-Oriented Database*): dirancang untuk mengakomodasi tipe data kompleks dan terintegrasi efisien dengan Bahasa Pemrograman

Berorientasi Objek (OOPL), didefinisikan sebagai koleksi objek yang merepresentasikan instans entitas (Liu *et al.*, 2022). Contohya seperti basis data MITOMAP.

- Basis Data Relasional (*Relational Database*): Terkonseptualisasi sebagai kumpulan relasi atau tabel, dengan data diatur secara sistematis dalam format tabular, di mana setiap baris (tuple) memuat rekaman dan setiap kolom merepresentasikan atributnya. Contohya seperti basis data GDB dan SMART.
- *Extensible Markup Language* (XML): merupakan evolusi dari format Berkas Datar yang memberikan dukungan superior untuk merepresentasikan struktur data bersarang dan kompleks (Anza *et al.*, 2022). Contohya seperti basis data GO dan Swissport.

Klasifikasi Data Biologis

Basis Data Primer

Basis data primer adalah tempat penyimpanan data sekuens mentah dan data beranotasi yang diperoleh secara eksperimental seperti sekuens nukleotida, struktur tiga dimensi protein, serta menandakan sifat-sifat penting dari setiap sekuens (Khenifi *et al.*, 2023a). Basis data primer ini dapat diakses secara bebas melalui internet melalui *World Wide Web* (WWW). Basis data primer dapat diklasifikasikan lebih lanjut seperti yang diberikan di bawah ini:

1. Basis data urutan: Basis data sekuens menyimpan informasi mengenai sekuens DNA/nukleotida dan protein.
2. Basis data DNA/nukleotida: Database DNA/nukleotida menyimpan data tentang urutan DNA/nukleotida. Setiap

basis data memiliki perangkat pengiriman dan pengambilan data sendiri, tetapi mereka bertukar data setiap hari, sehingga semua basis data harus berisi rangkaian urutan yang sama. Beberapa contoh penting dari pangkalan data DNA/nukleotida diberikan di bawah ini.

Basis Data DNA/Nukleotida

GenBank

GenBank merupakan repositori akses terbuka yang teranotasi, mengkompilasi sekuens nukleotida dan translasi proteinnya. Datanya meliputi mRNA dengan daerah pengkodean, segmen DNA genomik (gen tunggal atau multipel), serta kelompok gen RNA ribosom. GenBank dikelola oleh National Center for Biotechnology Information (NCBI) sebagai bagian dari kolaborasi internasional dengan EMBL-EBI dan DDBJ. Submisi sekuens dapat dilakukan oleh laboratorium individu atau pusat sekuensing skala besar menggunakan BankIt (berbasis web) atau Sequin (perangkat lunak). Setelah submisi, staf GenBank akan menetapkan Nomor Aksesi dan melakukan pemeriksaan kualitas sebelum sekuens dirilis ke basis data. Pengambilan data GenBank dapat dilakukan melalui Entrez atau File Transfer Protocol (FTP). Repositori ini menyimpan informasi seperti Expressed Sequence Tag (EST), Sequence Tagged Site (STS), Genome Survey Sequence (GSS), High-Throughput Genome Sequence (HTGS), serta sekuens genom mikroba lengkap. Informasi GenBank dapat diakses di http://www.ncbi.nlm.nih.gov/genbank/. Metode pencarian dan pengambilan data mencakup Entrez Nucleotide, BLAST, dan NCBI E-utilities:

1. Cari GenBank untuk pengidentifikasi sekuens dan anotasi dengan Entrez Nucleotide, yang dibagi menjadi tiga divisi: CoreNucleotide (koleksi utama), dbEST (Expressed Sequence Tags), dan dbGSS (Genome Survey Sequences).
2. Mencari dan menyelaraskan sekuens GenBank dengan sekuens kueri menggunakan BLAST.
3. Mencari, menautkan, dan mengunduh sekuens secara terprogram menggunakan NCBI E-utilities.

DNA Data Base Japan (DDBJ)

Bank data sekuens nukleotida (DDB) mengumpulkan data sekuens nukleotida terutama dari peneliti Jepang, tetapi juga menerima data dan memberikan nomor akses kepada peneliti dari negara lain. Pada tahun 1986, DDBJ menjadi bank data di National Institute of Genetics (NIG). Saat ini, DDBJ beroperasi di NIG di Mishima, Jepang (Khenifi *et al*., 2023c). Informasi tentang DDBJ dapat ditemukan di servernya di http://www.ddbj.nig.ac.jp

Kegiatan utama DDBJ adalah sebagai berikut:

1. Sebagai anggota INSDC, DDBJ mengumpulkan data sekuens nukleotida dari para peneliti, memberikan nomor akses kepada para pengirim data, dan bertukar data ini setiap hari dengan EMBL-Bank dan GenBank.
2. DDBJ mengembangkan perangkat pengiriman dan pengambilan data bioinformatika.
3. DDBJ mengembangkan perangkat untuk analisis data biologi, serta
4. DDBJ mengadakan kursus pelatihan bioinformatika dalam bahasa Jepang untuk mengajarkan cara menganalisis data biologi.

European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI)

European Molecular Biology Laboratory (EMBL) adalah bagian dari European Bioinformatics Institute (EBI). EMBL-EBI, sekarang dikenal sebagai EMBL-Bank, didirikan pada tahun 1980 di EMBL di Heidelberg, Jerman. Menurut Kristianingsih *et al*. (2024), EMBL adalah basis data pertama di dunia yang mencakup urutan nukleotida. EMBL-EBI melakukan penelitian dasar dalam biologi komputasi, menyediakan data yang tersedia secara bebas dari eksperimen ilmu kehidupan, serta menyediakan program pelatihan pengguna yang luas bagi para peneliti. EMBL-EBI menyimpan data seperti DNA dan RNA (gen, genom, dan variasi), ekspresi gen (RNA, protein, dan ekspresi metabolit), protein (urutan, famili, dan motif), struktur (seluler dan molekuler), sistem (reaksi, interaksi, jalur), biologi kimiawi (kemogenomik dan metabolomik), ontologi (taksonomi dan kosakata yang terkontrol), dan literatur (publikasi ilmiah dan paten). EMBL-EBI dapat diakses melalui server http://www.ebi.ac.uk.

EnsEMBL

Ensembl adalah proyek bersama EBI, EMBL, dan Welcome Trust Sanger Institute. Tujuannya adalah untuk mengembangkan sistem perangkat lunak yang dapat menghasilkan dan memelihara anotasi genom eukariotik yang dipilih secara otomatis. Pada tahun 1999, EnsEMBL didirikan dengan tujuan untuk membuat anotasi genom secara otomatis, mengintegrasikan anotasi ini dengan data biologis lainnya yang tersedia, serta memberikan informasi kepada para peneliti melalui web. EnsEMBL menghasilkan basis data genom untuk vertebrata dan spesies eukariotik lainnya, sekaligus membuat

informasi ini tersedia secara bebas secara online. EnsEMBL dapat diakses secara gratis melalui server http://www.asia.ensembl.org. Banyak proyek penelitian di seluruh dunia menyumbangkan sekuens DNA dan data rakitan mereka ke EnsEMBL (Elkhalifa *et al.*, 2023)..

Sumber Informasi Protein (PIR)

Pada tahun 1984, National Biomedical Research Foundation (NBRF) mengembangkan PIR untuk membantu para peneliti menemukan dan memahami informasi tentang sekuens protein. Basis data ini adalah sumber daya informatika protein terintegrasi yang mendukung penemuan ilmiah dan penelitian genomik dan proteomik (Muslikh *et al.*, 2022). PIR terdiri dari tiga komponen yang berbeda. PIR1 mengandung entri yang diklasifikasikan dan dianotasi sepenuhnya, PIR2 mengandung entri pendahuluan yang belum ditinjau secara menyeluruh dan mengandung redundansi, dan PIR3 mengandung entri yang belum diverifikasi. Selain itu, basis data ini juga mencakup berbagai jenis sekuens, seperti terjemahan konseptual dari sekuens nukleotida, terjemahan dari sekuens yang belum ditranskripsi atau diterjemahkan secara eksperimental, sekuens protein hasil rekayasa genetika, serta sekuens yang tidak dikodekan secara genetik dan tidak disintesis melalui mekanisme ribosom. Basis data Sekuens Protein (PSD), yang berisi lebih dari 283.000 sekuens, dikelola oleh PIR http://www.pir.georgetown.edu.

UniProtKB/Swiss-Prot

UniProtKB/SwissProt adalah bagian yang dianotasi dan ditinjau secara manual dari basis pengetahuan UniProt (UniProtKB). Basis data ini berisikan sekuens protein yang

beranotasi dan tidak berlebihan (artinya sekuens yang kurang identik ada dalam basis data) (Chigurupati *et al.*, 2021). Swiss-Prot didirikan pada tahun 1986 dan dikelola secara kolaboratif oleh EMBL Outstation (EBI) dan Institut Bioinformatika Swiss (SIB). Basis data ini menyediakan informasi tentang struktur domain protein, fungsinya, modifikasi pasca-translasi, varian, dll. Basis data Swiss-Prot berbeda dengan basis data sekuens protein lainnya berdasarkan tiga kriteria utama, yaitu tingkat redundansi yang minimal, anotasi yang terkurasi secara manual dengan tingkat akurasi tinggi, serta integrasi yang kuat dengan basis data lain. Pada tahun 1996, tambahan beranotasi komputer untuk SWISS-PROT dibuat dan diberi nama TrEMBL (Terjemahan database urutan nukleotida EMBL). TrEMBL terdiri dari entri beranotasi komputer yang berasal dari terjemahan semua urutan pengkodean (*coding sequence*, CDS) dalam basis data EMBL, kecuali CDS yang telah ada di Swiss-Prot (Bairoch dan Apweiler, 2000). Server untuk SwissProt dan TrEMBL adalah www.ebi.ac.uk/uniprot dan www.ebi.ac.

Basis Data Sekuens Protein (PSD)

PIR-International Protein Sequence Database (PIR-PSD) adalah basis data pertama di dunia untuk sekuens protein yang diklasifikasikan dan diberi anotasi fungsional yang berkembang dari Atlas Sekuens dan Struktur Protein (1965-1978) yang disunting oleh Margaret Dayhoff. PIR-PSD diproduksi dan didistribusikan oleh Protein Information Resource, yang bekerja sama dengan MIPS (Munich Information Center for Protein Sequences) dan JIPID (Japan International Protein Information Database). PIR-PSD telah menjadi basis data sekuens protein yang paling komprehensif dan dikuratori oleh para ahli selama

lebih dari 20 tahun. URL untuk PSD adalah http:// www.pir.georgetown.edu

Basis data struktur

Basis data struktur dibuat untuk menyimpan informasi struktur makromolekul biologis, misalnya asam nukleat dan protein. Beberapa basis data struktur yang penting diberikan di sini di bawah ini:

Protein Data Bank (PDB)

Bank Data Protein didirikan di National Lab of Brookhaven (BNL) pada tahun 1971. PDB, yang dikelola oleh Research Collaboratory for Structural Bioinformatics (RCSB) di Rutgers University, berisi struktur protein 3D yang dibuat dengan kristalografi sinar-x dan resonansi magnetik nuklir (NMR). Struktur protein yang tersedia di PDB memberikan informasi tentang koordinat atom asam amino dalam protein, fragmen protein, ataupun protein yang terikat pada substrat atau inhibitor (M. Liu *et al.*, 2024). PDB dapat ditemukan di link berikut: http://www.rcsb.org/pdb/.

Nucleic Acid Database (NDB)

Tujuan dari Proyek Basis Data Asam Nukleat (NDB) adalah untuk menyimpan dan menyebarkan informasi struktural tentang asam nukleat. Database Asam Nukleat adalah basis data yang berisi informasi menyeluruh tentang struktur asam nukleat dalam tiga dimensi. David Beveridge dari Universitas Wesleyan dan Helen M. Berman dan Wilma K. Olson dari Universitas Rutgers mendirikan NDB pada tahun 1992 (Chakrabarty *et al.*, 2019).

Cambridge Crystallographic Data Centre (CCDB/CSD)

CCDC adalah pusat data kristalografi nirlaba yang berlokasi di Cambridge, Inggris. Basis data ini mengkompilasi dan mendistribusikan Basis data Struktur Cambridge (CSD), yang merupakan basis data molekul-molekul kecil yang menyimpan data struktur kristal organik dan logam-organik yang ditentukan secara eksperimental (Chen *et al.*, 2025). CSD berlokasi di www.ccdc.cam.ac.uk.

Basis data sekunder

Basis data sekunder berisi hasil analisis data sekuens primer dalam bentuk pola reguler, sidik jari, blok, profil, atau Model Markov Tersembunyi. Dengan demikian, basis data sekunder berisi sekuens yang dilestarikan, sekuens *marker*, dan residu situs aktif famili protein yang diperoleh dari penyejajaran sekuens dari sekumpulan protein terkait. Basis data sekunder termasuk ProSite, Profil, Cetakan, Pfam, dan REBase. Empat jenis basis data sekunder adalah sekuens, genom, struktur, dan jalur (Susianti *et al.*, 2025).

ProSite

ProSite adalah database protein sekunder yang menyimpan profil protein, pola asam amino, domain, dan situs fungsional protein. Tim dari Institut Bioinformatika Swiss menyusun datanya secara manual. Amos Bairoch membuat ProSite pada tahun 1988. ProSite memungkinkan analisis sekuens protein dan deteksi motif. Deskripsi domain ProSite menciptakan basis data ProRule, yang menyediakan informasi tentang asam amino fungsional dan struktural yang utama (Yan *et al.*, 2025). Akses ke ProSite dapat ditemukan di lin berikut http://www.prosite.expasy.org.

Pfam

Pfam adalah database famili protein yang menyediakan penyelarasan beberapa urutan dan Model Markov Tersembunyi (HMM) untuk domain protein dari setiap jenis organisme. HMM digunakan untuk mengumpulkan data penjajaran sekuens ganda (el-dien *et al.*, 2024). Pfam dapat dicari untuk struktur domain protein, distribusi spesies, mengikuti basis data lain, dan struktur protein yang diketahui. Pfam terdiri dari dua komponen, yaitu: Pfam-A dan Pfam-B. Pfam-A menyimpan entri berkualitas tinggi yang dikurasi secara manual. Untuk setiap entri, penyelarasan urutan protein dan HMM juga disimpan (Pinipay *et al.*, 2023). Famili Pfam-B yang berkualitas rendah dapat digunakan karena entri Pfam-A tidak mencakup semua protein yang diketahui, sehingga entri berkualitas rendah secara otomatis disimpan di Pfam-B (Tran *et al.*, 2021).

ExplorEnz

ExplorEnz dirancang, dikembangkan, dan dikelola oleh Andrew McDonald di School of Biochemistry and Immunology di Trinity College di Dublin, Irlandia. Dia menjalankan ExplorEnz untuk mengakses data dari daftar nomenklatur enzim International Union of Biochemistry and Molecular Biology (IUBMB) (Ding *et al.*, 2022).

REBase

Sebelum tahun 1980, Richard J. Roberts telah mengelola REBase, yang merupakan basis data enzim restriksi (Roberts, 1980). Informasi tentang enzim restriksi dan protein yang terkait dapat ditemukan dalam basis data ini. Informasi ini termasuk

referensi yang dipublikasikan dan tidak dipublikasikan, situs pengenalan dan pembelahan, *isoschizomer*, ketersediaan komersial, sensitivitas metilasi, kristal, genom, data sekuens, DNA metiltransferase, endonuklease *homing*, enzim *nicking*, subunit spesifisitas, dan protein kontrol (Jadhav *et al.*, 2025). Website resmi REBase adalah http://www.rebase.neb.com.

Basis data terkait genom

Mendelian Inheritance in Man (OMIM), yang dimulai pada awal tahun 1960an, adalah sebuah basis data online yang menyimpan informasi tentang gen manusia, beserta kelainan genetik, variasi genetik, dan gambar. Basis data ini juga menyediakan referensi untuk penelitian lebih lanjut dan alat untuk analisis genom dari gen yang dikatalogkan (Hassanin *et al.*, 2024).

Basis Data Faktor Transkripsi Tanaman (PlnTFDB)

PlnTFDB adalah basis data publik yang mengidentifikasi dan menyimpan semua gen tanaman yang terlibat dalam kontrol transkripsi. *The Plant Transcription Factor Database* (PlnTFDB; http://plntfdb.bio.uni-potsdam.de/v3.0/) adalah sebuah basis data integratif yang menyediakan kumpulan lengkap faktor transkripsi (TF) dan pengatur transkripsi (TR) lainnya pada spesies tanaman yang genomnya telah diurutkan secara lengkap dan diberi keterangan. Set lengkap 84 famili TF dan TR dari 19 spesies mulai dari ganggang merah dan hijau uniseluler hingga angiospermae termasuk dalam PlnTFDB, yang mewakili >1,6 miliar tahun evolusi jaringan regulasi gen (Kamau *et al.*, 2024)

Famili Protein yang Mirip Secara Struktural (FSSP)

Basis data protein yang mirip secara struktural FSSP dibuat menggunakan algoritme "*Alignment* Distance-matrix" (DALI) dan bermanfaat untuk perbandingan struktur protein (Abraham *et al.*, 2021).

Distance Matrix Alignment

Dali adalah basis data struktur sekunder, berdasarkan perbandingan struktur protein 3D dari Protein Data Bank (PDB).

Basis data informasi jalur

Basis data informasi jalur memberikan informasi tentang interaksi molekuler dalam proses biologis, informasi tentang gen dan protein, serta informasi tentang senyawa kimia dan reaksinya(el-dien *et al.*, 2024).

Basis data komposit

Basis data komposit menyimpan data dari basis data primer yang berbeda, sehingga tidak perlu mencari beberapa basis data primer untuk urutan nukleotida, urutan protein, struktur protein, dll. Contoh dari beberapa basis data komposit adalah:

1. nrdb (basis data nonredundan) menggabungkan dan menyimpan sekuens dari GenBank (terjemahan CDS), PDB, Swiss-Prot, PIR, dan PRF.
2. INSD (Basis Data Urutan Nukleotida Internasional) menyimpan urutan nukleotida dari EMBL, GenBank, dan DDBJ.
3. UniProt (basis data sekuens protein universal) adalah kumpulan sekuens protein dari PIR-PSD, Swiss-Prot, dan TrEMBL.

Lainnya

Dalam kategori basis data lainnya, terdapat basis data berbasis literatur dan basis data keanekaragaman hayati. Basis data khusus organisme memuat informasi detail tentang spesies non-manusia, seperti virus, bakteri, jamur, invertebrata (misalnya Drosophila), termasuk peta genom, ekspresi gen, mutasi, dan ciri-ciri morfologis. Sementara itu, basis data berbasis literatur mengkompilasi artikel-artikel ilmiah, baik dalam bentuk abstrak maupun makalah lengkap, yang dipublikasikan dalam berbagai jurnal. PubMed adalah salah satu basis data literatur terkemuka. Secara keseluruhan, basis data ini menyediakan informasi esensial mengenai keragaman hayati organisme.

Sistem pengambilan basis data biologi

Tiga sistem pengambilan data yang paling penting adalah Entrez (di NCBI), Sistem Pencarian Segmen (di EBI), SRS (di EBI), dan DBGET/LinkDB (di Jepang). Sistem pencarian ini tidak hanya mengembalikan kecocokan kueri, tetapi juga memberikan informasi tambahan yang penting dalam basis data terkait (Abraham *et al.*, 2021).

ENTREZ

Entrez dikembangkan oleh National Center for Biotechnology (NCBI) dan merupakan sistem pencarian dan pangkalan data biologi molekuler. Entrez (http://www.ncbi.nlm.nih.gov/Entrez/) menawarkan akses ke berbagai pangkalan data, termasuk pangkalan data urutan nukleotida seperti GenBank/DDBJ/EBI; pangkalan data urutan protein seperti Swiss-Prot, PIR, PRF, PDB, dan sekuens protein

yang diterjemahkan dari pangkalan data urutan DNA; data pemetaan genom dan kromosom; pangkalan data struktur 3-D pemodelan molekuler; serta pangkalan data kepustakaan PubMed. Salah satu fitur paling penting Entrez adalah kemampuan untuk memberikan akses ke artikel-artikel yang terkait dari basis data yang relevan (Perdana Istyastono, 2015).

***Sequence Retrieval System* (SRS)**

Dalam biologi molekuler, SRS adalah peramban jaringan untuk basis data. SRS mencari delapan puluh pangkalan data biologi yang dikembangkan di *European Bioinformatics Institute* (EBI) di Hinxton, Inggris. SRS memberikan akses ke database yang mencakup sekuens dan urutan, jalur metabolisme, faktor transkripsi, struktur protein 3D, genom, pemetaan, mutasi, dan mutasi lokus tertentu. Keuntungan pencarian kueri cepat SRS adalah memungkinkan pengguna mengambil, menautkan, dan mengakses entri dari semua sumber daya yang saling berhubungan (Subhan *et al.*, 2025).

DBGET/LinkDB

GenomeNet, yang dikembangkan oleh Institute for Chemical Research di Universitas Kyoto dan Human Genome Center di Universitas Tokyo, adalah sistem pengambilan basis data bioinformatika yang terintegrasi. Anda dapat mengakses sistem ini melalui URL berikut: http://www.genome.ad.jp/dbget/. DBGET dapat mengakses sekitar 20 basis data biologi molekuler yang dapat ditelusuri secara individual. Setelah melakukan kueri pada salah satu pangkalan data, DBGET memberikan tautan untuk menghubungkan data ke daftar hasil. DBGET dapat dikaitkan

dengan database Kyoto Encyclopedia of Genes and Genomes (KEGG) (Muslikh *et al.*, 2022).

DAFTAR PUSTAKA

Abraham, J., Aft, R., Agnese, D., Allison, K. H., Anderson, B., Blair, S. L., Burstein, H. J., Dang, C., Elias, A. D., Giordano, S. H., Kumar, N. R., Burns, J., Goetz, M. P., Goldstein, L. J., Hurvitz, S. A., Isakoff, S. J., Jankowitz, R. C., Javid, S. H., Krishnamurthy, J., … Young, J. S. (2021). *NCCN clinical practice guidelines in oncology: Breast cancer (Version 5.2021)*. National Comprehensive Cancer Network.

Agustini, K., Sangande, F., Nuralih, N., Harahap, A. M., Ningsih, S., & Bahtiar, A. (2024). Molecular mechanism elucidation of *Ocimum basilicum* as anticancer using system bioinformatic approach supported by in vitro assay. *Pharmacia, 71*, 1–12. https://doi.org/10.3897/pharmacia.71.e127395

Anza, M., Endale, M., Cardona, L., Cortes, D., Eswaramoorthy, R., Cabedo, N., Abarca, B., Zueco, J., Rico, H., Domingo-Ortí, I., & Palomino-Schätzlein, M. (2022). Cytotoxicity, antimicrobial activity, molecular docking, drug likeness and DFT analysis of benzo[c]phenanthridine alkaloids from roots of *Zanthoxylum chalybeum*. *Biointerface Research in Applied Chemistry, 12*(2), 1569–1586. https://doi.org/10.33263/BRIAC122.15691586

Chakrabarty, N., Haque, M. T., Ershad, M., Kabir, M. A., Shams, M. R., Tahsin, F., Milonuzzaman, M., Al Haque, M. M., Hasan, M. Z., Al Mahabub, A., Rahman, M. M., & Patwary, M. R. U. (2019). Antidiarrheal activity of methanol extract of *Piper sylvaticum* (Roxb.) stem in mice and *in silico* molecular docking of its isolated compounds. *Discovery Phytomedicine, 6*(3). https://doi.org/10.15562/phytomedicine.2019.97

Chen, F., Cai, Y., Chen, C., & Zhou, J. (2025). Network pharmacology integrated with molecular docking and

experimental validation elucidates the therapeutic potential of *Forsythiae Fructus* extract against hepatitis B virus-related hepatocellular carcinoma. *Frontiers in Oncology, 15*. https://doi.org/10.3389/fonc.2025.1571537

Chigurupati, S., Alharbi, F. S., Almahmoud, S., Aldubayan, M., Almoshari, Y., Vijayabalan, S., Bhatia, S., Chinnam, S., & Venugopal, V. (2021). Molecular docking of phenolic compounds and screening of antioxidant and antidiabetic potential of *Olea europaea* L. ethanolic leaves extract. *Arabian Journal of Chemistry, 14*(11). https://doi.org/10.1016/j.arabjc.2021.103422

Ding, J., Wu, J., Wei, H., Li, S., Huang, M., Wang, Y., & Fang, Q. (2022). Exploring the mechanism of hawthorn leaves against coronary heart disease using network pharmacology and molecular docking. *Frontiers in Cardiovascular Medicine, 9*. https://doi.org/10.3389/fcvm.2022.804801

El-Dien, R. T. M., Maher, S. A., Hisham, M., Saber, E. A., Khedr, A. I. M., Fouad, M. A., Kamel, M. S., & Mahmoud, B. K. (2024). Network pharmacology, molecular docking study, and in vivo validation of the wound healing activity of the Red Sea soft coral *Paralemnalia thyrsoides* (Ehrenberg, 1834) ethanolic extract and bioactive metabolites. *Beni-Suef University Journal of Basic and Applied Sciences, 13*(1). https://doi.org/10.1186/s43088-024-00512-x

Elkhalifa, A. E. O., Banu, H., Khan, M. I., & Ashraf, S. A. (2023). Integrated network pharmacology, molecular docking, molecular simulation, and in vitro validation revealed the bioactive components in soy-fermented food products and the underlying mechanistic pathways in lung cancer. *Nutrients, 15*(18). https://doi.org/10.3390/nu15183949

Hassanin, S. O., Hegab, A. M. M., Mekky, R. H., Said, M. A., Khalil, M. G., Hamza, A. A., & Amin, A. (2024). Combining in vitro, in vivo, and network pharmacology assays to identify targets and molecular mechanisms of spirulina-derived biomolecules against breast cancer. *Marine Drugs, 22*(7). https://doi.org/10.3390/md22070328

Jadhav, P. A., Thomas, A. B., Pathan, M. K., Chaudhari, S. Y., Wavhale, R. D., & Chitlange, S. S. (2025). Unlocking the therapeutic potential of unexplored phytocompounds as hepatoprotective agents through integration of network pharmacology and *in silico* analysis. *Scientific Reports, 15*(1), 8425. https://doi.org/10.1038/s41598-025-92868-y

Jamroz, M., Kolinski, A., & Kmiecik, S. (2014). CABS-flex predictions of protein flexibility compared with NMR ensembles. *Bioinformatics, 30*(15), 2150–2154. https://doi.org/10.1093/bioinformatics/btu184

Jia, J., Cui, E., Li, S., Li, B., & Ge, Q. (2025). An exploratory study on the molecular targets and interaction mechanisms of *Citri reticulatae pericarpium* in the treatment of chronic obstructive pulmonary disease based on GEO combined with bioinformatics. *Chinese Journal of Analytical Chemistry*. https://doi.org/10.1016/j.cjac.2025.100516

Kamau, S. W., Ngugi, M. P., Mwitari, P. G., & Njeru, S. N. (2024). Network pharmacology, molecular docking and experimental approaches of the anti-proliferative effects of *Rhamnus prinoides* ethyl-acetate extract in cervical cancer cells. *Heliyon, 10*(17). https://doi.org/10.1016/j.heliyon.2024.e37324

Khenifi, M. L., Serseg, T., Migas, P., Krauze-Baranowska, M., Özdemir, S., Bensouici, C., Alghonaim, M. I., Al-Khafaji, K.,

Alsalamah, S. A., Boudjeniba, M., Yousfi, M., Boufahja, F., Bendif, H., & Mahdid, M. (2023). HPLC-DAD-MS characterization, antioxidant activity, α-amylase inhibition, molecular docking, and ADMET of flavonoids from fenugreek *seed*s. *Molecules, 28*(23). https://doi.org/10.3390/molecules28237798

Kristianingsih, A., Soetrisno, R., Reviono, R., & Wasita, B. (2024). Molecular docking study of quercetin from ethanol extract of *Mimosa pudica* Linn on asthma biomarkers. *Tropical Journal of Natural Product Research, 8*(10), 8640–8645. https://doi.org/10.26538/tjnpr/v8i10.4

Li, Y., Kuhn, M., Gavin, A. C., & Bork, P. (2020). Identification of metabolites from tandem mass spectra with a machine learning approach utilizing structural features. *Bioinformatics, 36*(4), 1213–1218. https://doi.org/10.1093/bioinformatics/btz736

Liu, H., Mohammed, S. A. D., Lu, F., Chen, P., Wang, Y., & Liu, S. (2022). Network pharmacology and molecular docking-based mechanism study to reveal antihypertensive effect of Gedan Jiangya decoction. *BioMed Research International, 2022*. https://doi.org/10.1155/2022/3353464

Liu, M., Wang, Y., Deng, W., Xie, J., He, Y., Wang, L., Zhang, J., & Cui, M. (2024). Combining network pharmacology, machine learning, molecular docking and molecular dynamic to explore the mechanism of Chufeng Qingpi decoction in treating schistosomiasis. *Frontiers in Cellular and Infection Microbiology, 14*. https://doi.org/10.3389/fcimb.2024.1453529

Muslikh, F. A., Samudra, R. R., Ma'arif, B., Ulhaq, Z. S., Hardjono, S., & Agil, M. (2022). *In silico* molecular docking and ADMET

analysis for drug development of phytoestrogens compound with its evaluation of neurodegenerative diseases. *Borneo Journal of Pharmacy, 5*(4), 357–366. https://doi.org/10.33084/bjop.v5i4.3801

Perdana Istyastono, E. (2015). Uji *in silico* senyawa emodin sebagai ligan pada reseptor estrogen alfa. *[Nama jurnal tidak tersedia], 12*(2), 48–53.

Pinipay, F., Rokkam, R., Botcha, S., & Tamanam, R. R. (2023). Evaluating the anti-urolithiasis potential of *Ficus religiosa seed* GC–MS evaluated phytoconstituents based on their in vitro antioxidant properties and *in silico* ADMET and molecular docking studies. *Clinical Phytoscience, 9*(1). https://doi.org/10.1186/s40816-023-00359-2

Song, J., Cui, Q., & Gao, J. (2024). Roles of lncRNAs related to the p53 network in breast cancer progression. *Frontiers in Oncology, 14*. https://doi.org/10.3389/fonc.2024.1453807

Souaifi, M., Dhahbi, W., Jebabli, N., Ceylan, H. İ., Boujabli, M., Muntean, R. I., & Dergaa, I. (2025). Artificial intelligence in sports biomechanics: A scoping review on wearable technology, motion analysis, and injury prevention. *Bioengineering, 12*(8). https://doi.org/10.3390/bioengineering12080887

Subhan, M., Sanachai, K., Sungthong, B., Datham, S., Ratha, J., & Puthongking, P. (2025). Comparison in vitro and *in silico* studies of phenolic acids and flavonoids on α-glucosidase inhibition. *Journal of Pharmacy and Pharmacognosy Research, 13*(1), 311–323. https://doi.org/10.56499/jppres24.2081_13.1.311

Sun, Z., & Sun, P. (2022). Attitude monitoring algorithm for volleyball sports training based on machine learning in the

context of artificial intelligence. *Security and Communication Networks, 2022*. https://doi.org/10.1155/2022/2907393

Susianti, S., Bahri, S., Hadi, S., Setiawansyah, A., Rachmadi, L., Fadilah, I., & Ratna, M. G. (2025). Integrating the network pharmacology and molecular docking confirmed with in vitro toxicity to reveal potential mechanism of non-polar fraction of *Cyperus rotundus* Linn as anti-cancer candidate. *Journal of Multidisciplinary Applied Natural Science, 5*(1), 56–73. https://doi.org/10.47352/jmans.2774-3047.228

Sympli, H. D. (2021). Estimation of drug-likeness properties of GC–MS separated bioactive compounds in rare medicinal *Pleione maculata* using molecular docking technique and SwissADME *in silico* tools. *Network Modeling Analysis in Health Informatics and Bioinformatics, 10*(1). https://doi.org/10.1007/s13721-020-00276-1

Tran, V., Kim, R., Maertens, M., Hartung, T., & Maertens, A. (2021). Similarities and differences in gene expression networks between the breast cancer cell line Michigan Cancer Foundation-7 and invasive human breast cancer tissues. *Frontiers in Artificial Intelligence, 4*. https://doi.org/10.3389/frai.2021.674370

Yan, K., Zhang, R. K., Wang, J. X., Chen, H. F., Zhang, Y., Cheng, F., Jiang, Y., Wang, M., Wu, Z., Chen, X. G., Chen, Z. N., Li, G. J., & Yao, X. M. (2025). Using network pharmacology and molecular docking technology, proteomics and experiments were used to verify the effect of Yigu decoction (YGD) on the expression of key genes in osteoporotic mice. *Annals of Medicine, 57*(1). https://doi.org/10.1080/07853890.2024.2449225

BAB 4
Analisis Sekuens DNA dan RNA

Kiky Mey Putranty

Pengantar Analisis Sekuens DNA dan RNA

Analisis sekuens DNA dan RNA merupakan salah satu pilar utama dalam bioinformatika modern (Mount, 2004; Brown, 2016). Sekuens biologis menyimpan informasi genetik yang berperan penting dalam menentukan struktur, fungsi, serta regulasi proses biologis pada tingkat molekuler. Perkembangan teknologi sekuensing berkecepatan tinggi telah menghasilkan data sekuens dalam jumlah yang sangat besar, sehingga pendekatan komputasional menjadi esensial untuk mengelola, menganalisis, dan menginterpretasikan data tersebut.

DNA berfungsi sebagai penyimpan informasi genetik jangka panjang, sedangkan RNA berperan sebagai perantara dan regulator ekspresi gen. Analisis sekuens DNA umumnya bertujuan untuk memahami organisasi genom, mengidentifikasi gen dan elemen regulator, serta mendeteksi variasi genetik. Sementara itu, analisis sekuens RNA lebih menitikberatkan pada dinamika ekspresi gen dan respons sel terhadap kondisi tertentu (Altschul *et al.*, 1990).

Dalam bioinformatika, analisis sekuens tidak hanya melibatkan penggunaan perangkat lunak, tetapi juga pemahaman biologis yang mendalam. Peneliti perlu mampu

menghubungkan hasil analisis komputasional dengan fenomena biologis yang berkaitan. Oleh karena itu, bab ini disusun untuk memberikan gambaran menyeluruh mengenai konsep, metode, dan alur kerja analisis sekuens DNA dan RNA sebagai pengantar bagi pembaca pemula.

Jenis Data Sekuens DNA dan RNA

Data sekuens DNA dan RNA dihasilkan dari berbagai teknologi sekuensing yang terus berkembang (Goodwin *et al.*, 2016; Mardis, 2017). Metode sekuensing generasi pertama, seperti *Sanger sequencing*, menghasilkan data dengan akurasi tinggi tetapi berkapasitas rendah. Sebaliknya, teknologi *next-generation sequencing (NGS)* memungkinkan pembacaan jutaan hingga miliaran fragmen sekuens secara paralel dengan biaya yang lebih efisien.

Hasil sekuensing DNA dan RNA umumnya disimpan dalam format standar seperti FASTA dan FASTQ. Format FASTA berisi informasi urutan nukleotida, sedangkan FASTQ menyertakan informasi tambahan berupa skor kualitas setiap basa. Nilai kualitas ini penting untuk mengetahui keandalan data sebelum dilakukan analisis lanjutan.

Pada analisis RNA-seq, data sekuens merepresentasikan fragmen RNA yang telah dikonversi menjadi cDNA. Data ini menunjukkan tingkat ekspresi gen dalam suatu kondisi biologis tertentu. Pemahaman terhadap jenis data, format file, dan karakteristik teknologi *sequencing* merupakan langkah awal yang penting dalam analisis sekuens DNA dan RNA.

Pra-Pemrosesan Data Sekuens (*Pre-processing*)

Data sekuens DNA dan RNA yang dihasilkan oleh teknologi sekuensing umumnya masih berupa data mentah (*raw data*) yang mengandung berbagai pengotor teknis. Pengotor tersebut dapat berupa sekuens *adaptor*, basa dengan kualitas rendah, maupun kesalahan pembacaan selama proses sekuensing. Oleh karena itu, sebelum dilakukan analisis lanjutan, data sekuens perlu melalui tahap pra-pemrosesan untuk memastikan kualitas dan keandalan hasil analisis bioinformatika.

Quality Control (QC) Data Sekuens

Quality control (QC) bertujuan untuk mengevaluasi kualitas data sekuens sebelum dilakukan analisis lebih lanjut. Pada tahap ini, berbagai parameter kualitas diperiksa untuk mengidentifikasi potensi masalah pada data mentah hasil sekuensing.

Beberapa parameter kualitas yang umum dianalisis meliputi distribusi kualitas basa, panjang sekuens, kandungan GC, serta keberadaan *adaptor sequencing*. Nilai kualitas basa biasanya dinyatakan dalam skor Phred, yang mencerminkan tingkat kepercayaan terhadap ketepatan pembacaan basa tertentu. Semakin tinggi skor Phred, semakin tinggi pula kualitas sekuens tersebut (Xiong, 2008).

Hasil *quality control* umumnya disajikan dalam bentuk grafik dan ringkasan statistik, sehingga memudahkan peneliti untuk menilai apakah data sekuens layak digunakan atau memerlukan pemrosesan lanjutan. Tahap QC tidak mengubah data, melainkan memberikan gambaran awal mengenai kondisi dan kualitas data sekuens.

Trimming dan Filtering Sekuens

Setelah dilakukan *quality control*, tahap berikutnya adalah *trimming*(pemotongan) dan *filtering* data sekuens. *Trimming* bertujuan untuk menghilangkan bagian sekuens yang tidak diinginkan, seperti *adaptor* sekuensing dan basa berkualitas rendah di ujung sekuens. Sementara itu, *filtering* dilakukan untuk membuang sekuens yang tidak memenuhi kriteria kualitas tertentu, misalnya sekuens dengan panjang terlalu pendek atau kualitas rata-rata yang rendah.

Proses *trimming* dan *filtering* sangat penting untuk meningkatkan akurasi analisis selanjutnya. Sekuens yang mengandung banyak kesalahan dapat menyebabkan pemetaan yang keliru pada tahap *alignment* atau menghasilkan *assembly* genom yang tidak akurat. Dalam analisis RNA-seq, data berkualitas rendah juga dapat memengaruhi perhitungan tingkat ekspresi gen.

Kriteria *trimming* dan *filtering* dapat disesuaikan dengan tujuan analisis dan karakteristik data sekuensing yang digunakan. Oleh karena itu, peneliti perlu memahami dampak dari setiap parameter yang diterapkan agar tidak menghilangkan informasi biologis yang penting.

Penghilangan Sekuens yang tidak diperlukan dan Kontaminasi

Selain *adaptor* dan basa berkualitas rendah, data sekuens juga dapat mengandung senyawa lain, seperti kontaminasi dari organisme lain, sekuens ribosomal RNA pada data RNA-seq, atau duplikasi sekuens akibat proses amplifikasi. Pengotor ini dapat mengganggu interpretasi biologis jika tidak ditangani dengan baik.

Sebagai contoh pada analisis RNA-seq, keberadaan rRNA dalam jumlah besar dapat mengurangi proporsi sekuens mRNA yang informatif. Oleh karena itu, beberapa *workflow* analisis RNA-seq menyertakan tahap khusus untuk mengidentifikasi dan mengurangi kontaminasi rRNA. Penghilangan pengotor dan kontaminasi bertujuan untuk memastikan bahwa data yang dianalisis benar-benar merepresentasikan target biologis yang diinginkan.

Perangkat Lunak untuk Pra-Pemrosesan Data Sekuens

Berbagai perangkat lunak bioinformatika telah dikembangkan untuk mendukung tahap pra-pemrosesan data sekuens DNA dan RNA. Beberapa perangkat lunak yang umum digunakan antara lain FastQC untuk *quality control*, Trimmomatic dan Cutadapt untuk *trimming adaptor* dan basa berkualitas rendah (Xiong, 2008).

Meskipun setiap perangkat lunak memiliki pendekatan dan fitur yang berbeda, prinsip dasar pra-pemrosesan tetap sama, yaitu meningkatkan kualitas data sebelum analisis lanjutan. Pemilihan perangkat lunak dan parameter analisis sebaiknya disesuaikan dengan jenis data, teknologi sekuensing, serta tujuan penelitian.

Analisis Sekuens DNA

Analisis sekuens DNA bertujuan untuk mendapatkan informasi biologis dari data genomik yang telah melalui tahap pra-pemrosesan. Pada tahap ini, sekuens DNA dianalisis untuk mengidentifikasi struktur genom, lokasi gen, variasi genetik, serta hubungan evolusioner antar organisme. Pendekatan analisis sekuens DNA sangat bergantung pada tujuan penelitian,

ketersediaan genom referensi, dan karakteristik data sekuensing yang digunakan.

Secara umum, analisis sekuens DNA mencakup beberapa tahapan utama, yaitu pemetaan sekuens terhadap referensi (*alignment*), perakitan genom (*assembly*), dan anotasi genom. Tahapan-tahapan ini saling berkaitan dan membentuk alur analisis genomik yang utuh (Mount, 2004; Brown, 2016).

Alignment Sekuens DNA (Gambaran Konseptual)

Alignment sekuens DNA merupakan proses mencocokkan sekuens hasil sekuensing dengan sekuens referensi atau dengan sekuens lain untuk mengidentifikasi kesamaan dan perbedaan urutan nukleotida. Tujuan utama *alignment* adalah menentukan posisi sekuens dalam genom serta mengidentifikasi variasi genetik, seperti substitusi, insersi, dan delesi (Durbin *et al.*, 1998; Altschul *et al.*, 1990).

Dalam analisis sekuens DNA, *alignment* berperan sebagai langkah awal untuk berbagai analisis lanjutan, termasuk deteksi mutasi, analisis variasi genetik, dan anotasi genom. Secara konseptual, *alignment* dapat dibedakan menjadi *alignment* global dan *alignment* lokal, tergantung pada cakupan wilayah sekuens yang dibandingkan.

Meskipun *alignment* merupakan komponen penting dalam analisis DNA, pembahasan teknis dan algoritmik mengenai *alignment* sekuens akan dibahas secara lebih mendalam pada chapter tersendiri. Pada chapter ini, *alignment* dipahami sebagai bagian dari *workflow* analisis genom yang berfungsi sebagai jembatan menuju analisis struktural dan fungsional genom.

Assembly Genom

Assembly genom adalah proses menyusun fragmen-fragmen sekuens DNA menjadi urutan genom yang lebih panjang dan runtut. Proses ini sangat penting terutama ketika genom referensi belum tersedia atau ketika tujuan penelitian adalah merekonstruksi genom organisme secara utuh.

Terdapat dua pendekatan utama dalam *assembly* genom, yaitu *de novo assembly* dan *reference-based assembly*. De novo assembly dilakukan tanpa menggunakan genom referensi dan mengandalkan tumpang tindih antar sekuens untuk membentuk kontig dan *scaffold*. Pendekatan ini sering digunakan pada organisme non-model atau genom yang belum terkarakterisasi dengan baik.

Sebaliknya, *reference-based assembly* menggunakan genom referensi sebagai panduan untuk menyusun sekuens hasil sekuensing. Pendekatan ini umumnya lebih efisien dan akurat jika genom referensi tersedia dan memiliki kemiripan tinggi dengan organisme yang dianalisis. Pemilihan metode *assembly* sangat dipengaruhi oleh tujuan analisis, kualitas data, serta ketersediaan referensi genom.

Analisis Variasi Genetik

Salah satu aplikasi penting analisis sekuens DNA adalah identifikasi variasi genetik. Variasi genetik mencakup perubahan urutan nukleotida yang dapat terjadi antar individu, populasi, atau spesies. Variasi ini dapat berupa *single nucleotide polymorphism (SNP)*, insersi, delesi, maupun variasi struktural yang lebih kompleks.

Analisis variasi genetik banyak digunakan dalam berbagai bidang, seperti genetika populasi, pemuliaan tanaman,

epidemiologi molekuler, dan studi penyakit genetik. Dengan membandingkan sekuens DNA terhadap referensi, peneliti dapat mengidentifikasi variasi yang berpotensi memengaruhi fungsi gen atau fenotipe organisme.

Interpretasi variasi genetik memerlukan kehati-hatian, karena tidak semua variasi memiliki dampak biologis yang signifikan. Oleh karena itu, hasil analisis variasi genetik perlu dikaitkan dengan informasi fungsional dan konteks biologis yang relevan.

Anotasi Genom

Anotasi genom merupakan tahap penting dalam analisis sekuens DNA yang bertujuan untuk mengidentifikasi dan memberi makna biologis pada elemen-elemen dalam genom. Anotasi genom mencakup penentuan lokasi gen, struktur gen (ekson dan intron), serta prediksi fungsi gen berdasarkan kemiripan dengan sekuens yang telah diketahui (Xiong, 2008).

Secara umum, anotasi genom dapat dibedakan menjadi anotasi struktural dan anotasi fungsional. Anotasi struktural berfokus pada identifikasi elemen fisik dalam genom, sedangkan anotasi fungsional bertujuan untuk memprediksi peran biologis gen atau produk gen tersebut. Proses anotasi sering kali melibatkan pemanfaatan basis data biologis dan perbandingan sekuens dengan database referensi. Tanpa anotasi yang memadai, data genom hanya berupa rangkaian nukleotida tanpa makna biologis yang jelas. Oleh karena itu, anotasi genom merupakan tahap krusial untuk menjembatani analisis komputasional dengan pemahaman biologis.

Integrasi Hasil Analisis Sekuens DNA

Tahap akhir analisis sekuens DNA adalah integrasi berbagai hasil analisis untuk memperoleh gambaran genom yang komprehensif. Hasil *alignment*, *assembly*, analisis variasi genetik, dan anotasi genom perlu dipadukan agar dapat menjawab pertanyaan penelitian secara menyeluruh.

Integrasi hasil analisis juga melibatkan interpretasi biologis yang mempertimbangkan konteks organisme, tujuan penelitian, dan keterbatasan data. Dengan pendekatan yang terintegrasi, analisis sekuens DNA tidak hanya menghasilkan data teknis, tetapi juga informasi biologis yang bermakna dan dapat dimanfaatkan untuk penelitian lanjutan.

Analisis Sekuens RNA

Analisis sekuens RNA bertujuan untuk memahami pola ekspresi gen dan mekanisme regulasi genetik dalam suatu kondisi biologis tertentu. Berbeda dengan analisis sekuens DNA yang bersifat relatif statis, analisis sekuens RNA bersifat dinamis karena mencerminkan respons sel terhadap lingkungan, tahap perkembangan, atau perlakuan eksperimen. Oleh karena itu, analisis sekuens RNA memainkan peran penting dalam studi transkriptomik dan biologi sistem.

Pendekatan yang paling umum digunakan dalam analisis sekuens RNA adalah RNA *sequencing* (RNA-seq). Melalui RNA-seq, peneliti dapat mengidentifikasi gen yang diekspresikan, mengukur tingkat ekspresi gen, serta mendeteksi perubahan ekspresi antar kondisi biologis yang berbeda.

Konsep Dasar RNA-Seq

RNA-seq merupakan teknik berbasis sekuensing yang digunakan untuk menganalisis transkriptom, yaitu kumpulan seluruh molekul RNA yang diekspresikan dalam suatu sel atau jaringan pada waktu tertentu. Data RNA-seq dihasilkan dari proses sekuensing molekul RNA yang telah dikonversi menjadi *complementary* DNA (cDNA) (Wang *et al.*, 2009; Stark *et al.*, 2019).

Dalam *workflow* RNA-seq, data mentah hasil sekuensing terlebih dahulu melalui tahap pra-pemrosesan untuk memastikan kualitas data. Selanjutnya, sekuens RNA dipetakan ke genom atau transkriptom referensi untuk mengidentifikasi asal sekuens dan menghitung jumlah *read* yang merepresentasikan tingkat ekspresi gen.

Keunggulan utama RNA-seq dibandingkan metode konvensional adalah kemampuannya untuk mendeteksi ekspresi gen secara komprehensif tanpa memerlukan informasi gen sebelumnya. Hal ini memungkinkan identifikasi gen baru, isoform transkrip, dan perubahan ekspresi gen yang kompleks.

Pemetaan Sekuens RNA dan Quantification

Tahap pemetaan (*mapping*) dalam analisis RNA-seq bertujuan untuk menentukan lokasi sekuens RNA pada genom atau transkriptom referensi. Proses ini berbeda dengan *alignment* DNA karena sekuens RNA dapat mengalami *splicing*, sehingga pemetaan harus mampu mengenali sambungan antar ekson.

Setelah pemetaan, dilakukan proses kuantifikasi untuk menghitung jumlah *read* yang terasosiasi dengan setiap gen atau transkrip. Jumlah *read* ini digunakan sebagai indikator tingkat ekspresi gen. Namun, karena panjang gen dan kedalaman sekuensing dapat berbeda antar sampel, data ekspresi perlu

dinormalisasi agar dapat dibandingkan secara adil. Hasil kuantifikasi biasanya disajikan dalam bentuk matriks ekspresi gen, yang menjadi dasar untuk analisis lanjutan, seperti analisis ekspresi gen diferensial dan visualisasi pola ekspresi.

Analisis Ekspresi Gen Diferensial

Analisis ekspresi gen diferensial bertujuan untuk mengidentifikasi gen-gen yang menunjukkan perbedaan tingkat ekspresi secara signifikan antara dua atau lebih kondisi biologis. Contohnya adalah perbandingan antara sampel kontrol dan perlakuan, atau antara kondisi sehat dan sakit (Love *et al.*, 2014; Conesa *et al.*, 2016).

Dalam analisis ini, pendekatan statistik digunakan untuk menentukan apakah perbedaan ekspresi gen bersifat signifikan secara biologis dan bukan akibat variasi teknis. Gen yang menunjukkan peningkatan ekspresi disebut sebagai gen terinduksi, sedangkan gen yang mengalami penurunan ekspresi disebut sebagai gen teredam.

Hasil analisis ekspresi gen diferensial sering divisualisasikan dalam bentuk *heatmap* dan volcano plot untuk memudahkan interpretasi. Visualisasi ini membantu peneliti dalam mengidentifikasi pola ekspresi gen dan memilih gen kandidat untuk analisis lanjutan.

Analisis Isoform dan Splicing Alternatif

Salah satu keunggulan analisis RNA-seq adalah kemampuannya untuk mendeteksi isoform transkrip dan peristiwa *splicing* alternatif. *Splicing* alternatif memungkinkan satu gen menghasilkan beberapa transkrip dengan struktur yang

berbeda, sehingga meningkatkan keragaman protein yang dihasilkan oleh organisme.

Analisis isoform bertujuan untuk mengidentifikasi dan mengkuantifikasi transkrip alternatif yang dihasilkan dari satu gen. Informasi ini sangat penting dalam studi regulasi gen dan dapat memberikan wawasan baru mengenai mekanisme biologis yang tidak terdeteksi pada tingkat DNA.

Kompleksitas analisis isoform menuntut kualitas data RNA-seq yang tinggi serta pendekatan analisis bioinformatika yang tepat. Oleh karena itu, analisis ini biasanya dilakukan setelah tahap analisis ekspresi gen dasar.

Interpretasi Biologis Data RNA

Tahap akhir analisis sekuens RNA adalah interpretasi biologis hasil analisis. Pada tahap ini, hasil analisis ekspresi gen dan isoform dikaitkan dengan fungsi biologis gen, jalur metabolik, dan respons seluler. Interpretasi yang tepat memerlukan pemahaman konteks biologis, desain eksperimen, dan keterbatasan data.

Perlu ditekankan bahwa perubahan ekspresi gen tidak selalu mencerminkan perubahan aktivitas protein secara langsung. Oleh karena itu, hasil analisis RNA-seq sebaiknya dipadukan dengan informasi lain, seperti data proteomik atau hasil eksperimen laboratorium, untuk memperoleh kesimpulan yang lebih komprehensif.

Workflow Analisis Sekuens DNA dan RNA

Workflow analisis sekuens DNA dan RNA menggambarkan rangkaian langkah sistematis dari data mentah hingga interpretasi hasil (Köster & Rahmann, 2018; Grüning *et al.*, 2018).

Workflow ini membantu peneliti memahami keterkaitan antar tahap analisis dan memastikan konsistensi proses.

Workflow umumnya mencakup tahap kontrol kualitas, preprocessing data, analisis utama, dan visualisasi hasil. Dalam praktik bioinformatika modern, workflow sering diotomatisasi menggunakan *pipeline* komputasional untuk meningkatkan efisiensi dan reprodusibilitas.

Pemahaman *workflow* sangat penting bagi pemula agar tidak hanya berfokus pada penggunaan perangkat lunak, tetapi juga memahami alur logika analisis secara keseluruhan.

Interpretasi dan Visualisasi Hasil Analisis

Interpretasi dan visualisasi merupakan tahap akhir yang menentukan nilai biologis dari analisis sekuens DNA dan RNA (Conesa *et al.*, 2016; Mount, 2004). Data hasil analisis bioinformatika umumnya bersifat kompleks dan sulit dipahami tanpa visualisasi yang tepat.

Visualisasi seperti *heatmap*, volcano plot, dan *genome browser* membantu menyederhanakan data kompleks menjadi representasi yang mudah dipahami. Melalui visualisasi, pola dan hubungan biologis dapat dikenali dengan lebih jelas. Interpretasi hasil harus selalu mempertimbangkan konteks biologis, kualitas data, dan keterbatasan metode analisis. Dengan pendekatan interpretasi yang kritis dan visualisasi yang tepat, analisis sekuens DNA dan RNA dapat menghasilkan wawasan biologis yang bermakna dan dapat diandalkan.

DAFTAR PUSTAKA

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local *alignment* search tool. *Journal of Molecular Biology, 215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Brown, T. A. (2016). *Genomes* (4th ed.). Garland Science.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., … Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology, 17*, 13. https://doi.org/10.1186/s13059-016-0881-8

Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics, 17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., & Köster, J. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature Methods, 15*(7), 475–476. https://doi.org/10.1038/s41592-018-0046-7

Köster, J., & Rahmann, S. (2018). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics, 34*(20), 3600. https://doi.org/10.1093/bioinformatics/bty350

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*, 550. https://doi.org/10.1186/s13059-014-0550-8

Mardis, E. R. (2017). DNA sequencing technologies: 2006–2016. *Nature Protocols, 12*(2), 213–218. https://doi.org/10.1038/nprot.2016.182

Mount, D. W. (2004). *Bioinformatics: Sequence and genome analysis* (2nd ed.). Cold Spring Harbor Laboratory Press.

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: The teenage years. *Nature Reviews Genetics, 20*(11), 631–656. https://doi.org/10.1038/s41576-019-0150-2

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics, 10*(1), 57–63. https://doi.org/10.1038/nrg2484

Xiong, J. (2008). *Essential bioinformatics*. Cambridge University Press.

BAB 5
Sequence Alignment: Konsep dan Algoritme

Rini Hafzari

Sequence Alignment merupakan salah satu bagian fundamental dalam ilmu bioinformatika dan biologi molekuler komputasi yang digunakan untuk membandingkan dua atau lebih urutan DNA, RNA, maupun protein (Ren *et al.*, 2018). Melalui proses *alignment*, posisi-posisi homolog antar sekuens dapat diidentifikasi, sehingga memungkinkan dilakukan analisis kesamaan struktural, fungsional maupun evolusioner. Dalam era biologi modern yang didukung dengan melimpahnya data sekuens hasil dari sekuensing, sekuens *alignment* mejadi bagian utama dalam interpretasi data biomolekuler (Sabonsolin & Lao, 2026).

Pada penggunaanya, hasil sekuens *alignment* sering digunakan sebagai dasar pengambilan keputusan ilmiah, misalnya dalam penentuan homologi gen, analisis variasi genetik, identifikasi spesies berbasis DNA *barcoding*, hingga prediksi fungsi protein dan penemuan target obat. Oleh karena itu, dibutuhkan pemahaman yang baik tentang interpretasi hasil *alignment* agar tidak terjadi kesalahan dalam pembacaan hasil data biologis.

Pengertian Sekuens *Alignment*

Sequence Alignment adalah proses penyusunan dua atau lebih sekuens biologis, baik berupa sekuens DNA, RNA, ataupun protein ke dalam suatu representasi sejajar, sehingga residu-residu yang bersesuaian secara evolusioner berada pada posisi kolom yang sama (Sabonsolin & Lao, 2026). Keselarasan ini memungkinkan perbandingan langsung antar sekuens untuk mengidentifikasi kesamaan, perbedaan, dan pola konservasi molekuler.

Secara konseptual, *Sequence Alignment* merepresentasikan upaya untuk merekonstruksi sejarah evolusi molekul dengan asumsi bahwa perbedaan antar sekuens muncul akibat peristiwa mutasi yang terjadi secara bertahap, seperti substitusi nukleotida atau asam amino, insersi, dan delesi (Rosenberg, 2009). Oleh karena itu, *alignment* tidak hanya berfungsi sebagai alat komputasional, tetapi juga sebagai pendekatan biologis untuk memahami dinamika perubahan molekul sepanjang waktu evolusi.

Dalam praktik bioinformatika, *alignment* biasanya direpresentasikan dalam bentuk barisan karakter yang disusun sejajar, dengan penambahan simbol *gap* ("-") untuk merepresentasikan insersi atau delesi yang terjadi selama evolusi. Penempatan *gap* ini dilakukan secara sistematis berdasarkan aturan skoring tertentu agar keselarasan yang dihasilkan tetap mencerminkan kemungkinan biologis yang benar. Dengan demikian, kualitas suatu *alignment* sangat bergantung pada keseimbangan antara kesesuaian sekuens dan jumlah *gap* yang digunakan (Redelings *et al.*, 2024).

Konsep Dasar Sekuens *Alignment*

Identitas, similaritas dan homologi

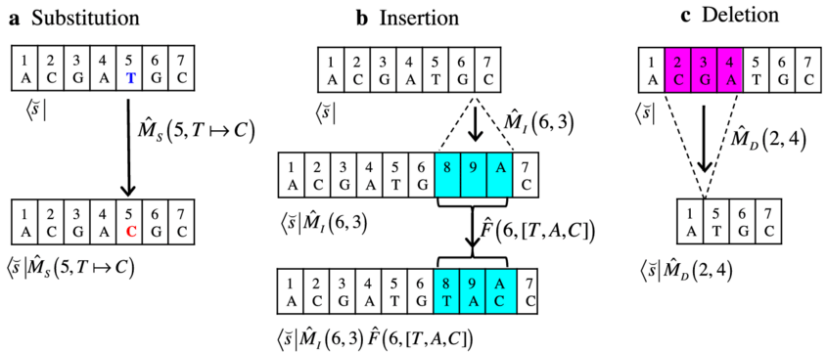
Identitas mengacu pada proporsi posisi dalam *alignment* yang memiliki karakter identik. Similaritas memperhitungkan kemiripan sifat fisikokimia residu, khususnya pada sekuens protein, sehingga dua asam amino berbeda dapat dianggap mirip jika memiliki karakter kimia yang sama. Homologi, merupakan konsep evolusioner yang menyatakan adanya nenek moyang bersama dan tidak dinyatakan dalam bentuk persentase.

Pemisahan antara ketiga istilah ini penting karena kesalahan konseptual, seperti menyatakan "persentase homologi", masih sering ditemukan dalam literatur ilmiah. *Alignment* hanya menyediakan bukti kuantitatif berupa identitas dan similaritas, sedangkan kesimpulan homologi memerlukan interpretasi biologis lebih lanjut yang mempertimbangkan hubungan filogenetik, panjang dan pola kesamaan sekuens, serta dukungan data biologis lain selain hasil *alignment* (Kanduc, 2012; Phillips, 2006).

Mutasi, gap dan Penalti

Dalam *Sequence Alignment*, perbedaan antar sekuens biologis dimodelkan sebagai hasil dari peristiwa mutasi yang terjadi selama proses evolusi. Mutasi yang paling umum meliputi substitusi, insersi, dan delesi (Gambar 5.1). Mutasi substitusi direpresentasikan sebagai penggantian satu nukleotida atau asam amino dengan karakter lain pada posisi yang sama, sedangkan mutasi insersi dan delesi direpresentasikan dalam *alignment* sebagai *gap*. *gap* digunakan untuk menjaga keselarasan posisi homolog antar sekuens, sehingga perubahan

panjang akibat mutasi dapat dimodelkan secara eksplisit (Goonesekere & Lee, 2004; Phillips, 2006).



Sumber: Ezawa (2016)

Gambar 5.1. Subtitusi, Insersi dan Delesi

Penggunaan *gap* dalam *alignment* tidak dapat dilakukan secara bebas karena dapat menghasilkan keselarasan yang tidak realistis secara biologis. Tanpa pembatasan, algoritme *alignment* cenderung menambahkan banyak *gap* untuk meningkatkan kecocokan karakter, meskipun hal tersebut tidak mencerminkan peristiwa evolusioner yang sebenarnya. Oleh karena itu, setiap penambahan *gap* dikenai penalti (*gap penalty*) yang berfungsi sebagai pengendali agar jumlah dan panjang *gap* tetap masuk kriteria (Goonesekere & Lee, 2004).

Secara umum, terdapat dua jenis penalti *gap* yang digunakan dalam algoritme *Sequence Alignment*, yaitu penalti *gap* linear dan penalti *gap* affine. Penalti *gap* linear memberikan nilai penalti yang sama untuk setiap posisi gap, sehingga penambahan *gap* meningkat secara proporsional dengan panjang gap. Meskipun sederhana dan mudah diterapkan, pendekatan ini kurang merepresentasikan realitas biologis

karena mengasumsikan bahwa setiap posisi *gap* memiliki dampak evolusioner yang sama.

Sebaliknya, penalti *gap affine* membedakan antara pembukaan *gap* (*gap opening penalty*) dan perpanjangan *gap* (*gap extension penalty*). Dalam model ini, pembukaan *gap* diberi penalti yang lebih besar dibandingkan perpanjangannya. Pendekatan ini lebih realistis secara biologis karena peristiwa insersi atau delesi umumnya terjadi sebagai satu kejadian yang melibatkan beberapa residu sekaligus, bukan sebagai banyak kejadian terpisah. Dengan demikian, penalti *affine* lebih mampu merepresentasikan mekanisme evolusi yang sebenarnya dan banyak digunakan dalam algoritme *alignment* modern (Marco-Sola *et al.*, 2023).

Pemilihan jenis penalti *gap* sangat memengaruhi hasil akhir *alignment*. Penalti *gap* yang terlalu rendah dapat menghasilkan *alignment* dengan banyak celah yang tidak biologis, sedangkan penalti yang terlalu tinggi dapat menghambat penyelarasan daerah homolog yang sebenarnya mengalami insersi atau delesi. Oleh karena itu, penentuan parameter *gap* harus disesuaikan dengan jenis sekuens, tujuan analisis, dan konteks biologis dari data yang dianalisis.

Sistem skoring sekuens alignment

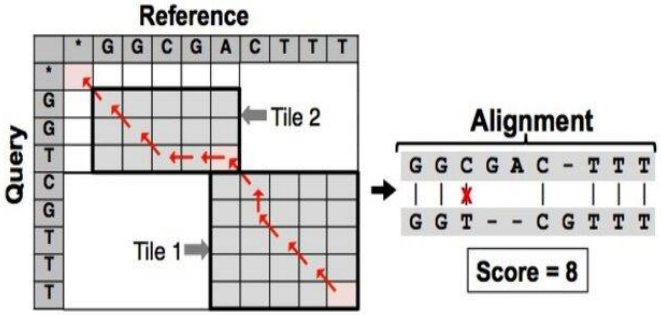
Sistem skoring merupakan inti dari seluruh algoritme *Sequence Alignment* karena menentukan bagaimana kesesuaian dan perbedaan antar sekuens dinilai secara kuantitatif. Dalam konteks *alignment*, skor tidak hanya berfungsi sebagai nilai numerik, tetapi juga sebagai representasi dari kemungkinan biologis bahwa dua residu pada posisi tertentu berasal dari peristiwa evolusioner yang sama. Dengan kata lain, sistem

skoring menjadi penghubung antara perhitungan komputasional dan interpretasi biologis.

Secara umum, skor *alignment* merupakan hasil penjumlahan dari tiga komponen utama, yaitu kecocokan (*match*), ketidakcocokan (*mismatch*), dan *gap*. Kecocokan diberikan nilai positif karena menunjukkan kesamaan residu, sedangkan ketidakcocokan dan *gap* umumnya diberi nilai negatif sebagai penalti. Kombinasi ketiga komponen ini menentukan *alignment* dengan skor total tertinggi yang dianggap paling optimal oleh algoritme (Liu & Schmidt, 2015; Reali *et al.*, 2018).

Skoring pada sekuens DNA

Pada sekuens DNA, sistem skoring relatif sederhana karena hanya melibatkan empat jenis nukleotida, yaitu adenin (A), timin (T), sitosin (C), dan guanin (G). Umumnya, kecocokan antar nukleotida identik diberi skor positif, sedangkan ketidakcocokan diberi skor negatif (Gambar 5.2). Penalti *gap* diterapkan untuk merepresentasikan insersi atau delesi yang terjadi selama evolusi. Kesederhanaan sistem skoring DNA membuat *alignment* nukleotida relatif mudah diinterpretasikan. Namun, pendekatan ini memiliki keterbatasan karena tidak semua substitusi nukleotida memiliki probabilitas yang sama. Sebagai contoh, substitusi transisi lebih sering terjadi dibandingkan dengan substitusi transversi, tetapi perbedaan ini tidak selalu tercermin dalam sistem skoring sederhana (Edi *et al.*, 2025).

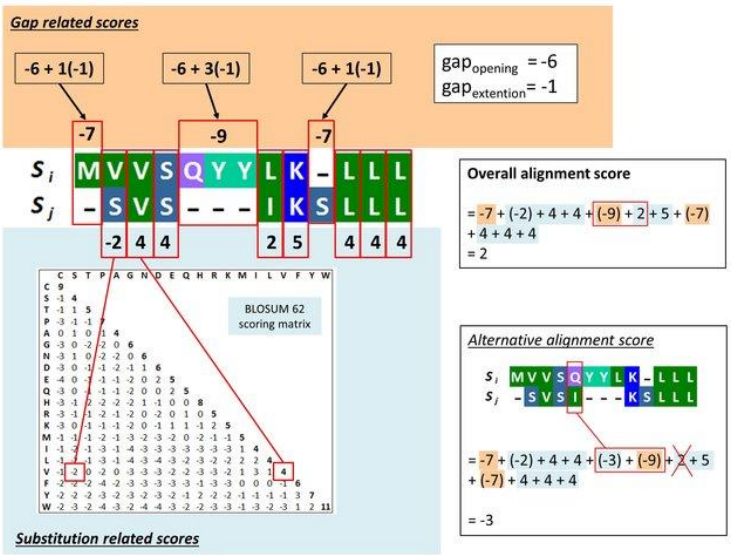


Sumber: Turakhia *et al.* (2017)

Gambar 5.2. Skema perhitungan skoring pada sekuens DNA

Skoring pada sekuens protein

Berbeda dengan DNA, *alignment* protein memerlukan skoring yang lebih kompleks karena melibatkan 20 jenis asam amino dengan sifat fisikokimia yang beragam. Dalam *alignment* protein, dua asam amino yang berbeda masih dapat dianggap serupa apabila memiliki sifat kimia yang setara, seperti muatan, ukuran, atau polaritas. Oleh karena itu, sistem skoring protein dirancang untuk merepresentasikan tingkat kesamaan tersebut. Sistem skoring protein tidak lagi menggunakan pendekatan biner antara *match* dan *mismatch*, tetapi memanfaatkan matriks substitusi yang memberikan skor berbeda untuk setiap pasangan asam amino (Gambar 5.3). Matriks ini memungkinkan algoritme *alignment* untuk mengenali substitusi konservatif yang secara biologis lebih mungkin terjadi dibandingkan substitusi non-konservatif (Ranwez & Chantret, 2020).



Sumber: Ranwez & Chantret (2020)

Gambar 5.3. Skema skoring pada sekuens protein

Matriks substitusi Point Accepted Mutation (PAM) dan Blocks Substitution Matrix (BLOSUM)

Matriks PAM (*Point Accepted Mutation*) dan BLOSUM (*Blocks Substitution Matrix*) merupakan dua matriks substitusi yang paling umum digunakan dalam *alignment* protein. Kedua matriks ini dibangun berdasarkan data empiris dari sekuens protein yang mengalami perubahan evolusioner. Matriks PAM dikembangkan dengan memodelkan laju mutasi berdasarkan jarak evolusioner tertentu, sehingga cocok digunakan untuk *alignment* protein yang memiliki hubungan evolusioner yang relatif dekat atau jauh (tergantung pada nilai PAM yang dipilih). Sebaliknya, matriks BLOSUM dibangun berdasarkan blok sekuens konservatif dan lebih sering digunakan dalam praktik karena dianggap lebih stabil untuk berbagai tingkat kesamaan sekuens (Jia & Jernigan, 2021).

Pemilihan matriks substitusi yang tepat sangat penting. Hal ini dikarenakan matriks yang terlalu permisif dapat menghasilkan *alignment* yang tampak baik secara matematis, tetapi tidak mencerminkan hubungan biologis yang sebenarnya. Sebaliknya, matriks yang terlalu ketat dapat gagal mendeteksi kesamaan yang relevan secara fungsional.

Algoritme Sekuens Alignment

Pendekatan *Dynamic programming*

Dalam konteks *Sequence Alignment*, pendekatan ini digunakan untuk menentukan keselarasan terbaik antara dua sekuens dengan mempertimbangkan seluruh kemungkinan penjajaran karakter secara sistematis, sehingga solusi yang diperoleh bersifat optimal berdasarkan sistem skoring yang ditetapkan. Prinsip utama *dynamic programming* adalah bahwa solusi optimal dari suatu masalah dapat dibangun dari solusi optimal submasalahnya, sehingga perhitungan tidak perlu diulang untuk kasus yang sama (Nalbantoğlu Ö, 2014).

Pada *Sequence Alignment*, *dynamic programming* diimplementasikan dalam bentuk matriks dua dimensi, dengan baris dan kolom merepresentasikan posisi karakter pada masing-masing sekuens yang dibandingkan. Setiap sel dalam matriks menyimpan skor terbaik yang dapat dicapai hingga posisi tersebut. Hal ini didasarkan pada tiga kemungkinan jalur utama, yaitu penyelarasan dua karakter (*match* atau *mismatch*), penambahan *gap* pada sekuens pertama, ataupun penambahan *gap* pada sekuens kedua. Dengan cara ini, algoritme secara eksplisit mengevaluasi

seluruh jalur *alignment* yang mungkin dan memilih jalur dengan skor tertinggi sesuai aturan skoring.

Proses *dynamic programming* dalam *Sequence Alignment* umumnya terdiri atas tiga tahap utama. Tahap pertama adalah inisialisasi matriks, yaitu pemberian nilai awal pada baris dan kolom pertama yang merepresentasikan *alignment* sekuens dengan *gap* penuh. Nilai inisialisasi ini sangat bergantung pada jenis *alignment* yang digunakan, apakah global atau lokal. Tahap kedua adalah pengisian matriks (*matrix filling*). Tahap ini dilakukan perhitungan setiap sel secara berurutan menggunakan relasi rekursif yang membandingkan skor dari ketiga kemungkinan jalur *alignment*. Tahap ini merupakan inti dari *dynamic programming*, karena di sinilah seluruh alternatif *alignment* dievaluasi secara kuantitatif. Tahap ketiga adalah *traceback*, yaitu proses penelusuran kembali jalur dari sel dengan skor optimal untuk merekonstruksi *alignment* akhir yang dihasilkan oleh algoritme (Nalbantoğlu Ö, 2014).

Keunggulan utama pendekatan *dynamic programming* adalah kemampuannya untuk menjamin solusi *alignment* yang optimal secara matematis. Namun, keunggulan ini disertai dengan keterbatasan komputasional, karena kebutuhan waktu dan memori meningkat secara kuadratik seiring bertambahnya panjang sekuens. Oleh karena itu, meskipun *dynamic programming* menjadi dasar teoritis algoritme *alignment* klasik seperti global dan lokal *alignment*, pendekatan ini kurang efisien untuk analisis basis data sekuens berskala besar. Keterbatasan inilah yang kemudian mendorong pengembangan algoritme heuristik

yang lebih cepat, meskipun tidak selalu menjamin solusi optimal (Nalbantoğlu Ö, 2014).

Algoritme Global *Alignment* (Needlemen-Wunsch)

Algoritme Needleman–Wunsch merupakan algoritme klasik dalam *Sequence Alignment* yang digunakan untuk melakukan global *alignment*, yaitu penyelarasan dua sekuens secara menyeluruh dari awal hingga akhir (Muhamad *et al.*, 2018). Algoritme ini dikembangkan dengan pendekatan *dynamic programming* dan dirancang untuk menemukan *alignment* dengan skor total maksimum berdasarkan sistem skoring yang telah ditentukan. Global *alignment* sangat sesuai digunakan ketika dua sekuens yang dibandingkan memiliki panjang relatif sama dan diperkirakan memiliki tingkat kesamaan yang tinggi, seperti gen homolog dari organisme yang berkerabat dekat.

Secara konseptual, algoritme Needleman–Wunsch bekerja dengan membangun sebuah matriks dua dimensi. Setiap baris merepresentasikan posisi karakter pada sekuens pertama dan kolom merepresentasikan posisi karakter pada sekuens kedua. Setiap sel dalam matriks menyimpan skor terbaik yang dapat dicapai hingga posisi tersebut. Matriks ini memungkinkan algoritme untuk mengevaluasi seluruh kemungkinan penjajaran karakter secara sistematis dan terstruktur, sehingga tidak ada alternatif *alignment* yang terlewatkan (Muhamad *et al.*, 2018).

Tahap pertama dalam algoritme Needleman–Wunsch adalah inisialisasi matriks, dengan baris dan kolom pertama diisi dengan nilai kumulatif penalti *gap*. Inisialisasi ini

mencerminkan kondisi *alignment* ketika salah satu sekuens disejajarkan sepenuhnya dengan *gap* pada sekuens lainnya. Tahap ini sangat penting karena menentukan bahwa *alignment* akan dipaksakan mencakup seluruh panjang kedua sekuens, sesuai dengan konsep global *alignment*.

Tahap kedua adalah pengisian matriks (*matrix filling*). Pada tahap ini, skor setiap sel dihitung berdasarkan nilai maksimum dari tiga kemungkinan jalur, yaitu penyelarasan dua karakter (*match* atau *mismatch*), penambahan *gap* pada sekuens pertama, atau penambahan *gap* pada sekuens kedua. Setiap pilihan jalur mencerminkan kemungkinan peristiwa evolusioner yang berbeda, dan algoritme akan memilih jalur yang menghasilkan skor tertinggi sesuai sistem skoring yang digunakan. Dengan pendekatan ini, algoritme secara implisit membandingkan seluruh alternatif *alignment* dan menyimpan solusi optimal pada setiap submasalah.

Tahap terakhir adalah *traceback*, yaitu proses penelusuran kembali jalur skor optimal dari sudut kanan bawah matriks menuju sudut kiri atas. Jalur *traceback* ini merepresentasikan *alignment* akhir antara kedua sekuens, termasuk posisi kecocokan, ketidakcocokan, dan *gap*. Hasil *traceback* inilah yang kemudian ditampilkan sebagai *alignment* global yang optimal.

Algoritme Local *Alignment* (Smith-Waterman)

Algoritme Smith–Waterman merupakan modifikasi dari pendekatan global *alignment* yang dirancang untuk menemukan keselarasan lokal terbaik antara dua sekuens. Algoritme ini sangat berguna ketika hanya sebagian kecil

sekuens yang homolog, seperti pada identifikasi domain protein atau motif konservatif (Zhou & Wang, 2014).

Perbedaan utama dengan Needleman–Wunsch terletak pada penerapan nilai nol sebagai batas bawah skor. Jika perhitungan skor menghasilkan nilai negatif, maka nilai tersebut diatur menjadi nol, sehingga *alignment* hanya diperluas pada wilayah yang memberikan kontribusi positif terhadap skor total.

Secara biologis, algoritme Smith–Waterman lebih realistis untuk banyak kasus karena tidak memaksakan keselarasan pada seluruh panjang sekuens. Namun, seperti global *alignment*, algoritme ini tetap memiliki keterbatasan komputasi untuk analisis skala besar.

Algoritme Heuristik dan Pendekatan Aproksimasi Sekuens *Alignment*

Algoritme heuristik dikembangkan sebagai salah satu solusi untuk mengatasi keterbatasan algoritme sekuens *alignment* berbasis *dynamic programming* yang memiliki kompleksitas waktu dan memori tinggi. Walaupun algoritme seperti Needlemen-Wunsch dan Smith-Waterman juga merupakan algoritme *alignment* optimal secara matematis, namun pendekatan tersebut menjadi tidak praktis ketika diterapkan pada basis data sekuens yang sangat besar. Contohnya seperti ketika diterapkan pada basis data genom atau protein berskala global. Dalam konteks ini, algoritme heuristik menawarkan solusi antara kecepatan komputasi dan kualitas biologis hasil *alignment* (Chao *et al.*, 2022).

Pendekatan heuristik tidak bertujuan untuk mengevaluasi seluruh kemungkinan *alignment* secara menyeluruh, melainkan untuk mengidentifikasi kandidat *alignment* yang paling baik berdasarkan pola-pola kesamaan awal yang sederhana. Prinsip dasarnya adalah bahwa sekuens memiliki hubungan biologis atau evolusioner yang cenderung berbagi potongan sekuens pendek yang identik atau sangat mirip. Potongan sekuens pendek ini dikenal sebagai kata kunci atau *seed*. Dengan memfokuskan pencarian pada *seed* terlebih dahulu, algoritme heuristik dapat mengurangi ruang pencarian yang harus dievaluasi.

Proses kerja algoritme heuristik umumnya diawali dengan pencarian *seed*, yaitu potongan sekuens pendek dengan panjang tertentu yang cocok atau hampir cocok antara sekuens kueri dan sekuens dalam basis data. *Seed* ini berfungsi sebagai titik awal yang menunjukkan adanya potensi kesamaan. Setelah *seed* ditemukan, algoritme kemudian melakukan proses perluasan (ekstensi) ke arah kiri dan kanan dari *seed* tersebut untuk membangun *alignment* yang lebih panjang. Proses perluasan ini biasanya dihentikan ketika skor *alignment* mulai menurun di bawah ambang batas tertentu, sehingga hanya daerah dengan kontribusi skor positif yang dipertahankan (Marçais *et al.*, 2018).

Keunggulan utama pendekatan heuristik terletak pada efisiensi waktu komputasi. Dengan membatasi pencarian hanya pada daerah yang mengandung *seed*, algoritme dapat melakukan pencarian *alignment* dalam waktu yang sangat singkat, bahkan pada basis data yang berisi jutaan sekuens. Hal ini menjadikan algoritme heuristik sangat populer dalam

praktik bioinformatika sehari-hari, khususnya untuk pencarian kesamaan sekuens secara cepat.

Namun demikian, pendekatan heuristik memiliki keterbatasan inheren. Hal ini dikarenakan pendekatan ini tidak menjamin ditemukannya *alignment* optimal secara global. Ada kemungkinan bahwa hubungan biologis yang nyata terlewatkan jika kesamaan awal antar sekuens tidak cukup kuat untuk membentuk *seed*. Oleh karena itu, hasil *alignment* dari algoritme heuristik harus dipahami sebagai pendekatan aproksimasi yang memberikan kandidat kesamaan paling mungkin, bukan sebagai representasi mutlak dari hubungan evolusioner.

Meskipun demikian, kualitas biologis hasil *alignment* heuristik umumnya sangat baik, terutama ketika dikombinasikan dengan sistem skoring yang tepat dan kriteria statistik seperti nilai E (*expectation value*). Dalam banyak aplikasi praktis, hasil yang diperoleh dari algoritme heuristik sudah cukup untuk keperluan identifikasi gen, anotasi protein, dan analisis hubungan evolusioner awal, sehingga pendekatan ini menjadi pilihan utama dalam analisis sekuens berskala besar.

Multiple Sequence *Alignment* (MSA)

Multiple *Sequence Alignment* merupakan perluasan dari *alignment* pasangan sekuens ke lebih dari dua sekuens. MSA sangat penting dalam analisis filogenetik dan identifikasi residu konservatif. Algoritme MSA umumnya menggunakan pendekatan progresif atau iteratif, yang masing-masing

memiliki kelebihan dan keterbatasan (Sabonsolin & Lao, 2026).

Pendekatan progresif menyusun *alignment* secara bertahap berdasarkan pohon panduan (*guide tree*), tetapi rentan terhadap propagasi kesalahan. Pendekatan iteratif berusaha memperbaiki *alignment* secara berulang untuk meningkatkan kualitas hasil akhir.

Tantangan dan Arah Perkembangan Sequence Alignment

Perkembangan teknologi sekuensing modern menghasilkan data biomolekuler dalam skala yang sangat besar dan kompleks, sehingga menimbulkan tantangan baru dalam *Sequence Alignment*, baik dari sisi komputasi maupun interpretasi biologis. Algoritme *alignment* konvensional berbasis *dynamic programming* menjadi kurang efisien untuk menangani volume data yang terus meningkat, sehingga diperlukan pendekatan yang lebih cepat dan hemat sumber daya melalui optimasi algoritmik dan pemanfaatan komputasi berperforma tinggi.

Selain itu, meningkatnya variasi biologis, seperti sekuens yang sangat divergen, variasi struktural genom, dan data non-linier, menuntut metode *alignment* yang lebih adaptif dan kontekstual. Hal ini dilakukan agar hasil yang diperoleh tetap akurat dan bermakna secara biologis. Dalam konteks ini, integrasi kecerdasan buatan dan *machine learning* mulai dikembangkan untuk meningkatkan efisiensi dan akurasi *alignment* dengan memanfaatkan pola-pola biologis yang dipelajari dari data dalam jumlah besar. Dengan demikian, arah perkembangan *Sequence Alignment* tidak hanya berfokus pada kecepatan dan akurasi komputasi, tetapi juga pada kemampuan algoritme untuk

menghasilkan keselarasan sekuens yang dapat diinterpretasikan secara ilmiah dan relevan dengan kompleksitas biologi modern.

DAFTAR PUSTAKA

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

Edi, E., Robet, R., & Harahap, N. (2025). Comparative analysis of DNA sequence alignment algorithms in SARS-CoV-2. *Sinkron: Jurnal dan Penelitian Teknik Informatika, 9*(4), 2318–2325. https://doi.org/10.33395/sinkron.v9i4.15323

Ezawa, K. (2016). General continuous-time Markov model of sequence evolution via insertions/deletions: Are alignment probabilities factorable? *BMC Bioinformatics, 17*, 304. https://doi.org/10.1186/s12859-016-1105-7

Goonesekere, N. C., & Lee, B. (2004). Frequency of gaps observed in a structurally aligned protein pair database suggests a simple gap penalty function. *Nucleic Acids Research, 32*(9), 2838–2843. https://doi.org/10.1093/nar/gkh610

Jia, K., & Jernigan, R. L. (2021). New amino acid substitution matrix brings sequence alignments into agreement with structure matches. *Proteins: Structure, Function, and Bioinformatics, 89*(6), 671–682. https://doi.org/10.1002/prot.26050

Kanduc, D. (2012). Homology, similarity, and identity in peptide epitope immunodefinition. *Journal of Peptide Science, 18*(8), 487–494. https://doi.org/10.1002/psc.2419

Liu, Y., & Schmidt, B. (2015). Pairwise DNA sequence alignment optimization. In J. Reinders & J. Jeffers (Eds.), *High performance parallelism pearls* (pp. 43–54). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-803819-2.00007-0

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S.

L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology, 14*(1), e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Marco-Sola, S., Eizenga, J. M., Guarracino, A., Paten, B., Garrison, E., & Moreto, M. (2023). Optimal gap-affine alignment in O(s) space. *Bioinformatics, 39*(2). https://doi.org/10.1093/bioinformatics/btad074

Muhamad, F. N., Ahmad, R. B., Asi, S. M., & Murad, M. N. (2018). Performance analysis of Needleman–Wunsch algorithm (global) and Smith–Waterman algorithm (local) in reducing search space and time for DNA sequence alignment. *Journal of Physics: Conference Series, 1019*(1), 012085. https://doi.org/10.1088/1742-6596/1019/1/012085

Nalbantoğlu, Ö. U. (2014). Dynamic programming. In *Methods in molecular biology* (Vol. 1079, pp. 3–27). https://doi.org/10.1007/978-1-62703-646-7_1

Phillips, A. J. (2006). Homology assessment and molecular sequence alignment. *Journal of Biomedical Informatics, 39*(1), 18–33. https://doi.org/10.1016/j.jbi.2005.11.005

Reali, G., Femminella, M., Nunzi, E., & Valocchi, D. (2018). Genomics as a service: A joint computing and networking perspective. *Computer Networks, 145*, 27–51. https://doi.org/10.1016/j.comnet.2018.08.005

Redelings, B. D., Holmes, I., Lunter, G., Pupko, T., & Anisimova, M. (2024). Insertions and deletions: Computational methods, evolutionary dynamics, and biological applications. *Molecular Biology and Evolution, 41*(9). https://doi.org/10.1093/molbev/msae177

Ren, J., Bai, X., Lu, Y. Y., Tang, K., Wang, Y., Reinert, G., & Sun, F.

(2018). Alignment-free sequence analysis and applications. *Annual Review of Biomedical Data Science, 1*, 93–114. https://doi.org/10.1146/annurev-biodatasci-080917-013431

Sabonsolin, J., & Lao, D. (2026). A comprehensive systematic literature review of multiple sequence alignment algorithms. *Discover Computing, 29*(1), 33. https://doi.org/10.1007/s10791-026-09911-3

Turakhia, Y., Zheng, K., Bejerano, G., & Dally, W. (2017). Darwin: A hardware-acceleration framework for genomic sequence alignment. *bioRxiv*. https://doi.org/10.1101/092171

Zhou, H., & Wang, H. (2014). Localized Smith–Waterman algorithm for fast and complete search of siRNA off-target sequences. *American Journal of Bioinformatics Research, 4*(1), 1–5. https://doi.org/10.5923/j.bioinformatics.20140401.01

BAB 6
Multiple Sequence Alignment (MSA) dan Analisis Evolusi

M. Khoiron Ferdiansyah

Landasan Teoretis *Multiple Sequence Alignment* (MSA)

Multiple Sequence Alignment (MSA) didefinisikan sebagai proses penyejajaran tiga atau lebih sekuens biologis (baik nukleotida maupun asam amino) yang memiliki hubungan homologi. Secara teoretis, MSA merupakan perluasan dari *pairwise alignment*, yang merupakan metode bioinformatika untuk menjajarkan dua sekuens biologis DNA, RNA, atau protein untuk mengidentifikasi wilayah kemiripan, perbedaan, serta hubungan fungsional, struktural, atau evolusioner di antara keduanya. Metode ini sering kali menggunakan sistem skor untuk mencocokkan karakter dan menghitung skor terbaik, tetapi dengan kompleksitas komputasi yang jauh lebih tinggi. Tujuan utama dari MSA dalam bioinformatika adalah untuk mengidentifikasi daerah konservasi (daerah yang tidak berubah), yang mencerminkan domain fungsional penting atau struktur sekunder protein yang dipertahankan selama proses evolusi.

Dalam konteks evolusi, MSA berfungsi sebagai rekonstruksi hubungan filogenetik. Penyejajaran yang akurat memungkinkan peneliti untuk membedakan antara mutasi titik (substitusi), insersi, dan delesi (indels). Celah (*gaps*) yang muncul dalam MSA

merepresentasikan peristiwa evolusioner: materi genetik ditambahkan atau dihilangkan dalam satu garis keturunan dibandingkan dengan yang lain.

Algoritme dan Pendekatan Komputasi dalam MSA

Kompleksitas MSA meningkat secara eksponensial seiring bertambahnya jumlah sekuens dan panjang sekuens (Gambar 6.1). Oleh karena itu, berbagai algoritme dikembangkan untuk menyeimbangkan antara akurasi dan efisiensi waktu:

- Metode Progresif: Algoritme ini (seperti yang digunakan pada ClustalW dan Clustal Omega) bekerja dengan cara membangun *guide tree* berdasarkan jarak genetik awal, kemudian menyejajarkan sekuens satu per satu mulai dari pasangan yang paling mirip.
- Metode Iteratif: Algoritme seperti MAFFT atau MUSCLE melakukan optimasi berulang terhadap hasil penyejajaran awal untuk memperbaiki kesalahan yang mungkin terjadi pada tahap progresif awal.
- Transformasi Fourier Cepat (FFT): Digunakan untuk mempercepat identifikasi homologi lokal pada sekuens yang sangat panjang melalui analisis korelasi numerik asam amino.

Gambar 6.1. Contoh Tampilan MSA

Analisis Evolusi Molekuler

Analisis evolusi berbasis MSA memungkinkan kita untuk memetakan hubungan kekerabatan antar spesies atau antar gen dalam satu genom (paralog).

Konservasi Evolusioner dan Arsitektur Genetik

Melalui MSA, peneliti dapat mengamati tingkat konservasi pada posisi residu tertentu. Residu yang terkonservasi secara absolut di seluruh taksonomi sering kali merupakan residu aktif yang bertanggung jawab atas aktivitas katalitik enzim atau stabilitas struktural. Sebaliknya, variasi pada daerah tertentu sering kali menunjukkan adaptasi evolusioner terhadap lingkungan yang berbeda.

Dalam studi genomik perbandingan, analisis evolusi juga mencakup pengamatan terhadap pengelompokan gen dalam operon. Evolusi sering kali mengatur gen-gen dengan fungsi

metabolisme yang berkaitan untuk berada dalam satu unit transkripsi yang sama (operon) guna meningkatkan efisiensi regulasi seluler. Perbedaan orientasi gen antar spesies (seperti arah transkripsi yang berlawanan atau searah) mencerminkan peristiwa rekombinasi genetik dan penataan ulang genom yang terjadi selama jutaan tahun.

Duplikasi Gen dan Divergensi Fungsi

Salah satu mekanisme evolusi yang paling signifikan adalah duplikasi gen. Setelah peristiwa duplikasi, satu salinan gen dapat mempertahankan fungsi aslinya, sementara salinan lainnya (paralog) bebas untuk mengalami mutasi dan mengembangkan fungsi baru atau spesifisitas substrat yang berbeda. Analisis MSA pada famili protein yang terdiversifikasi membantu ilmuwan memahami bagaimana satu enzim moyang berkembang menjadi berbagai varian dengan karakteristik kinetika yang berbeda, seperti perbedaan nilai konstanta Michaelis-Menten dan efisiensi katalitik.

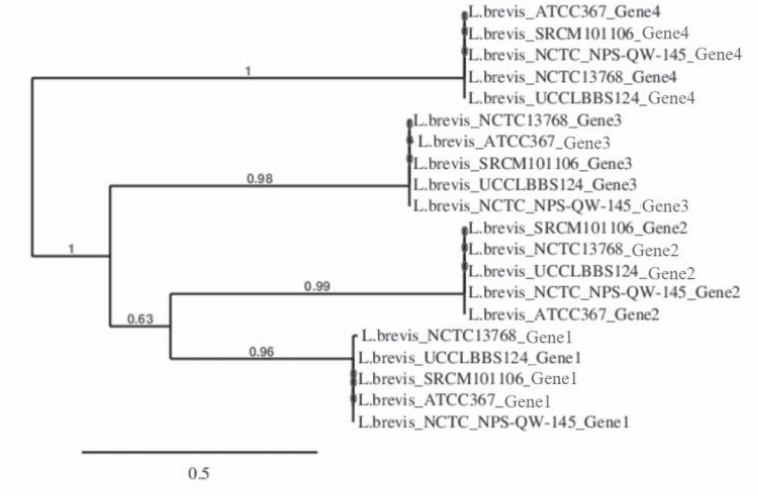
Studi Kasus: Implementasi MSA pada Bakteri Asam Laktat (BAL) dalam Produksi, Purifikasi, dan Karakterisasi Enzim Purine Nucleosidase dari *Levilactobacillus brevis* LABC170

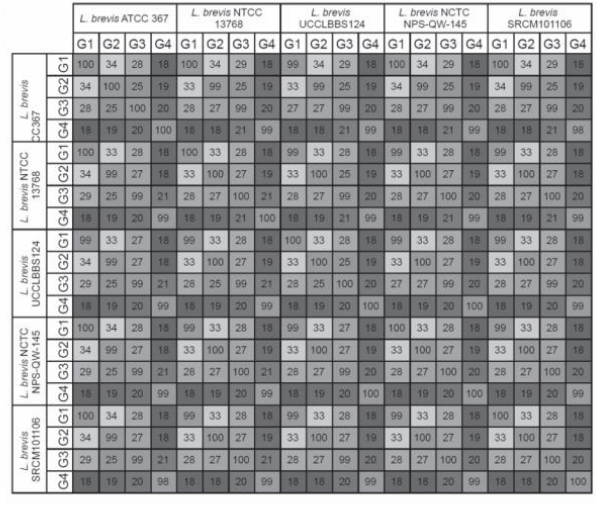
Purine nucleosidase (PNase) merupakan enzim yang berperan penting dalam metabolisme purin melalui reaksi hidrolisis nukleosida purin menjadi basa purin dan ribosa. Dalam konteks pengendalian hiperurisemia berbasis mikroba, bakteri asam laktat (BAL) menjadi kandidat yang menjanjikan karena aspek keamanan dan potensi aplikasinya pada pangan fungsional. Salah satu strain BAL yang menunjukkan aktivitas PNase tinggi adalah *Levilactobacillus brevis* LABC170.

Analisis genom *L. brevis* mengungkap keberadaan lebih dari satu gen putatif PNase. Keberadaan beberapa gen dengan anotasi serupa menimbulkan pertanyaan mengenai kontribusi masing-masing gen terhadap aktivitas enzimatik aktual. Oleh karena itu, pendekatan Multiple *Alignment* Sequences (MAS) digunakan sebagai langkah awal untuk mengevaluasi konservasi sekuens, variasi residu asam amino, serta potensi perbedaan fungsi antar gen PNase sebelum dilakukan validasi eksperimental (Gambar 6.2).

Multiple *alignment* sekuens asam amino PNase dari berbagai strain *L. brevis* dan *Lactobacillus fermentum* menunjukkan tingkat konservasi yang tinggi, mencerminkan peran katalitik yang esensial dan relatif terjaga secara evolusioner. Namun demikian, perbedaan jumlah gen PNase antar spesies menjadi temuan penting, di mana *L. brevis* memiliki empat gen PNase, sementara *L. fermentum* hanya memiliki tiga.

Perbedaan ini mengindikasikan kemungkinan diversifikasi fungsi atau efisiensi enzim di antara gen-gen tersebut. Analisis MAS memungkinkan identifikasi variasi residu spesifik yang berpotensi mempengaruhi afinitas substrat atau stabilitas struktur enzim, sehingga memberikan dasar rasional untuk mengekspresikan dan menguji setiap gen secara terpisah.

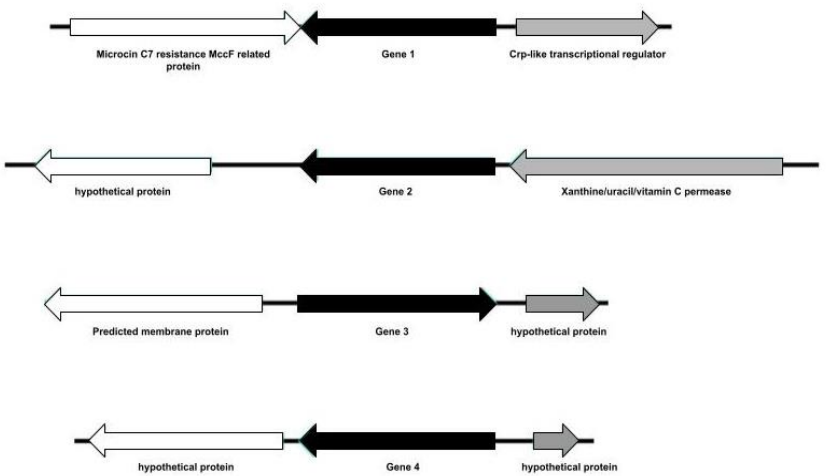




		L. brevis ATCC 367				*L. brevis* NTCC 13768				*L. brevis* UCCLBBS124				*L. brevis* NCTC NPS-QW-145				*L. brevis* SRCM101106			
		G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4
L. brevis ATCC367	G1	100	34	28	18	100	34	29	18	99	34	29	18	100	34	29	18	100	34	29	18
	G2	34	100	25	19	33	99	25	19	33	99	25	19	34	99	25	19	34	99	25	19
	G3	28	25	100	20	28	27	99	20	27	27	99	20	28	27	99	20	28	27	99	20
	G4	18	19	20	100	18	18	21	99	18	18	21	99	18	18	21	99	18	18	21	98
L. brevis NTCC 13768	G1	100	33	28	18	100	33	28	18	99	33	28	18	99	33	28	18	99	33	28	18
	G2	34	99	27	18	33	100	27	19	33	100	25	19	33	100	27	19	33	100	27	18
	G3	29	25	99	21	28	27	100	21	28	27	99	20	28	27	100	20	28	27	100	20
	G4	18	19	20	99	18	19	21	100	18	19	21	99	18	19	21	99	18	19	21	99
L. brevis UCCLBBS124	G1	99	33	27	18	99	33	28	18	100	33	28	18	99	33	28	18	99	33	28	18
	G2	34	99	27	18	33	100	27	19	33	100	25	19	33	100	27	19	33	100	27	18
	G3	29	25	99	21	28	25	99	21	28	25	100	20	27	27	99	20	28	27	99	20
	G4	18	19	20	99	18	19	20	99	18	19	20	100	18	19	20	100	18	19	20	99
L. brevis NCTC NPS-QW-145	G1	100	34	28	18	99	33	28	18	99	33	27	18	100	33	28	18	99	33	28	18
	G2	34	99	27	18	33	100	27	19	33	100	27	19	33	100	27	19	33	100	27	18
	G3	29	25	99	21	28	27	100	21	28	27	99	20	28	27	100	20	28	27	100	20
	G4	18	19	20	99	18	19	20	99	18	19	20	100	18	19	20	100	18	19	20	99
L. brevis SRCM101106	G1	100	34	28	18	99	33	28	18	99	33	28	18	99	33	28	18	100	33	28	18
	G2	34	99	27	18	33	100	27	19	33	100	27	19	33	100	27	19	33	100	27	18
	G3	29	25	99	21	28	27	100	21	28	27	99	20	28	27	100	20	28	27	100	20
	G4	18	19	20	98	18	18	20	99	18	18	20	99	18	18	20	99	18	18	20	100

Gambar 6.2. Analisis Filogenetik dan Kemiripan Sekuens Gen PNase Putatif pada Galur *L. brevis*. (A) Pohon filogenetik menggambarkan hubungan evolusioner di antara gen PNase putatif yang diidentifikasi pada berbagai galur *L. brevis*. (B) Analisis kemiripan sekuens menunjukkan persentase identitas di antara gen-gen PNase tersebut, dengan menyoroti daerah terkonservasi dan variabel yang dapat memengaruhi fungsi enzimatik.

Selain kesamaan sekuens, analisis susunan gen (*gene arrangement*) menunjukkan bahwa keempat gen PNase pada genom *L. brevis* memiliki orientasi dan jarak yang berbeda terhadap gen tetangganya (Gambar 6.3). Hal ini mengindikasikan bahwa sebagian besar gen PNase kemungkinan ditranskripsi secara independen, sehingga ekspresi dan aktivitasnya tidak selalu berkaitan langsung dengan konteks operon.



Gambar 6.3. Susunan gen purine nucleosidase pada genom *L. brevis*

Uji aktivitas degradasi guanosin dan inosine menggunakan HPLC menunjukkan perbedaan yang sangat kontras antar gen PNase. Dari keempat gen yang diekspresikan, hanya Gene 3 yang secara konsisten menunjukkan degradasi 100% terhadap kedua substrat pada seluruh bentuk preparasi, yaitu *whole-cell*, supernatan, enzim murni, dan enzim terkonsentrasi.

Gen lain menunjukkan aktivitas yang relatif rendah dan mendekati kontrol, meskipun memiliki tingkat kesamaan sekuens

yang tinggi berdasarkan analisis MAS. Temuan ini menegaskan bahwa konservasi sekuens tidak selalu berbanding lurus dengan aktivitas katalitik aktual, sehingga pendekatan MAS perlu selalu diintegrasikan dengan validasi eksperimental.

Studi kasus ini menegaskan bahwa implementasi Multiple *Alignment* Sequences (MAS) berperan penting sebagai tahap awal seleksi gen dalam penelitian enzim berbasis bakteri asam laktat. MAS memberikan gambaran konservasi dan variasi sekuens, namun tidak dapat berdiri sendiri tanpa konfirmasi melalui pendekatan bioteknologi molekuler dan uji aktivitas enzim. Integrasi bioinformatika dan eksperimen menjadi kunci dalam mengidentifikasi gen fungsional dengan potensi aplikasi nyata.

Kesimpulan

Multiple *Sequence Alignment* adalah instrumen fundamental dalam bioinformatika yang menjembatani data mentah sekuens dengan pemahaman biologis yang mendalam. Dengan mengintegrasikan algoritme penyejajaran yang tepat dan analisis evolusi yang komprehensif, peneliti dapat mengungkap rahasia adaptasi organisme, mekanisme kerja enzim pada tingkat molekuler, dan potensi pemanfaatan materi genetik untuk solusi kesehatan global.

DAFTAR PUSTAKA

ExPASy. (n.d.). EC 3.2.2.1: Purine nucleosidase. *ENZYME database*. Retrieved February 16, 2026, from https://enzyme.expasy.org/EC/3.2.2.1

Ferdiansyah, M. K., Ji, S. H., Park, B., Kwon, Y. H., Cha, M. S., Manasa, G., & Kim, K.-P. (2025). Cloning, biochemical identification, and characterization of purine nucleoside N-ribohydrolase in *Levilactobacillus brevis*. *Gene, 970*, 149804. https://doi.org/10.1016/j.gene.2025.149804

Ferdiansyah, M. K., Ji, S. H., Cha, M. S., Kwon, Y. H., Kim, G. Y., Park, B., Manasa, G., & Kim, K.-P. (2025). The newly isolated *Levilactobacillus brevis* LABC170 and *Limosilactobacillus fermentum* LABC37 with purine nucleosides degradation activity show probiotic efficacy in prevention and treatment of hyperuricemia in mice. *Biocatalysis and Agricultural Biotechnology, 50*, 103592. https://doi.org/10.1016/j.bcab.2025.103592

Ferdiansyah, M. K., Kang, H. S., Kim, G. Y., Park, B., Kularathna, R. M. R. E., Abraha, H. B., & Kim, K.-P. (2023). Purine nucleosidase (PNase) activity, probiotic potential, and food applicability of a newly isolated *Levilactobacillus brevis* LAB42. *Food Science and Technology International*. Advance online publication. https://doi.org/10.1177/10820132231219859

Ranjbarian, F., Rafie, K., Shankar, K., Krakovka, S., Svärd, S. G., Carlson, L.-A., & Hofer, A. (2023). *Giardia intestinalis* deoxyadenosine kinase has a unique tetrameric structure that enables high substrate affinity and makes the parasite

sensitive to deoxyadenosine analogues. *bioRxiv*. https://doi.org/10.1101/2023.12.18.572228

Wiltgen, M. (2019). Algorithms for structure comparison and analysis: Homology modelling of proteins. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of bioinformatics and computational biology* (pp. 38–61). Academic Press. https://doi.org/10.1016/B978-0-12-809633-8.20484-6

Zupok, A., Kozul, D., Schöttler, M. A., Niehörster, J., Garbsch, F., Liere, K., Fischer, A., Zoschke, R., Malinova, I., Bock, R., & Greiner, S. (2021). A photosynthesis operon in the chloroplast genome drives speciation in evening primroses. *The Plant Cell, 33*(8), 2583–2601. https://doi.org/10.1093/plcell/koab155

BAB 7
Filogenetika Komputasional

Rolina Anna Erica Sihombing

Filogenetika menjelaskan hubungan dan sejarah evolusi antar gen atau makromolekul biologis lain dalam suatu diagram menyerupai pohon. Pohon ini dapat mendeskripsikan hasil modifikasi akibat seleksi alam dan mutasi genetik tersebut, baik dalam bentuk penyatuan (konvergensi) dan/atau pemisahan (divergensi) suatu gen yang kemudian berevolusi menjadi 2 gen atau lebih. Data yang digunakan dalam konstruksi pohon filogenetik antara lain sekuens nukleotida dan sekuens protein.

Dalam konstruksi pohon filogenetika, terdapat beberapa asumsi sebagai berikut; (1) data molekuler yang digunakan dalam konstruksi pohon filogenetik bersifat homolog, yaitu berasal dari nenek moyang yang sama dan kemudian mengalami divergensi seiring berjalannya waktu; (2) divergensi dalam filogenetika bersifat bifurkasi, yaitu induk terbagi menjadi dua cabang anak pada titik tertentu; (3) sekuens yang dibandingkan berevolusi secara independen; (4) data sekuens yang dipilih mewakili unit molekuler yang akan dianalisis hubungan evolusinya (spesies, gen, protein); (5) sekuens molekuler berevolusi pada laju yang konstan sehingga jumlah mutasi yang terakumulasi berbanding lurus dengan waktu evolusi. Dengan asumsi ini, maka panjang

cabang dari suatu pohon dapat digunakan untuk memperkirakan waktu terjadinya divergensi antar takson.

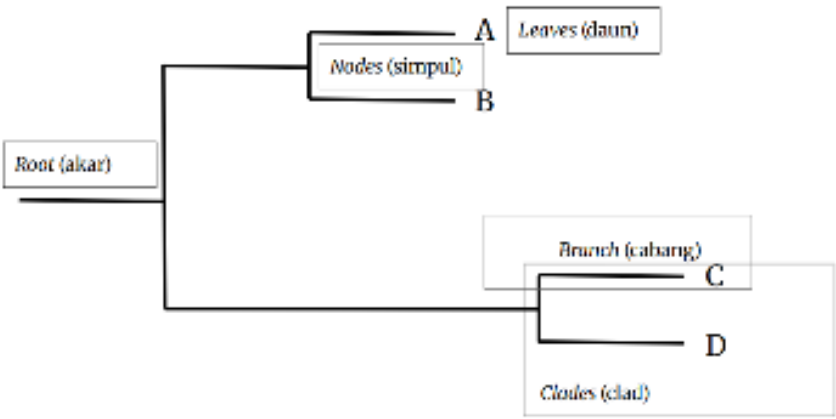
Terminologi dalam Filogenetika

Dalam penyusunan dan analisis pohon filogenetika, terdapat beberapa kosakata khusus untuk karakterisasi pohon filogenetik. Kosakata tersebut dapat dikelompokkan menjadi kosakata untuk menjelaskan grafis pohon filogenetika, topologi pohon filogenetika, dan jenis pohon filogenetika.

Grafis pohon filogenetika

Pohon filogenetika umumnya divisualisasikan seperti pada Gambar 1. Sesuai dengan namanya, pohon filogenetika memiliki **daun (*leaves*)** yang merupakan istilah lain untuk satuan takson (OTU, Operational Taxonomic Units) yaitu gen/protein/spesies yang ingin dipelajari hubungan kekerabatannya. Kemudian terdapat **simpul (*node*)** yang menjadi titik pertemuan dari dua cabang yang berdekatan. Dalam pohon filogenetika, simpul merepresentasikan nenek moyang terdekat dari dua takson (*Most Recent Common Ancestor*, MRCA). Selanjutnya terdapat **klad (*clade*)**, yang merupakan kelompok dari beberapa takson dengan satu nenek moyang yang sama. Taksa, kumpulan dari beberapa takson, yang tergabung dalam satu klad seringkali disebut sebagai **saudara** (*sister taxa*) satu sama lain. Lalu, terdapat **cabang** (***branch***) yang menjelaskan relasi antar klad atau antar OTU dengan bagian lain pada pohon filogenetik. Berikutnya, terdapat **akar (*root*)** yang merupakan nenek moyang yang sama dari semua takson pada pohon filogenetik. Terdapat pula pohon filogenetika tanpa akar (***unrooted***) yang tidak mengidentifikasi jalur evolusi, sehingga terdapat kemungkinan

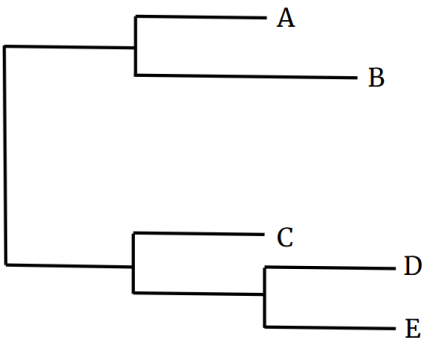
salah satu **simpul (*node*)** atau takson dapat berkerabat lebih dekat dengan nenek moyang bersama **(akar, *root*)** dibandingkan dengan simpul atau takson lainnya (Gambar 7.1).



Gambar 7.1. Diagram dan istilah dalam pohon filogenetika

Format komputasi pohon filogenetika

Informasi topologi pohon filogenetika dapat dimasukkan ke komputer dalam format teks, disebut sebagai format Newick (Gambar 7.2). Dalam format Newick, pohon direpresentasikan oleh takson yang ditulis dalam tanda kurung (). Setiap simpul internal ditulis sebagai pasangan tanda kurung () yang meliputi semua anggota kelompok monofiletik (satu klad) yang dipisahkan dengan koma (,). Untuk pohon berskala, panjang cabang dalam satuan tertentu dan ditempatkan tepat setelah nama takson yang dipisahkan oleh titik dua.



((A:1,B:1.5),((D:2,E:2),C:1))

Gambar 7.2. Contoh pohon filogenetika dan penulisan format Newick dari pohon tersebut

Prosedur Konstruksi Pohon Filogenetika

Secara garis besar, prosedur dalam konstruksi pohon filogenetika dapat dibagi menjadi 5 tahapan, yaitu: (1) pemilihan data; (2) penyejajaran data-data sekuens; (3) pemilihan model evolusi dan pembuatan matriks substitusi; (4) menentukan metode konstruksi pohon filogenetika; serta (5) evaluasi, penilaian, dan visualisasi pohon filogenetika. Pada sub bab berikutnya akan dibahas lebih detail mengenai tahapan (4) dan (5).

Metode Jarak dalam Konstruksi Pohon Filogenetika: Konversi Hasil Penyejajaran Sekuens menjadi Data Jarak

Metode jarak dalam konstruksi pohon filogenetika menghitung seberapa berbeda dua sekuens ketika disejajarkan. Metode jarak mengasumsikan semua sekuens yang dibandingkan berasal dari nenek moyang yang sama (homolog) dan panjang cabang pada pohon bersifat aditif, yaitu jarak evolusi antara dua organisme sama dengan jumlah seluruh

cabang yang menghubungkan kedua organisme tersebut dalam pohon filogenetika. Hasil perhitungan model evolusi dapat digunakan untuk membuat matriks jarak antar sekuens takson. Pohon filogenetika kemudian dibangun dari nilai jarak antar dua takson pada matriks. Algoritme konstruksi pohon filogenetika dengan metode jarak dapat dibagi menjadi berbasis klaster dan berbasis optimalitas. Algoritme berbasis klaster mengkonstruksi pohon berdasarkan matriks jarak dimulai dari pasangan yang paling identik, memiliki paling sedikit perbedaan pada sekuens. Algoritme berbasis klaster antara lain *unweighted pair group method using arithmetic average* (UPGMA) dan *Neighbor Joining*. Algoritme berbasis klaster umum dipakai dalam analisis filogenetika. Algoritme berbasis optimalitas membandingkan berbagai topologi pohon dan memilih pohon yang paling sesuai berdasarkan estimasi jarak pada pohon dan jarak evolusi sesungguhnya. Algoritme berbasis optimalitas antara lain Fitch-Margoliash dan algoritme evolusi minimum.

Unweighted Pair Group Method Using Arithmetic Average (UPGMA)

Sesuai dengan namanya, metode UPGMA membuat klaster dalam pohon filogenetika dengan menghitung jarak antar kelompok menggunakan rata-rata aritmatika dari semua nilai jarak antar kelompok, yang kemudian menggabungkan kelompok-kelompok tersebut menjadi suatu kelompok baru berdasarkan pasangan dengan jarak terdekat hingga semua takson tergabung dalam satu kelompok. UPGMA secara implisit mengasumsikan adanya laju substitusi residu yang konstan. Analisis UPGMA dimulai dengan pengelompokkan kedua takson dengan jarak genetik terdekat, contohnya OTU A dan OTU B.

Kedua takson tersebut dikelompokkan dan membentuk OTU AB. Selanjutnya, matriks jarak yang baru dihitung antara OTU AB dengan OTU lainnya, misalnya OTU X. Nilai rata-rata yang diperoleh dari matriks jarak baru ini menjadi nilai jarak evolusi antara OTU AB dengan OTUX. Jarak antara OTU AB dengan OTU X (atau OTU lain) dapat dihitung dengan persamaan sebagai berikut:

$$d_{(AB)\ X} = ½\ (d_{AX} + d_{BX})$$

Pada pengulangan selanjutnya, kedua takson dengan jarak terkecil kembali dikelompokkan, dan proses ini terus berulang hingga menyisakan 2 OTU.

Berikut merupakan rumus yang digunakan untuk menghitung rata-rata jarak dalam pembuatan klaster.

$$d_{(C1\ +\ C2)} = [n_1 / (n_1 + n_2)]d_{C1D} + [n_2 / (n_1 + n_2)]d_{C2D}$$

dengan:

C1, C2 : klaster dengan takson n1 dan n2

C1C2 : hasil penggabungan C1 dan C2 menjadi OTU baru (OTU C1C2)

d_{C1D}, d_{C2D} : jarak antara OTU C1 dengan OTU D, jarak antara OTU C2 dengan OTU D

Penghitungan juga dapat dilakukan dengan formula alternatif sebagai berikut.

$$d_{(C1C2)D} = ½\ (d_{C1D} + d_{C2D})$$

Neighbor Joining

UPGMA menggunakan asumsi bahwa semua takson memiliki laju evolusi konstan. Mengingat bahwa asumsi ini seringkali tidak terpenuhi pada kondisi sesungguhnya, maka metode *Neighbor-Joining* (NJ) dapat digunakan untuk konstruksi pohon filogenetika yang lebih akurat. NJ merupakan metode

berbasis klaster yang melakukan koreksi untuk laju evolusi yang tidak sama antar sekuens melalui tahapan konversi. Konversi ini memerlukan perhitungan nilai-r dan nilai-r yang ditransformasikan (r′) dengan persamaan sebagai berikut.

$$d'_{AB} = d_{AB} - ½ \times (r_A + r_B)$$

dengan:

d'_{AB} : jarak antara A dan B yang telah dikonversi

d_{AB} : jarak evolusi antara A dan B yang sesungguhnya

r_A, r_B : jumlah jarak dari A (atau B) ke semua takson yang lain

Nilai r secara umum dapat diekspresikan sebagai r_i, dan dapat dihitung dengan persamaan sebagai berikut.

$$r_i = \sum d_{ij}$$

dengan i dan j merupakan 2 takson yang berbeda. Nilai r ini kemudian digunakan untuk membuat matriks jarak. Nilai r yang ditransformasi (r′) digunakan untuk menentukan jarak dari suatu takson ke simpul (*node*) terdekat.

$$r'_i = r_i / n - 2$$

dengan n merupakan jumlah taksa. JIka takson A dan takson B membentuk simpul baru yang disebut sebagai U, maka jarak antara A dan U dapat dihitung dengan persamaan berikut.

$$d_{AU} = d_{AB} + [(r'_A + r'_B)]/2$$

Hal lain yang membedakan NJ dengan UPGMA adalah dalam prosedur konstruksi pohon filogenetika. UPGMA memulai konstruksi pohon dengan pasangan takson yang memiliki jarak paling dekat, sedangkan NJ memulai konstruksi pohon dengan struktur pohon filogenetika yang menyerupai bintang. Struktur ini dapat terbentuk karena NJ mengasumsikan semua takson langsung terhubung ke satu titik pusat tanpa hubungan kekerabatan yang jelas. Kemudian, struktur pohon dipisahkan

secara bertahap dengan memilih pasangan takson yang paling dekat berdasarkan jarak yang sudah dikoreksi. Pasangan takson tersebut akan digabung terlebih dulu membentuk satu simpul (*node*). Simpul ini akan dianggap sebagai satu OTU baru, sehingga jumlah takson dalam matriks jarak akan berkurang satu. Selanjutnya, takson lain yang paling dekat dapat bergabung dengan simpul yang sudah terbentuk. Langkah ini terus mengalami pengulangan hingga semua takson terhubung dalam pohon filogenetika.

Mengingat bahwa metode NJ tidak mengasumsikan semua takson memiliki laju evolusi yang konstan, maka pohon filogenetika yang dihasilkan dengan metode ini tidak dapat menunjukkan arah waktu evolusi, tetapi terbatas pada hubungan kekerabatan relatif antara satu takson dengan yang lainnya. Namun, akar pohon dari metode ini tetap dapat ditentukan dengan menambahkan ***outgroup***, yaitu sekuens pembanding yang tidak termasuk dengan sekuens kelompok utama. Pemilihan sekuens *outgroup* dilakukan berdasarkan pengetahuan biologis atau evolusi yang sudah ada. Umumnya, *outgroup* dipilih berdasarkan perbedaan dari karakteristik takson yang ingin dikomparasi. Misalnya, jika ingin membandingkan takson dari beberapa kelompok mamalia, maka *outgroup* yang digunakan dapat berasal dari reptil.

Limitasi dari algoritme NJ dalam konstruksi pohon filogenetika adalah hanya menghasilkan satu pohon dan tidak memberikan kemungkinan topologi pohon lainnya dari taksa yang dianalisis. Pada tahap awal konstruksi pohon dengan algoritme ini, dapat ditemukan lebih dari satu pasang OTU dengan jarak yang paling dekat. Untuk mengatasi limitasi tersebut, dapat digunakan metode *Generalized Neighbor-Joining*

(GNJ) yang menghasilkan banyak kemungkinan topologi pohon berdasarkan pengelompokkan awal takson yang paling dekat dan besaran jarak yang sama.

Metode Karakter dalam Konstruksi Pohon Filogenetika: Algoritme Penilaian Pohon Filogenetika

Metode berbasis karakter melakukan konstruksi pohon filogenetika mengacu pada karakter masing-masing sekuens, yaitu dengan menghitung kejadian mutasi yang terjadi pada sekuens dan mempertahankan informasi mutasi tersebut ketika karakter sekuens dikonversi menjadi jarak. Metode berbasis karakter yang umum digunakan antara lain Maximum Parsimony, Maximum *Likelihood*, dan Bayesian Inference.

Maximum Parsimony

Maximum Parsimony (MP) memilih pohon dengan perubahan evolusi paling sedikit atau total panjang cabang paling pendek, berdasarkan prinsip *Occam's razor*, penjelasan yang lebih sederhana biasanya paling mendekati kebenaran. Dalam konteks pohon filogenetika, pohon dengan jumlah substitusi paling sedikit memiliki kemungkinan terbaik untuk menjelaskan perbedaan antar taksa dalam pohon. Dengan demikian, pohon dengan perubahan paling sedikit adalah topologi yang paling baik untuk menggambarkan kondisi evolusi yang sesungguhnya.

Konstruksi pohon filogenetika dimulai dengan mencari semua topologi pohon yang mungkin terbentuk dan rekonstruksi sekuens nenek moyang yang memerlukan perubahan paling sedikit untuk berevolusi menjadi sekuens tersebut. Untuk menghemat proses komputasi, maka, untuk menentukan

topologi pohon, dipilih beberapa situs yang memiliki informasi filogenetika paling tinggi. Situs terpilih tersebut adalah situs informatif, yaitu situs yang memiliki sedikitnya 2 jenis karakter yang berbeda. Kemudian, situs selain situs informatif, disebut sebagai situs non-informatif atau situs konstan, dibuang dalam konstruksi pohon parsimony. Selanjutnya, jumlah substitusi terkecil dari setiap situs informatif dihitung. Total jumlah substitusi dari semua situs informatif dijumlahkan untuk setiap topologi pohon. Topologi pohon dengan jumlah substitusi terkecil dipilih sebagai topologi pohon terbaik.

Poin utama dalam menghitung jumlah substitusi minimum untuk situs tertentu adalah menentukan keadaan karakter nenek moyang pada simpul internal. Prinsip parsimony akan memilih keadaan karakter nenek moyang yang menghasilkan jumlah substitusi minimum. Prediksi sekuens nenek moyang dibuat dengan meninjau sekuens setiap takson (*leaves*), ke simpul internal (*nodes*), dan ke bagian akar (*root*) untuk menentukan semua kemungkinan keadaan karakter leluhur. Selanjutnya, kembali dari akar ke simpul internal dan ke daun untuk menentukan jumlah substitusi minimum yang dapat terjadi pada sekuens nenek moyang, sehingga berdivergensi menjadi taksa dalam pohon filogenetika.

Maximum ***Likelihood***

Maximum Likelihood (ML) menggunakan model peluang untuk memilih pohon dengan topologi terbaik yang memiliki peluang atau kemungkinan (*likelihood*) tertinggi untuk memproduksi data yang teramati. Berbeda *dengan Maximum Likelihood* yang membangun pohon filogenetika dari situs informatif dan mengabaikan situs non-informatif, ML merupakan

metode komprehensif yang mencari setiap topologi pohon yang mungkin dan mempertimbangkan setiap posisi di penyejajaran. Dengan menggunakan metode substitusi tertentu yang memiliki nilai probabilitas substitusi residu, ML menghitung total kemungkinan dari sekuens nenek moyang berevolusi ke simpul internal, dan kemudian ke sekuens yang ada.

ML membangun pohon filogenetika dengan menghitung probabilitas jalur evolusi tertentu untuk sekuens yang ada. Nilai probabilitas tersebut ditentukan dari model substitusi. Untuk situs tertentu, probabilitas jalur pohon adalah hasil perkalian dari akar ke semua takson, termasuk cabang perantara dalam topologi pohon. Hasil perkalian tersebut umumnya menghasilkan nilai yang sangat kecil. Untuk mempermudah proses komputasi, maka hasil perkalian tersebut dinyatakan dalam logaritma *natural likelihood* (lnL), sehingga operasi perkalian menjadi penjumlahan. Nilai probabilitas juga menghitung semua skenario yang mungkin untuk sekuens nenek moyang. Setelah dilakukan konversi logaritmik, nilai *likelihood* untuk topologi adalah jumlah logaritma *natural likelihood* dari setiap cabang pohon. Setelah menghitung semua kemungkinan jalur pohon dengan kombinasi sekuens nenek moyang yang mungkin, jalur pohon yang memiliki skor *likelihood* tertinggi adalah topologi akhir di situs tersebut. Skor log *likelihood* untuk setiap situs dihitung secara independen, karena semua karakter diasumsikan telah berevolusi secara independen. Skor log *likelihood* keseluruhan untuk jalur pohon tertentu pada semua sekuens merupakan jumlah log *likelihood* dari semua situs individual. Prosedur penghitungan jumlah log *likelihood* ini harus diulangi untuk semua kemungkinan topologi pohon. Topologi

pohon dengan skor *likelihood* tertinggi adalah pohon *Maximum Likelihood* terpilih.

Bayesian inference

Metode Bayesian merupakan pengembangan dari metode ML. Metode Bayesian menggunakan distribusi statistik untuk menghitung ketidakpastian pada parameter yang digunakan. Metode analisis ini menggunakan konsep yang dikenal sebagai probabilitas posterior, yaitu probabilitas yang direvisi dari ekspektasi sebelumnya, setelah mempelajari sesuatu yang baru tentang data. Dalam istilah matematika, analisis Bayesian menghitung probabilitas posterior dari dua peristiwa gabungan dengan menggunakan nilai probabilitas prior dan probabilitas kondisional dengan persamaan sebagai berikut.

Probabilitas posterior = (Probabilitas awal * Probabilitas kondisional)/Probabilitas total

Dalam pemilihan topologi pohon berdasarkan analisis Bayesian, probabilitas awal merupakan peluang untuk semua topologi pohon yang mungkin sebelum dilakukan analisis. Sebelum konstruksi pohon, peluang untuk semua topologi pohon tersebut adalah sama. Probabilitas kondisional adalah frekuensi substitusi dari karakter yang teramati dari hasil penyejajaran sekuens. Kedua informasi tersebut digunakan sebagai kondisi oleh algoritme Bayesian untuk menemukan topologi pohon yang paling mungkin.

Proses konstruksi pohon dengan metode Bayesian dimulai dengan pemilihan model evolusi berdasarkan sekuens yang digunakan dan informasi awal parameter seperti topologi pohon

dan panjang cabang. Menurut teori Bayes, mengkombinasikan informasi parameter awal dengan *likelihood* dari data sekuens akan menghasilkan informasi posterior dari parameter, yaitu distribusi peluang posterior dari parameter. Selanjutnya, pengambilan sampel berulang dengan prosedur Markov Chain Monte Carlo (MCMC) dilakukan. MCMC secara berulang mencari nilai *likelihood* yang lebih tinggi ketika mencari topologi pohon terbaik. Dengan prosedur tersebut, topologi pohon dengan nilai *likelihood* tertinggi akan lebih sering diambil oleh algoritme. Ketika prosedur MCMC mencapai topologi-topologi pohon dengan *likelihood* yang tinggi, kumpulan topologi pohon tersebut akan terpilih dan disertakan dalam konstruksi pohon konsensus.

Keunggulan metode Bayesian dalam konstruksi pohon filogenetika adalah dapat menangani dataset berukuran besar pada kecepatan komputasi yang lebih tinggi dibandingkan dengan metode ML. Selain itu, metode Bayesian dapat mengukur tingkat kepercayaan suatu topologi pohon dengan probabilitas posterior. Metode Bayesian dinilai dapat menghasilkan topologi pohon yang lebih baik dibandingkan dengan metode ML. Hal ini dikarenakan metode ML hanya mencari satu topologi pohon terbaik, sedangkan metode Bayesian melakukan konstruksi pohon konsensus dari kumpulan topologi pohon terbaik.

Evaluasi Hasil Konstruksi Pohon Filogenetika

Topologi-topologi pohon filogenetika yang dibangun dari berbagai algoritme di atas adalah hipotesis terkait evolusi, sehingga diperlukan beberapa pendekatan untuk menguji hipotesis tersebut. Pendekatan statistik umum digunakan untuk mengevaluasi pohon hasil analisis filogenetika, apakah topologi

pohon tersebut dapat dipercaya, sesuai, dan/atau mendekati kondisi evolusi yang sesungguhnya. Pertanyaan evaluasi tersebut umumnya dijawab dengan strategi *bootstrapping* dan Bayesian pada konstruksi pohon dengan metode Bayesian *Inference* seperti yang telah dijelaskan pada sub bab sebelumnya.

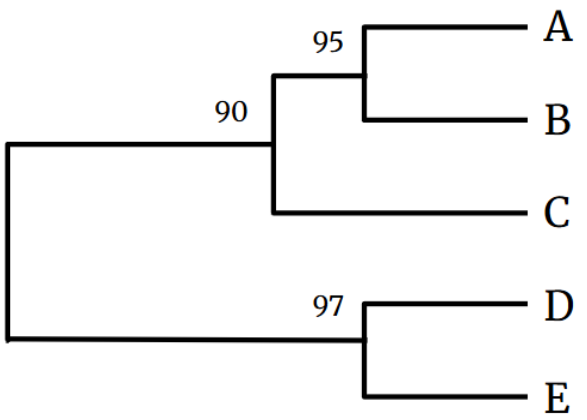
Bootstrapping

Bootstrap merupakan metode statistik yang menguji kesalahan pengambilan sampel pada konstruksi pohon filogenetika. Metode ini bekerja dengan mengambil sampel pohon secara berulang kali dari kumpulan data yang sedikit diubah. Dengan demikian, kekokohan pohon hasil konstruksi dapat dinilai. Untuk menentukan kekokohan atau reprodusibilitas dari topologi pohon yang terbentuk, pohon dibangun berulang kali dengan penyejajaran yang sedikit diubah dan hasil penyejajaran dipaparkan dengan beberapa fluktuasi acak (Gambar 7.3). Hubungan filogenetik yang kuat akan memiliki karakter yang cukup untuk mempertahankan hubungan tersebut, bahkan jika kumpulan data diubah sedemikian rupa. Jika hal ini tidak terpenuhi, maka data *noise* dalam proses pengambilan sampel ulang dapat menghasilkan pohon yang berbeda. Hal ini mengindikasikan bahwa topologi asli mungkin berasal dari sinyal filogenetik yang lemah. Dengan demikian, hasil analisis *bootstrap* memberikan gambaran tentang kepercayaan topologi pohon dengan pendekatan statistik.

Pengambilan sampel repetitif dalam *bootstrap* bergantung pada pengubahan set data sekuens asli. Terdapat dua strategi dalam mengubah dataset asli tersebut. Strategi pertama untuk menghasilkan perubahan pada set data adalah melalui penggantian situs secara acak, yang disebut sebagai strategi

***bootstrapping* nonparametrik**. Strategi kedua adalah menghasilkan set data sekuens yang baru berdasarkan distribusi sekuens tertentu dari data asal, yang disebut sebagai strategi ***bootstrapping* parametrik**. Dengan model distribusi yang tepat, maka replikasi situs untuk set data yang baru akan lebih wajar dibandingkan replikasi situs secara acak. Kemampuan ini menyebabkan strategi parametrik dinilai menghasilkan topologi pohon yang lebih kokoh dibandingkan strategi nonparametrik. Secara garis besar, kedua strategi ini dapat diaplikasikan dalam evaluasi pohon filogenetik dengan metode jarak, *parsimony*, dan *maximum likelihood*.

Perlu diingat bahwa *bootstrap*ping tidak menilai akurasi pohon, tetapi hanya mengindikasikan konsistensi dan stabilitas masing-masing cabang dari pohon. Akibatnya, pohon dengan nilai *bootstrap* yang tinggi tetap dapat dihasilkan jika terjadi galat secara sistematis dalam prosedur konstruksi pohon. Selain itu, jika dataset asli memiliki kandungan GC yang tinggi, biasanya akan dihasilkan laju evolusi yang tidak biasa dan model evolusi yang tidak wajar, sehingga menghasilkan estimasi *bootstrap* yang bias. Untuk meningkatkan akurasi pohon yang dihasilkan, jumlah *bootstrap* yang direkomendasikan untuk dilakukan pada pohon filogenetik adalah antara 500-1000 kali. Dalam beberapa kasus, operasi prosedur ini memerlukan waktu komputasi yang cukup lama.



Gambar 7.3. Contoh pohon filogenetika dengan nilai bootstrap. Cabang AB muncul dalam proses pengambilan sampel berulang sebanyak 95 kali, cabang DE muncul dalam proses pengambilan sampel berulang sebanyak 97 kali, cabang AB-C muncul dalam proses pengambilan sampel berulang sebanyak 90 kali

DAFTAR PUSTAKA

Christensen, H. (Ed.). (2018). *Introduction to bioinformatics in microbiology*. Springer International Publishing.

Claverie, J. M., & Notredame, C. (2011). *Bioinformatics for dummies*. John Wiley & Sons.

Nascimento, F. F., Reis, M. D., & Yang, Z. (2017). A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecology & Evolution, 1*(10), 1446–1454. https://doi.org/10.1038/s41559-017-0280-x

Pearson, W. R. (1990). The PAM250 scoring matrix. In *The Global Proteome Machine*. Retrieved 2024, from https://www.thegpm.org/BIOML/aainfo/pam250.htm

Weiß, M., & Göker, M. (2011). Molecular phylogenetic reconstruction. In K. Kurtzman, C. Fell, & T. Boekhout (Eds.), *The yeasts* (pp. 159–174). Elsevier.

Xiong, J. (2006). *Essential bioinformatics*. Cambridge University Press.

Zou, Y., Zhang, Z., Zeng, Y., Hu, H., Hao, Y., Huang, S., & Li, B. (2024). Common methods for phylogenetic tree construction and their implementation in R. *Bioengineering, 11*(5), 480. https://doi.org/10.3390/bioengineering11050480

BAB 8
Bioinformatika Struktur Protein

Lustyafa Inassani Alifia

Protein berperan penting dalam hampir semua proses biologi. Protein dapat menjalankan tugas-tugas pertahanan, transportasi, katalitik, dan struktural di dalam sel (Paiva *et al.*, 2022). Protein sangat penting bagi kehidupan, sehingga dengan memahami strukturnya akan dapat mempermudah pemahaman mekanistik tentang fungsinya (Ronneberger *et al.*, 2021).

Struktur Protein

Struktur protein memiliki ukuran yang bervariasi, dari puluhan hingga ribuan residu. Protein juga dapat diklasifikasikan berdasarkan ukuran fisiknya sebagai nanopartikel (1-100 nm). Asam amino adalah komponen fundamental dalam penyusunan protein, yang umumnya berjumlah 20 asam amino. Setiap satu asam amino terdiri atas satu gugus amino, satu gugus karboksil, satu atom hidrogen, dan satu rantai samping yang terikat pada atom karbon (Pereira & Alva, 2021). Molekul protein mengandung karbon, hidrogen, oksigen, dan nitrogen (Komari *et al.*, 2024), beserta unsur-unsur minor seperti Sulfur dan Fosfor (Liu, 2024). Penyusunan protein dapat dibagi menjadi empat struktur utama yang meliputi struktur primer, sekunder, tersier, dan kuartener. Peranan protein dapat diketahui setelah

memahami urutan asam amino yang menjalani proses pelipatan menjadi 3D sebagai struktur tersier.

Struktur Primer

Urutan asam amino yang membentuk rantai polipeptida menciptakan struktur dasar protein. Struktur dasar ini sering kali disebut sebagai struktur satu dimensi atau 1D. Struktur dasar memainkan peran penting dalam membentuk struktur 2D, 3D, 4D, beserta fungsi protein itu sendiri.

Struktur Sekunder

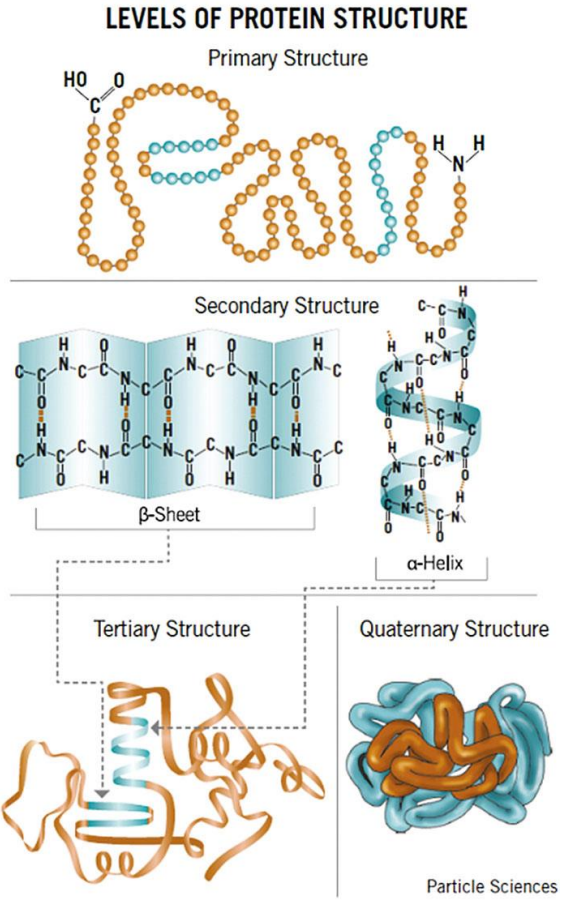
Daerah pada polipeptida yang distabilkan dengan jembatan ikatan hidrogen dan membentuk bentuk yang jelas. Ikatan hidrogen muncul antara atom-atom sepanjang tulang punggung polipeptida. Dalam sebagian besar protein globular, terdapat struktur heliks-α (α-heliks) dan lembaran β yang bergelombang (β-sheet) yang ditemukan berdekatan, yang menunjukkan struktur sekunder.

Struktur Tersier

Konformasi tiga dimensi dari protein yang telah terlipat dan memiliki aktivitas biologis disebut sebagai struktur tersier. Struktur tersier protein dihasilkan dari pelipatan rantai α-helix, β-sheet, atau *random coil* dari polipeptida, yang membentuk struktur globular dengan kompleksitas 3D yang lebih tinggi dibandingkan protein itu sendiri. Interaksi intramolekuler, seperti ikatan hidrogen, ikatan ion, gaya van der Waals, dan interaksi hidrofobik, juga berperan dalam menentukan orientasi struktur tiga dimensi protein.

Struktur Kuartener

Struktur kuartener menunjukkan subunit yang berbeda dan terorganisir bersama untuk membentuk struktur protein. Sebagian molekul protein mampu berinteraksi secara fisik tanpa adanya ikatan kovalen yang menciptakan oligomer yang stabil seperti dimer, trimer, atau tertramer. Kemampuan untuk menyusun diri dari sub-bagian ini menjadi karakteristik utama dari struktur kuartener dalam protein oligomer (Gambar 8.).



Sumber: Wijaya *et al.*, (2014)

Gambar 8.1. Struktur Protein

Pemodelan Protein dalam Bioinformatika

Struktur tiga dimensi dari protein dapat diidentifikasi melalui pendekatan bioinformatika (*in silico*) dengan bantuan komputer, yang memprediksi bentuk 3D protein berdasarkan urutan asam amino yang membentuknya. Ada tiga teknik dalam pemodelan struktur protein tiga dimensi, yaitu metode homologi/komparatif, metode pengenalan lipatan (*fold recognition*), dan metode *ab initio*. Proses homologi lebih sederhana dan lebih cepat dibandingkan dengan metode pengenalan lipatan dan *ab initio*. Pemodelan homologi adalah metode yang menggunakan *template* dari protein yang sudah diketahui struktur 3D-nya melalui percobaan di laboratorium. Selama terdapat kesamaan identitas antara dua protein (target dan *template*) yang serupa dan bertumpuk pada area yang sama, kedua protein tersebut pasti akan mengambil lipatan yang serupa (Makigaki & Ishida, 2020).

Metode Homologi

Metode pemodelan protein secara homologi dilakukan dengan menganalisis kedekatan dari susunan asam amino protein yang diinginkan dengan protein dari spesies lain yang terkait secara evolusi atau berasal dari nenek moyang yang identik. Protein homolog dari spesies lain tersebut merupakan protein yang telah memiliki struktur tiga dimensi yang diketahui melalui penelitian terdahulu, serta memiliki kesamaan susunan asam amino dengan protein target minimal 30%. Oleh karena itu, metode homologi tidak dapat menghasilkan model yang akurat jika *sequence identity* antara target dan *template* rendah (*sequence identity* <30%). Protein homolog ini selanjutnya dipakai sebagai acuan (*template*) untuk merancang model

protein yang diinginkan. Beberapa *web server* yang dapat dimanfaatkan untuk pemodelan homologi meliputi I-TASSER (https://aideepmed.com/I-TASSER/), MODELLER (https://salilab.org/modeller/), dan SWISS-MODEL (https://swissmodel.expasy.org/).

Fold Recognition

Metode *fold recognition* membentuk strukur 3D protein target berdasarkan pendeteksian fragmen yang sama yang didapat dari protein dari basis data Protein Data Bank RSCB PDB, yang memiliki kemiripan urutan asam amino dari protein target. Fragmen-fragmen lipatan yang sudah teridentifikasi ini kemudian dirangkai satu sama lain menjadi suatu struktur 3D protein yang utuh. *Web server* yang digunakan pada pemodelan metode *fold recognition* adalah Phyre2 *automatic fold recognition* (https://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index)

Ab initio

Ab initio adalah permodelan *de novo* dan *free template based*. *Ab initio* didasarkan pada informasi susunan residu asam amino penyusunnya. Memprediksi sruktur 3D protein target dengan berdasarkan informasi sekuens asam amino, yang biasanya dilakukan melalui analisis sifat fisikokimia dan energi struktural, merupakan prediksi yang sulit dijalankan dengan akurat dan masih didapatkan tingkat kesalahan yang tinggi. Hal ini disebabkan karena masih sedikitnya pengetahuan yang didapatkan mengenai mekanisme protein melipat ke dalam bentuk struktur 3D yang spesifik. Pelipatan ini didasarkan pada informasi urutan asam aminonya. Metode pemodelan yang

menerapkan pendekatan *Ab initio* adalah ROSETTA (https://rosettacommons.org/software/).

Teknik Pemodelan Protein

Pemilihan Sequens Protein Target

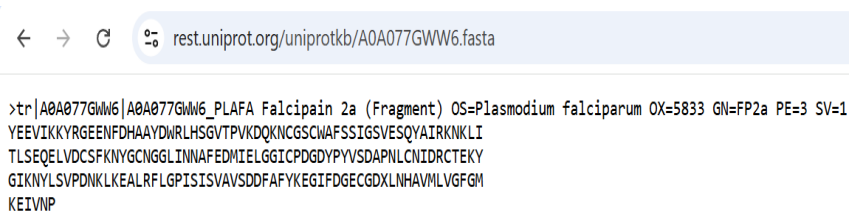
Sebagai contoh, protein yang dipilih adalah protein target *Plasmodium falciparum receptors falcipain-2* (FP-2). Sebuah protein pada spesies *Plasmodium falciparum* penyebab penyakit malaria, yang sekuennya telah ditentukan secara eksperimental, namun belum ditentukan struktur tiga dimensinya. Sekuens asam amino dari protein FP-2 dapat dicari pada basis data Uniprot (https://www.uniprot.org/). Data sekuens asam amino yang tersedia dan belum memiliki struktur 3D dapat disimpan bersamaan lalu diunduh dalam format FASTA.

Tabel 8.1 menunjukkan 6 protein FP-2 pada spesies *Plasmodium falciparum* dari basis data Uniprot. Sebagai contoh, protein FP-2 dengan kode akses A0A077GWW6 dipilih karena tidak ada informasi struktur 3D proteinnya. Sekuens protein FP-2 dengan kode akses A0A077GWW6 disimpan dalam format FASTA (Gambar 2).

Tabel 8.1. Pencarian Protein Target FP-2 pada Database Uniprot

No.	Kode akses	Nama protein	Gen	Spesies	Panjang asam amino
1	Q8I6U5	Falcipain -2b,	*FP2B*, *FP2'*, *PF3D7_1115300*	*Plasmodium falciparum* (isolate 3D7)	482 AA

No.	Kode akses	Nama protein	Gen	Spesies	Panjang asam amino
2	Q8I6U4	Falcipain -2a,	*FP2A*, *FP-2*, *FP2*, *PF3D7_1115700*	*Plasmodium falciparum* (isolate 3D7)	484 AA
3	Q8IIL0	Falcipain -3	*FP3*, *FP-3*, *PF3D7_1115400*	*Plasmodium falciparum* (isolate 3D7)	492 AA
4	Q9NAW3	Falcipain 2	Belum diketahui	*Plasmodium falciparum* (malaria 5parasite P. Fa6lciparum)	484 AA
5	Q9N6S8	Falcipain 2, EC:3.4.22.-	Belum diketahui	*Plasmodium falciparum* (malaria parasite P. falciparum)	484 AA
6	A0A077GWW6	Falcipain 2a	*FP2a*	*Plasmodium falciparum* (malaria parasite P. falciparum)	186 AA



Gambar 8.2. Sekuens Protein FP-2 dalam Format FASTA

Data ini digunakan sebagai dasar pembuatan model struktur 3D protein menggunakan Phyre2, hingga validasi struktur protein.

Analisis Protein Target menggunakan ProtParam

Analisis protein target dilakukan pada *web server* ProtParam *tool* (https://web.expasy.org/protparam/). Sekuens protein target dalam format FASTA disalin dan diisikan/disubmit pada kolom "*acid amino sequence*" lalu klik "*compute parameters"*. Contoh hasil analisis dapat dilihat pada Tabel 8.2.

Tabel 8.2. Sifat Fisika dan Kimia Protein FP-2 Kode A0A077GWW6

Sifat fisika dan kimia	Hasil
Jumlah asam amino	186
Berat molekul	20749.59
Titik isoelektrik (pl)	4.74
Waktu paruh	2.8 jam (retikulosit mamalia); 10 menit (jamur, in vivo); 2 menit (E. Coli, in vivo)
Indeks instabilitas	34.82
Aliphatic index	78.60
Grand Average of hydropathicity (GRAVY)	-0.295
Total number of negatively charged residues (Asp + Glu)	28
Total number of positively charged residues (Arg + Lys)	28

Titik isoelektrik (pI) merupakan suatu ukuran pH pada saat jumlah keseluruhan muatan pada kelompok bermuatan positif dan negatif mencapai titik setimbang (yang menjadikan jumlah

muatan asam amino menjadi netral). Saat level pH berada di atas titik isoelektrik, protein memiliki muatan negatif, sementara saat berada di bawah titik isoelektrik, protein memiliki muatan positif. Indeks instabilitas (II) pada protein target yaitu sebesar 34,82. Indeks instabilitas adalah suatu tolok ukur pada protein yang memberikan prediksi mengenai kestabilan suatu protein. Protein dianggap stabil jika indeks instabilitasnya kurang dari 40. Namun, jika lebih besar dari itu, kemungkinan protein tersebut tidak stabil. Oleh karena itu, FP-2 ini termasuk protein yang tidak stabil, yang akan mempengaruhi bentuk 3D dari protein target.

GRAVY (*grand average of hydropathicity*) atau nilai rata-rata hidropati adalah suatu parameter yang dipakai untuk menentukan karakter hidrofobik sebuah protein: semakin positif nilai indeks hidropati asam amino, maka asam amino tersebut semakin bersifat hidrofobik. Kebalikannya, jika nilai indeks hidropati asam amino semakin negatif, maka asam amino tersebut semakin bersifat hidrofilik. Nilai *grand average of hydropathicity* (GRAVY) untuk protein target adalah -0,295. Nilai ini menandakan bahwa protein target memiliki sifat hidrofilik. Hal ini memiliki pengaruh besar terhadap mekanisme pelipatan sebuah asam amino dalam pembentukan struktur tersiernya. Umumnya, protein yang memiliki banyak residu asam amino dengan sifat hidrofilik mengalami kesulitan dalam membentuk lipatan menjadi struktur tersier (3D) yang baik.

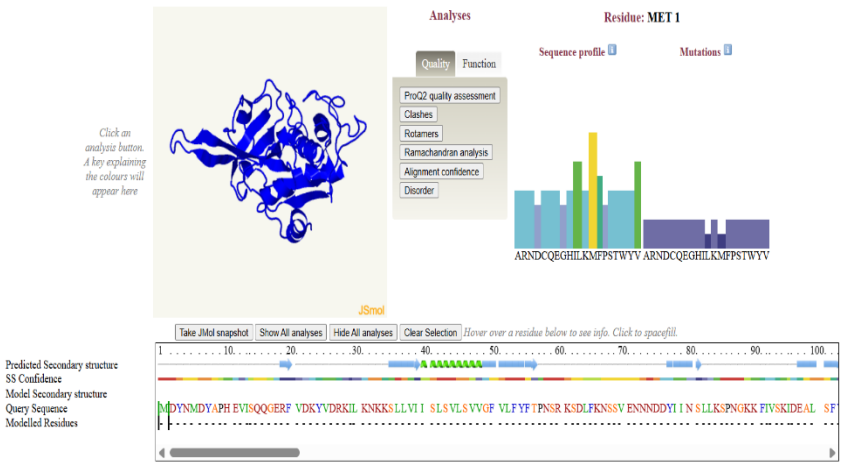
Identifikasi Template pada Swiss-Model

Identifikasi *template* dengan menggunakan Swiss-Model atau dengan Phyre2. Permodelan diawali dengan memilih "*Start Modelling*", lalu sekuens protein FP-2 dengan format FASTA dimasukkan. Setelah itu akan muncul beberapa alternatif

template. *Template* c2oulA dipilih sebagai *template* yang akan dievaluasi lebih lanjut.

Evaluasi Model Protein Menggunakan Swiss-Model

Evaluasi model protein dapat menggunakan *web server* Swiss-Model dan dilakukan validasi menggunakan Phyre2. *Template* protein yang dipilih dilakukan "*Run investigator*" dan "*Confirm Phyre*". Penjajaran sekuens protein target dan *template* dapat dilakukan menggunakan SWISS-MODEL. Hal ini bertujuan untuk mengetahui kemiripan antara sekuens asam amino target dengan *template*. Tampilan evaluasi model protein Swiss-Model dapat dilihat pada Gambar 8.3).



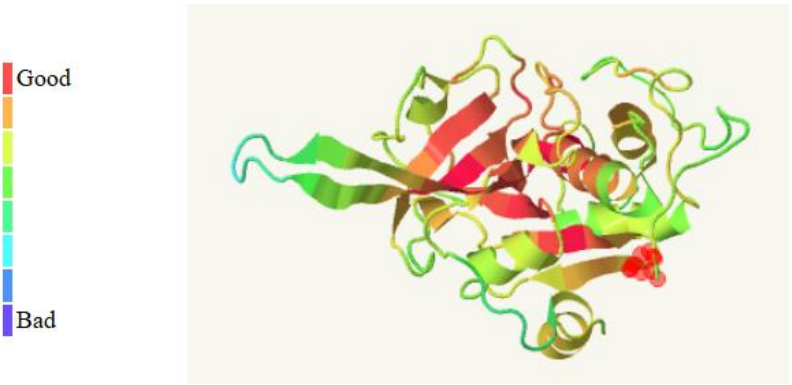
Gambar 8.3. Tampilan tools untuk evaluasi model pada Phyre2

Evaluasi model dilakukan dengan dua parameter, yaitu: kualitas dan fungsi. Kualitas terdiri dari tujuh alat untuk menilai model, yaitu: penilaian kualitas ProQ2 *quality assessment*, Clashes, Rotamers, analisis Ramachandran, *Alignment confidence*, dan Disorder. Fungsi mencakup tiga alat untuk mengevaluasi model, yaitu: konservasi, *pocket detection*, dan sensitivitas mutasi.

Setiap *tools* dari hasil analisis Phyre2 dapat dieksplor untuk mengetahui evaluasi model protein.

Kualitas Model Protein FP-2

Kualitas model protein dapat ditentukan dengan memilih ProQ2 *quality assessment* pada tampilan tools untuk evaluasi model pada Phyre2. Jika nilainya mendekati atau sama dengan 1, maka menunjukkan nilai yang kurang baik. Jika nilainya mendekati atau sama dengan 0, maka menunjukkan nilai yang baik. Tampilan kualitas model proteinnya dapat dilihat pada Gambar 8.4).



Gambar 8.4. Kualitas Model Protein FP-2

Indikator warna dari hijau ke merah menunjukkan daerah residu yang berarti skor *ProQ2* yang baik. Semakin menuju warna merah, maka semakin baik (*Good)*. Sedangkan indikator warna dari hijau ke ungu berarti skor *ProQ2* yang kurang baik. Semakin menuju warna ungu maka semakin kurang baik (*Bad)*.

Validasi Model Protein

Validasi model dilakukan dengan menggunakan program PROCHECK pada *web server* SAVESv6.0 (https://saves.mbi.ucla.edu/). Model protein FP-2 dengan *template* c2oulA diunduh dari *web server* Phyre2, lalu disimpan dalam format (.pdb). *Web server* diakses dan *file* model dalam format .pdb diunggah untuk proses validasi model. Hasil yang akan ditunjukkan meliputi plot Ramachandran *plot*, Chi1-chi2 plots, *Side-chain* params, *Residue properties,* G-factor, dan *Planar groups*.

DAFTAR PUSTAKA

Komari, N., Iskandar, & Suhartono, E. (2024). *Mengungkap misteri struktur protein: Teknik pemodelan in silico untuk penelitian biokimia*. ULM Press.

Liu, D. (2024). Protein structure and function. In *Handbook of molecular biotechnology* (pp. 195–201). CRC Press. https://doi.org/10.1201/9781003055211-22

Makigaki, S., & Ishida, T. (2020). Sequence alignment generation using intermediate sequence search for homology modeling. *Computational and Structural Biotechnology Journal, 18*, 2043–2050. https://doi.org/10.1016/j.csbj.2020.07.012

Paiva, V. de A., Gomes, I. de S., Monteiro, C. R., Mendonça, M. V., Martins, P. M., Santana, C. A., Gonçalves-Almeida, V., Izidoro, S. C., Melo-Minardi, R. C. de, & Silveira, S. de A. (2022). Protein structural bioinformatics: An overview. *Computers in Biology and Medicine, 147*, 105695. https://doi.org/10.1016/j.compbiomed.2022.105695

Pereira, J., & Alva, V. (2021). How do I get the most out of my protein sequence using bioinformatics tools? *Acta Crystallographica Section D: Structural Biology, 77*, 1116–1126. https://doi.org/10.1107/S2059798321007907

Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., … Kavukcuoglu, K. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Wijaya, C. H., Wijaya, W., & Mehta, B. M. (2014). *Handbook of food*

chemistry. Springer. https://doi.org/10.1007/978-3-642-41609-5

BAB 9
Analisis Genom dan *Next Generation Sequencing*

Lolita

Saat ini, perkembangan studi farmakogenomik begitu pesatnya dengan adanya dukungan teknologi *Next Generation Sequencing* (NGS) yang mampu menganalisis hubungan berbagai variasi genetik terhadap respon biologis maupun klinis. Namun demikian, penentuan metode analisis NGS perlu mempertimbangkan aspek ketersediaan sumber daya manusia, teknologi komputasi, dan biaya. Hal ini perlu dilakukan agar lebih efisien, rasional, dan sesuai dengan kebutuhan kontekstual, terutama di negara berkembang dengan akses sumber daya yang terbatas. Oleh sebab itu, pemahaman prinsip genomik dalam menginterpretasikan data NGS sangat penting agar diperoleh hasil yang bermakna. Setelah membaca bab ini hingga tuntas, penulis berharap agar pembaca mampu memahami konsep dasar, metode analisis genomik berbasis NGS, serta penerapannya pada praktik terapi individual, sehingga hasil penelitian dapat dipertanggung jawabkan secara valid.

Pendekatan NGS Dalam Studi Genom

Pendekatan NGS pada analisis genom mengidentifikasi varian penyebab penyakit, penemuan target obat baru, serta

fenomena jalur biologis yang kompleks dan variabel lain (Satam *et al.*, 2023). Perbedaan dalam pendekatan sekuensing akan memengaruhi kualitas, kuantitas, dan pilihan aplikasi sekuensing tersebut. Beberapa pendekatan NGS dalam studi genomik dijelaskan pada paragraf berikut.

Whole Genome Sequencing (WGS)

WGS digunakan saat data mekanisme biologis yang mendasari variabel sifat atau penyakit belum diketahui dengan jelas. Pendekatan ini menyajikan semua data materi genetik, baik di wilayah gen pengkode protein, maupun di pengaturan fungsi gen yang semuanya memiliki probabilitas netral, setara, dan berpotensi secara klinis. Hasil temuan WGS ini sangat melimpah dan dapat digunakan untuk merancang hipotesis baru yang bermakna secara biologis sebagai dasar analisis genomik lanjutan (Brlek *et al.*, 2024).

Whole Exome Sequencing (WES)

Perubahan fungsi dan penyakit sangat erat kaitannya dengan kelainan struktur protein. Analisis sequens WES berfokus pada pembacaan genom yang berfungsi secara langsung dalam pembentukan protein (*exome*). Hal ini dilatarbelakangi oloeh adanya perubahan biologis yang bermakna akibat perubahan genom yang berefek pada struktur dan fungsi protein (Brancato *et al.*, 2025). Oleh sebab itu, hasil data WES juga bersifat parsial dari keseluruhan struktur genom, karena tidak digunakan untuk mendeteksi variasi mekanisme regulasi gen. Namun demikian, WES tetap menyeimbangkan dua faktor utama, yaitu: cakupan biologis dan efisiensi analisis (Glotov *et al.*, 2023).

Targeted Gene Panel

Pendekatan analisis genom ini menargetkan beberapa panel gen spesifik yang sudah ditentukan sebelumnya secara selektif. Analisis *targeted gene panel* ini mengkonfirmasi fungsi gen yang terpilih saja, bukan menggali mekanisme baru di luar gen yang tidak sesuai. Hasil dari sekuensing ini tidak bisa mengeksporasi genom secara keseluruhan seperti pada WGS (Yu *et al.*, 2019).

Targeted Resequencing SNP atau Lokus Spesifik

Berbeda dengan *targeted gene panel* yang menganalisis sekumpulan gen, *targeted resequencing SNP* ini difokuskan pada satu atau beberapa genom tertentu (satu SNP, promoter, atau segmen gen) yang telah ditentukan sebelumnya. Tentu saja, pemilihan gen tersebut mempertimbangkan kedalaman dan relevansi biologis atau klinis, sehingga proses analisis akan lebih berfokus pada pengujian posisi dan pola variasi yang sudah diketahui (J. Zhang *et al.*, 2020).

Posisi NGS dalam Farmakogenomik

Ilmu farmakogenomik bertujuan untuk menilai hubungan antara variasi sekumpulan genetik terhadap respon biologis, klinis, terapi (absorpsi, distribusi, metabolisme, dan ekskresi), maupun keamanan obat. Berbeda dengan studi asosiasi genetik konvensional, NGS berperan dalam:

1. Mengidentifikasi varian langka (*rare variants*). Varian langka memiliki frekuensi rendah pada populasi, sehingga jarang dianalisis dengan pendekatan asosiasi umum. Namun demikian, varian ini dapat mempengaruhi secara signifikan terhadap efek biologis struktur dan

fungsi gen. Efek biologisnya yang kuat dapat berdampak pada perubahan fenotipe (Posey *et al.*, 2019).

2. Analisis multigen dan jalur biologis (*pathway-based analysis*). Analisis ini menekankan pada pendekatan interaksi berbagai gen yang berada pada jalur biologis yang sama. Perubahan fenotipe dipengaruhi bukan hanya satu gen tunggal saja, namun juga interaksi sejumlah gen yang saling mempengaruhi satu sama lain dalam jaringan dan jalur fungsional. Pendekatan ini berfokus pada fungsi kolektif dari jalur biologis, sehingga menghasilkan hipotesis genom yang lebih komprehensif dan kontekstual (Zhao *et al.*, 2015).
3. Pendekatan pengobatan presisi (*precision medicine*) berbasis populasi lokal. Dinamika genetik populasi menandakan bahwa variasi genetik dan respon biologis tidak bersifat universal. Pengobatan presisi berbasis populasi lokal sangat erat kaitannya dengan pola variasi genetik yang spesifik dan berbeda khas di setiap populasi tersebut. Adaptasi dan penyesuaian respon biologis dan klinis perlu mempertimbangkan beberapa faktor khusus seperti frekuensi varian, struktur genetik, serta latar belakang biologis lokal pada populasi setempat (Igl *et al.*, 2009; Schwarz *et al.*, 2019).
4. Metode *targeted sequencing* dan *whole exome sequencing* (WES) menjadi pilihan yang paling relevan secara kontekstual untuk diaplikasikan pada negara berkembang dengan keterbatasan sumber daya pada cakupan gen, biaya, dan interpretabilitas klinis (Schwarze *et al.*, 2018).

Desain Studi Analisis Genomik

Desain studi farmakogenomik perlu ditentukan pada tahap persiapan atau pra-analisis. Metode studi harus sesuai dengan rumusan dan tujuan luaran biologis dan klinis yang diharapkan. Beberapa pertimbangan hal dalam penyusunan desain studi farmakogenomik antara lain:

1. Fenotipe, merupakan variabel yang diukur sebagai hasil interaksi genetik terhadap respon biologis. Fenotipe diterapkan sebagai dasar studi genomik, sehingga harus ditentukan secara jelas dan terukur (Mori *et al.*, 2009).
2. Gen kandidat. Pemilihan gen kandidat dalam studi farmakogenomik didasarkan pengaruh gen pada jalur biologis yang paling signifikan. Secara teoretis, pemilihan gen kandidat bukan semata pada ketersediaan data genetik, namun pada gen dengan fungsi jalur biologis yang kuat sebelumnya. Pendekatan gen kandidat akan memungkinkan interpretasi yang lebih terarah dan relevan. Pemilihan gen kandidat juga menjadi bagian terpenting analisis genom agar lebih selektif, bukan eksploratif (Dickson *et al.*, 2010; Wang *et al.*, 2015).
3. Kualitas dan sumber DNA, sangat mempengaruhi hasil analisis genomik. DNA dapat diperoleh dari berbagai sumber seperti darah perifer, air liur, ataupun biobank. Masing-masing sumber DNA memiliki karakteristik biologis tertentu yang mempengaruhi teknik ekstraksi dan kualitas bahan awal studi genomik. DNA yang berasal dari darah perifer memiliki kualitas dan konsistensi lebih baik dibandingkan dari air liur atau koleksi biobank. Namun demikian, akses DNA dari koleksi biobank lebih mudah dan cakupannya lebih luas. DNA yang memiliki

kualitas baik juga menjadi syarat utama pra-analitik agar variasi genetik dapat terdeteksi dengan benar dan mewakili karakteristik individu (Peng *et al.*, 2024; Zhu *et al.*, 2015).

4. Ukuran sampel dan strategi replikasi yang sesuai merupakan dasar kekuatan analisis inferensi pada studi farmakogenomik. Kedua faktor tersebut dipengaruhi oleh tujuan analisis, distribusi fenotipe, dan pengaruh biologis yang diharapkan. Strategi replikasi juga didesain secara ilmiah untuk menentukan hubungan variasi genetik dan fenotipe yang konsisten, dan tidak bersifat kebetulan atau spesifik pada satu kelompok sampel saja. Replikasi yang sesuai menunjukkan konsistensi temuan pada set data atau populasi yang berbeda, sehingga meningkatkan keandalan interpretasi dan kesimpulan genomik yang valid (Igl *et al.*, 2009; Liu *et al.*, 2008).

Umumnya, kesalahan yang sering dijumpai pada tahap pra-analitis ini yaitu hanya mengandalkan panel gen tanpa justifikasi biologis atau klinis yang kuat.

Analisis Genom Berbasis NGS

Tahapan analisis genomik berbasis NGS dapat dijelaskan dibawah ini:

Quality Control (QC) Data Mentah

Tahap awal analisis genomik dilakukan dengan memeriksa kualitas data mentah NGS dalam bentuk file FASTQ. Kualitas data sekuensing yang dievaluasi secara cermat antara lain yaitu basa, adaptor, dan duplikasi. Hal ini dilakukan agar valid dan layak untuk dianalisis informasi genetik lebih lanjut. Sumber kesalahan

sekuensing dapat berasal dari kualitas pembacaan yang rendah, kontaminasi, atau bias teknis.

Tahap awal proses filter data FASTQ yaitu memeriksa kualitas pembacaan basa di sepanjang *read* dengan menganalisis distribusi skor *Phred*, kualitas basa di ujung *read*, dan variabilitas kualitas basa dan panjang antar *read*. Hasil pemeriksaan ini akan menentukan apakah data *file* membutuhkan proses trimming atau tidak. Selanjutnya, adaptor sekuensing yang masih menempel dan sekuens teknis non biologis diperiksa untuk mencegah varians palsu dan kerusakan *alignment*. Kualitas basa yang jelek, adaptor yang masih menempel, serta *read* di bawah panjang minimum perlu dihilangkan agar menghasilkan FASTQ yang bersih yang siap untuk dianalisis *alignment* (Guo *et al.*, 2014).

Alignment ke Genom Referensi

Analisis *alignment* dilakukan dengan memetakan hasil FASTQ sesuai urutan genom standar/referensi. Fragmen DNA disejajarkan pada posisi dan spesies yang paling sesuai di genom referensi, sehingga variasi genetik bisa diidentifikasi dengan menemukan perbedaan sampel dan referensi. Pada tahap ini, keakuratan proses *alignment* menjadi penentu dalam ketepatan penafsiran variasi genetik pada data mentah sampel. Konsistensi genom referensi sebagai standar baku sangat penting agar hasil *alignment* dapat diintegrasikan dengan genom antarindividu, data anotasi, dan studi lain yang lebih luas (Li *et al.*, 2016).

Variant Calling

Tahap ini bertujuan untuk menginferensikan *single nucleotide polymorphisms* (SNP) dan *insertions–deletions* (indel)

pada antara sampel dan genom referensi menggunakan algoritme yang valid. Varian SNP diinterpretasikan secara cermat untuk menghasilkan perbedaan biologis yang benar dan bermakna agar bisa dianalisis genomik tahap lanjutan. Oleh sebab itu, algoritme yang dipilih harus mampu membedakan variasi genetik yang nyata dari kesalahan pembacaan sekuens (Li *et al.*, 2016).

Algoritme *GATK HaplotypeCaller* merupakan contoh algoritme yang paling banyak digunakan dan menjadi standar pada beberapa studi genomik. Pendekatan berbasis *local re-assembly* yang digunakan pada *GATK HaplotypeCaller* bertujuan untuk meningkatkan akurasi deteksi varian, khususnya pada lokasi genom yang rumit (McKenna *et al.*, 2010).

Algoritme *FreeBayes* berbasis pemodelan variasi genetik juga banyak digunakan dalam studi populasi dan analisis multisaluran. Hal ini dikarenakan algoritme ini lebih fleksibel dalam menangani berbagai desain data (Bassano *et al.*, 2023). *SAMtools/BCFtools* berbasis pendekatan pileup juga sering digunakan sebagai pembanding atau alat validasi silang untuk mendeteksi perbedaan basa sampel dan genom referensi (Danecek *et al.*, 2021).

Dalam konteks klinis dan riset translasi, *VarScan* sering digunakan untuk mendeteksi varian dengan frekuensi alel rendah, terutama pada data dengan kedalaman sekuensing yang tinggi. Ketersediaan berbagai algoritme tervalidasi ini menggambarkan bahwa penggunaan algoritme harus diperkuat dengan bukti ilmiah: tidak ada satu metode algoritme tunggal yang sempurna. Keandalan *variant calling* didukung oleh penggunaan algoritme yang telah diuji secara luas dan diakui oleh komunitas genomik internasional (Koboldt *et al.*, 2009).

Filtering dan Quality Assurance Varian.

Hasil dari tahap *variant calling* dilakukan penyaringan dan kontrol kualitas agar memiliki keandalan yang adekuat dan bisa diproses ke analisis genomik tahap berikutnya. Proses penyaringan dan kontrol kualitas varian meliputi: kedalaman pembacaan (*depth*), skor kualitas varian, serta konsistensi sinyal sekuensing, yang digunakan untuk membedakan variasi genetik yang nyata dari kesalahan teknis. Selain itu, dilakukan juga analisis frekuensi varian dalam populasi. Hal ini bertujuan untuk mengelompokkan varian menjadi varian umum dan varian jarang (sesuai dengan tujuan analisis). Tahap *filtering* dan *quality assurance* ini memastikan bahwa varian memiliki dasar teknis yang kuat dan relevansi biologis yang lebih jelas, sehingga interpretasi genomik dapat dilakukan secara lebih andal (Florian *et al.*, 2023; Patel & Jain, 2012).

Anotasi, Asosiasi, dan Interpretasi Biologis Klinis

Tahap analisis selanjutnya adalah anotasi dan interpretasi biologis dengan menghubungkan varian genetik dengan konteks biologisnya. Varian SNP dianalisis untuk menentukan korelasi varian tersebut terhadap struktur dan regulasi gen, fungsi dan perubahan aktivitas protein, serta keterlibatannya dalam proses seluler tertentu. Selain menentukan lokasi dan sifat varian, anotasi juga ditujukan untuk memahami implikasi kombinasi varian terkait dengan perubahan fenotipe/luaran klinis seperti efektivitas terapi, keamanan, parameter farmakokinetik, dan farmakodinamik zat aktif (Nasr & Al-Mubaid, 2023; Pers *et al.*, 2015)

Varian SNP *missense* menyebabkan perubahan satu asam amino pada protein, sehingga mengganggu fungsi protein secara signifikan dari segi stabilitas, aktivitas, atau interaksi protein (Zhang *et al.*, 2012). Sementara itu, varian *nonsense* dapat memicu sinyal berhenti prematur dalam proses translasi, sehingga protein menjadi terpotong dan kehilangan fungsi biologisnya (Vitale *et al.*, 2025). Varian *splice-site* memengaruhi proses penyambungan RNA (*splicing*), yang menentukan bagaimana ekson dan intron disusun dalam mRNA matang. Gangguan pada *splicing* dapat menghasilkan protein dengan struktur yang berubah, hilangnya *domain* penting, atau bahkan kegagalan pembentukan protein yang fungsional (Blijlevens *et al.*, 2021; van den Hoogenhof *et al.*, 2016). Oleh karena itu, varian *missense*, *nonsense*, dan *splice-site* merupakan varian dengan potensi dampak fungsional yang berbeda yang penting dalam menilai konsekuensi biologis dan klinis dari variasi genetik.

Hasil anotasi dan asosiasi tersebut diintegrasikan ke dalam variabel klinis yang relevan, seperti usia, jenis kelamin, kondisi komorbid, dan terapi pendamping. Oleh sebab itu, analisis studi genomik dan klinis diharapkan dapat mendukung pengambilan keputusan klinis yang rasional dan utuh berbasis bukti.

Penutup

Analisis genomik berbasis NGS yang tepat dapat menghasilkan riset yang bermakna, relevan, dan berdaya saing global. Oleh sebab itu, NGS membutuhkan analisis genomik multidisiplin yang menuntut pemahaman biologis, bioinformatika agar dapat diintegrasikan pada konteks klinis.

DAFTAR PUSTAKA

Bassano, I., Ramachandran, V. K., Khalifa, M. S., Lilley, C. J., Brown, M. R., van Aerle, R., Denise, H., Rowe, W., George, A., Cairns, E., Wierzbicki, C., Pickwell, N. D., Carlile, M., Holmes, N., Payne, A., Loose, M., Burke, T. A., Paterson, S., Wade, M. J., & Grimsley, J. M. S. (2023). Evaluation of variant calling algorithms for wastewater-based epidemiology using mixed populations of SARS-CoV-2 variants in synthetic and wastewater samples. *Microbial Genomics, 9*(4), mgen000933. https://doi.org/10.1099/mgen.0.000933

Blijlevens, M., Li, J., & van Beusechem, V. W. (2021). Biology of the mRNA splicing machinery and its dysregulation in cancer providing therapeutic opportunities. *International Journal of Molecular Sciences, 22*(10), 5110. https://doi.org/10.3390/ijms22105110

Brancato, D., Treccarichi, S., Bruno, F., Coniglio, E., Vinci, M., Saccone, S., Calì, F., & Federico, C. (2025). NGS approaches in clinical diagnostics: From workflow to disease-specific applications. *International Journal of Molecular Sciences, 26*(19), 9597. https://doi.org/10.3390/ijms26199597

Brlek, P., Bulić, L., Bračić, M., Projić, P., Škaro, V., Shah, N., Shah, P., & Primorac, D. (2024). Implementing whole genome sequencing (WGS) in clinical practice: Advantages, challenges, and future perspectives. *Cells, 13*(6), 504. https://doi.org/10.3390/cells13060504

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience, 10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Florian, K., Benet-Pagès, A., Berner, D., Teubert, A., Eck, S., Arnold, N., Bauer, P., Begemann, M., Sturm, M., Kleinle, S., Haack, T. B., & Eggermann, T. (2023). Quality assurance within the context of genome diagnostics (A German perspective). *Medizinische Genetik, 35*(2), 91–104. https://doi.org/10.1515/medgen-2023-2028

Glotov, O. S., Chernov, A. N., & Glotov, A. S. (2023). Human exome sequencing and prospects for predictive medicine: Analysis of international data and own experience. *Journal of Personalized Medicine, 13*(8), 1236. https://doi.org/10.3390/jpm13081236

Guo, Y., Ye, F., Sheng, Q., Clark, T., & Samuels, D. C. (2014). Three-stage quality control strategies for DNA re-sequencing data. *Briefings in Bioinformatics, 15*(6), 879–889. https://doi.org/10.1093/bib/bbt069

Igl, B.-W., Konig, I. R., & Ziegler, A. (2009). What do we mean by "replication" and "validation" in genome-wide association studies? *Human Heredity, 67*(1), 66–68. https://doi.org/10.1159/000164400

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., Weinstock, G. M., Wilson, R. K., & Ding, L. (2009). VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics, 25*(17), 2283–2285. https://doi.org/10.1093/bioinformatics/btp373

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Mori, S., Chang, J. T., Andrechek, E. R., Potti, A., & Nevins, J. R. (2009). Utilization of genomic signatures to identify phenotype-specific drugs. *PLOS ONE, 4*(8), e6772. https://doi.org/10.1371/journal.pone.0006772

Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: A toolkit for quality control of next generation sequencing data. *PLOS ONE, 7*(2), e30619. https://doi.org/10.1371/journal.pone.0030619

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G., & Malonia, S. K. (2023). Next-generation sequencing technology: Current trends and advancements. *Biology, 12*(7), 997. https://doi.org/10.3390/biology12070997

Schwarz, U. I., Gulilat, M., & Kim, R. B. (2019). The role of next-generation sequencing in pharmacogenetics and pharmacogenomics. *Cold Spring Harbor Perspectives in Medicine, 9*(2), a033027. https://doi.org/10.1101/cshperspect.a033027

Schwarze, K., Buchanan, J., Taylor, J. C., & Wordsworth, S. (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genetics in Medicine, 20*(10), 1122–1130. https://doi.org/10.1038/gim.2017.247

BAB 10
Transkriptomik dan Analisis Ekspresi Gen

Chindy Nur Rosmeita

Multiomik telah berevolusi dari tren penelitian laboratorium menjadi pilar fundamental dalam transformasi sistem biologi dan kedokteran modern dalam beberapa tahun terakhir. Pendekatan ini mengintegrasikan data biologis yang beragam, termasuk genomik, transkriptomik, proteomik, metabolomik, dan *omics* lainnya untuk memberikan gambaran komprehensif dan akurat dalam identifikasi *biomarker*, pemahaman penyakit, dan penentuan terapi yang lebih tepat.

Dalam hierarki biologi molekuler, genomik berperan sebagai fondasi yang memetakan potensi genetik dan variasi yang mendasari risiko penyakit pada organisme. Sebagai sebuah cetak biru DNA, genomik menyediakan informasi statis dan stabil, sedangkan transkriptomik (regulasi ekspresi gen), proteomik (protein yang disintesis oleh sel), dan metabolomik (aktivitas metabolit yang dihasilkan dari reaksi biokimia seluler) lebih dinamis. Konsep ini terus berubah sebagai respons terhadap aktivitas biologis.

Tanskriptomik memegang peran krusial di antara berbagai pendekatan *omics*. Perannya sebagai jembatan antara potensi genetik dan fungsi biologis. Oleh karena itu, karakteristik data

transkriptomik yang kompleks dan dinamis tidak hanya menimbulkan tantangan analitis, tetapi juga menawarkan peluang signifikan untuk pengembangan dan penerapan bioinformatika modern.

Pendahuluan Transkriptomik

Definisi transkriptomik dan transkriptom

Transkriptomik merupakan studi mengenai RNA yang tidak terpisahkan dari istilah dogma sentral yang dicetuskan oleh Francis Crick pada 1958, serta menjadi salah satu bidang yang berkembang pesat di era pasca-genomik. Transkriptomik berkaitan erat dengan proses ekspresi gen dan protein yang dihasilkannya. Pendekatan ini digunakan untuk memprediksi perubahan pada tingkat protein dan aktivitas biologisnya. Secara definisi, transkriptom adalah kumpulan lengkap semua transkrip RNA, termasuk RNA pesan (mRNA), dan *non-coding* RNA (Dong *et al*, 2013).

Transkriptom merupakan kumpulan berbagai jenis RNA yang ada di dalam sel. Hasil analisis transkriptom digunakan untuk memahami perkembangan dan patogenesis penyakit. Secara umum, transkriptomik memiliki tiga tujuan utama, yaitu: (1) untuk mendata semua jenis transkrip (mRNA dan *non-coding* RNA); (2) untuk menentukan struktur transkripsi detail gen, termasuk identifikasi situs inisiasi, ujung 5′ dan 3′, pola *splicing*, dan modifikasi pasca-transkripsi; dan (3) untuk mengukur perubahan dinamis tingkat ekspresi setiap transkrip selama tahap perkembangan dan kondisi biologis yang berbeda (Wang, 2008).

mRNA dan non-coding RNA

Dalam studi transkriptomik, RNA seluler diklasifikasikan menjadi dua kelompok berdasarkan perannya dalam sintesis protein: mRNA dan *non-coding* RNA. Tahap pertama dalam proses ekspresi gen ialah mRNA yang membawa informasi genetik untuk proses pembentukan protein (translasi), sedangkan *non-coding* RNA tidak diterjemahkan tetapi berperan dalam regulasi ekspresi gen.

Non-coding RNA terlibat dalam regulasi transkripsi dan translasi, pengaturan epigenetik, modifikasi pasca transkripsi, interaksi protein, dan perekrutan kofaktor. *Non-coding* RNA diklasifikasikan berdasarkan panjang nukelotida: *small non-coding* RNA (sncRNA) (termasuk *microRNA* (miRNA) dengan panjang kurang dari 200 nukleotida) dan *long non-coding* RNA (lncRNA) (dengan panjang lebih dari 200 nukleotida). Bruxel *et al* (2021) menyatakan bahwa satu molekul miRNA dapat mengatur lebih dari seribu transkrip dan mengontrol lebih dari 50% protein, sehingga berpengaruh penting dalam regulasi gen.

Ribosomal RNA (rRNA) dan *transfer* RNA (tRNA) merupakan bagian dari kelompok *non-coding* RNA yang memainkan peran esensial dalam translasi protein. rRNA adalah jenis RNA yang paling melimpah di dalam sel. RNA ini membentuk ribosom bersama protein ribosom untuk mengikat mRNA dan mengkatalisis perakitan rantai polipeptida dari asam amino. Sementara itu, tRNA berfungsi membawa asam amino dan mengandung antikodon untuk menerjemahkan kode genetik pada translasi mRNA (Bruxel *et al*, 2021; Li & Liu, 2019)

Hubungan transkripsi dengan ekspresi gen

Ekspresi gen adalah proses biologis yang melibatkan penggunaan informasi genetik di dalam DNA untuk

menghasilkan produk protein fungsional. Proses ini terjadi melalui dua tahap utama, yaitu transkripsi dan translasi. Transkripsi merupakan tahap krusial dalam ekspresi gen yang mengatur aktivasi dan inaktivasi gen, sehingga menentukan kondisi fungsional suatu sel. Misalnya, perubahan ekspresi gen CFTR yang menyebabkan protein CFTR tidak berfungsi dengan baik dapat mengakibatkan penyakit *cytic fibrosis*. Hal ini menunjukkan bagaimana jalur dari DNA ke protein dapat mempengaruhi fenomena penyakit dalam kehidupan (Guo, 2014; Young, 2011).

Tingkat ekspresi gen berkorelasi dengan laju transkripsi karena jumlah mRNA yang dihasilkan mencerminkan aktivitas gen yang diekspresikan. Gen dengan laju transkripsi tinggi akan menghasilkan mRNA dalam jumlah besar dan memiliki potensi yang lebih besar untuk produksi protein. Oleh karena itu, kelimpahan mRNA berfungsi sebagai indikator ekspresi gen (Albert *et al*, 2002; Wang *et al*, 2008).

Regulasi transkripsi dan pasca transkripsi

Ekspresi gen dikendalikan melalui dua mekanisme regulasi, yaitu regulasi transkripsi dan regulasi pasca-transkripsi. Regulasi transkripsi mengatur aktivasi atau represi gen melalui interaksi dengan faktor transkripsi, elemen regulator DNA (*promoter*, *enchancer*, *operator*, dan *silencer*), dan kondisi kromatin (modifikasi histon dan metilasi DNA). Modifikasi histon seperti asetilasi dan metilasi memfasilitasi atau menghambat pengikatan faktor transkripsi, yang pada akhirnya memodulasi ekspresi gen (Lee & Young, 2019). Mekanisme ini menentukan apakah suatu gen akan ditranskripsi dan jumlah RNA yang dihasilkan, sehingga mempengaruhi tingkat ekspresi gen awal dan fungsi seluler.

Regulasi pasca-transkripsi terjadi setelah sintesis mRNA dan menyempurnakan ekspresi gen melalui *splicing* alternatif, stabilitas dan degradasi mRNA, serta interaksi dengan *non-coding* RNA (miRNA dan lncRNA) (Bartel, 2018; Chen & Shyu, 2017). Dengan demikian, ekspresi gen merupakan hasil integrasi antara regulasi transkripsi yang menentukan produksi RNA dan regulasi pasca transkripsi yang mengontrol pemrosesan dan pemanfaatan RNA.

Teknologi *Profiling* Transkriptomik

Teknologi transkriptomik bertujuan untuk mengidentifikasi dan mengukur seluruh RNA yang diekspresikan. Secara umum, teknologi transkriptomik berkembang pesat dari metode berbasis hibridisasi (*Microarray*) hingga metode berbasis sekuensing beresolusi tinggi (RNA-seq).

Microarray

Microarray merupakan teknologi transkripktomik generasi awal yang memungkinkan analisis simultan ribuan ekspresi gen menggunakan hibridisasi antara mRNA (atau cDNA) dan *probe* DNA spesifik yang terikat pada chip. Intensitas fluoeresensi yang dihasilkan mencerminkan tingkat ekspresi gen. *Microarray* secara luas diterapkan untuk ekspresi gen berskala besar (Schena *et al*, 1995)

RNA Sequencing (RNA-seq)

RNA *sequencing* (RNA-seq) berbeda dengan *microarray* (lihat tabel 10.1). Hal ini dikarenakan RNA-seq menggunakan teknologi *Next Generation Sequencing* (NGS) untuk mendeteksi, mengidentifikasi, dan menghitung jumlah molekul RNA. NGS

dapat mengurutkan jutaan hingga miliaran fragmen asam nukleat secara paralel dalam satu kali proses sekuensing. Dalam RNA-seq, molekul RNA diisolasi dan dikonversi menjadi fragmen cDNA untuk meningkatkan stabilitas selama sekuensing (Lowe *et al*, 2017). Aplikasi utama RNA-seq adalah analisis perbedaan ekspresi gen antara jaringan sehat dan kanker dengan mengembangkan *biomarker* spesifik untuk diagnosis dan terapi yang lebih presisi.

Berdasarkan panjang fragmen, teknologi NGS dikategorikan menjadi *long-read sequencing* dan *short-read sequencing*. *Long-read sequencing* digunakan pada platform Oxford Nanopore atau PacBio, yang memungkinkan pengurutan fragmen DNA hingga 1-50 kb. Hal ini memudahkan dalam mendeteksi *genome-wide repeat* pada genom dan variasi struktural yang kompleks. *Short-read sequencing* dapat mengurutkan fragmen DNA berkisar antara 50-400 bp. Salah satu platform terkenal yang menyediakan teknologi ini adalah Illumina dan mempunyai akurasi yang lebih tinggi (Verma *et al*, 2025).

Single-cell RNA-seq (scRNA-seq) ialah pengembangan teknologi RNA-seq berbasis NGS yang dirancang untuk menganalis transkriptom pada tingkat sel tunggal. Dalam proses ini, setiap molekul RNA diberi penanda unik berupa *cell-barcode* sebelum proses sekuensing, sehingga dapat membedakan asal molekul dari setiap sel. scRNA-seq menyediakan profil ekspresi gen beresolusi tinggi pada tingkat sel tunggal, tetapi tidak menyediakan informasi mengenai posisi sel dalam jaringan. Teknologi transkriptomik spasial mengatasi keterbatasan tersebut dengan mempertahankan informasi lokasinya. Transkriptomik spasial adalah pendekatan yang menganalisis ekspresi gen dengan mempertahankan informasi lokasi sel di

dalam jaringan. Pendekatan ini digunakan untuk mempelajari heterogenitas jaringan, interaksi antar sel, dan mikro-lingkungan jaringan. Oleh karena itu, data scRNA-seq sering digunakan sebagai referensi untuk mengidentifikasi dan memetakan tipe sel pada data transkriptomik spasial.

Small RNA-seq adalah teknik sekuensing berkapasitas tinggi untuk menganalisis RNA kecil dengan panjang 18-30 nukleotida. Molekul RNA kecil ini mencakup beberapa *regulator* penting ekspresi gen, seperti *microRNA* (miRNA), *small interfering RNA* (siRNA), *piwi-interacting RNA* (piRNA), dan *non-coding* RNA kecil lainnya yang terlibat dalam berbagai proses seluler.

Tabel 10.1. Perbandingan Metode Transkriptomik

Metode	Mircoarray	RNA-seq
Kapasitas	Lebih Tinggi	Tinggi
Jumlah input RNA	Tinggi ~1 µg dari total RNA	Rendah ~1 ng dari total RNA
Intensitas persiapan sampel	Rendah	Tinggi
Pengetahuan sebelumnya	Transkrip referensi diperlukan untuk probe	Tidak diperlukan
Akurasi kuantifikasi	>90% (dibatasi oleh akurasi deteksi *fluoresence*)	~90%
Resolusi sekuens	*Array* khusus dapat mendeteksi varian *splicing* (terbatas oleh desain probe dan hibridisasi silang)	Dapat mendeteksi SNP dan varian *splicing* (terbatas oleh akurasi sekuensing ~90%)

Sensitivitas	10^{-3} (terbatas oleh deteksi *fluoresence*)	10^{-6} (terbatas oleh *coverage* dari sekuens)
Dynamic range	10^{3}-10^{4} (terbatas oleh deteksi *fluoresence*)	$>10^{5}$ (terbatas oleh *coverage* dari sekuens)
Reproduksibilitas teknis	>99%	>99%

Pengumpulan Data Transkriptomik

Data transkrip RNA diperoleh melalui dua prinsip utama, yaitu penyusunan sekuens transkrip individu seperti *Expressed Sequence Tag* (EST) atau RNA-Seq, serta hibridisasi transkrip ke dalam probe nukleotida seperti *microarray*.

Isolasi RNA

Setiap metode transkriptomik memerlukan isolasi RNA sebagai langkah awal untuk mendapatkan data transkriptomik yang dibutuhkan. RNA yang diisolasi mengandung 98% RNA ribosom (rRNA), sehingga diperlukan *probe* dengan sekuens spesifik atau metode afinitas poly-A untuk mendapatkan jumlah mRNA yang lebih tinggi (Zhao *et al*, 2014).

EST, SAGE, dan CAGE

Expressed Sequence Tag (EST) adalah sekuens nukleotida pendek berukuran 200-800 bp yang dihasilkan dari transkrip RNA tunggal dan diubah menjadi cDNA oleh enzim transkriptase. cDNA akan memberikan informasi tentang EST untuk mengidentidikasi transkrip gen, penemuan gen, dan penentuan sekuens. EST digunakan untuk mengidentifikasi gen yang

ditranskripsikan tanpa memerlukan pengetahuan sebelumnya mengenai kandungan gen atau genom yang telah diurutkan.

Serial Analysis of Gene Expression (SAGE) merupakan pengembangan dari metode EST dengan bantuan *tag* berkapasitas tinggi untuk kuantifikasi. Pada metode SAGE, cDNA yang dihasilkan dari mRNA dipotong menggunakan enzim restriksi spesifik, sehingga menghasilkan fragmen berukuran 10-11 bp. Selanjutnya, *tag* ini digabungkan dan diurutkan menjadi *long strand* (>400-600 bp), serta disejajarkan dengan gen referensi untuk mengidentifikasi gen yang sesuai. *Cap Analysis of Gene Expression* (CAGE) adalah jenis dari SAGE yang mengurutkan *tag* mulai dari ujung 5' pada transkrip mRNA dan digunakan untuk analisis promotor atau kloning cDNA. SAGE dan CAGE menghasilkan lebih banyak informasi tentang gen dibandingkan dengan EST (Lowe *et al*, 2017; Magar *et al*, 2022).

Microarrays

Microarray terdiri dari oligomer nukleotida pendek yang disebut *probe*. Jumlah transkrip yang tinggi ditentukan oleh proses hibridisasi transkrip yang diberi label *fluoresence* dengan *probe*. Metode *microarray* memerlukan pengetahuan sebelumnya mengenai organisme yang diteliti untuk membuat *probe*, sehingga tidak efektif digunakan untuk gen yang memiliki ekspresi sangat tinggi atau rendah. Selain itu, hibridisasi sillang antar sekuens dapat mungkin terjadi dan menyebabkan penurunan deteksi yang tepat. Hasil dari *microarray* perlu divalidasi menggunakan qRT-PCR, *northen blot*, atau immunohistokimia menggunakan referensi gen yang sesuai (Lowe *et al*, 2017; Romanov *et al*, 2014).

RNA-Seq

RNA-Seq mengacu pada kuantifikasi transkriptom menggunakan NGS dan metode komputasi. Metode ini digunakan untuk membaca banyak fragmen pendek RNA, yang kemudian dianalisis secara komputasional untuk menyusun kembali transkrip RNA asli dan dicocokan dengan genom referensi atau dirakit tanpa genom acuan (perakitan *de novo*).

RNA-Seq memiliki rentang pengukuran yang sangat luas (sekitar 5 orde magnitude), sehingga lebih unggul dibandingkan *microarray*. Metode ini memerlukan jumlah RNA yang sangat kecil (nanogram) dan memungkinkan hasil analisis yang lebih detail, termasuk sel tunggal dengan amplifikasi cDNA. RNA-Seq dapat mengidentifikasi gen dalam genom dan menentukan aktivasi gen pada waktu tertentu. Jumlah pembacaan RNA memberikan pengukuran dengan tingkat akurasi tinggi (Lowe *et al*, 2017).

Analisis Data Transkriptomik

Alur pengolahan data transkriptomik mencakup *quality control* (QC), *Sequence Alignment*, kuantifikasi ekspresi, penyimpanan data, dan anotasi data. Tahap QC bertujuan untuk menilai kualitas *raw data* hasil sekuensing. Contohnya misalnya menggunakan FastQC untuk menilai kualitas sekuens, kandungan GC, tingkat duplikasi, distribusi panjang, kandungan K-mer, dan kontaminasi *adaptor*. Proses *trimming* berguna dalam tahapan ini untuk menghilangkan bacaan berkualitas rendah dan sekuens adaptor yang terkontaminasi (terjadi ketika panjang sekuens DNA lebih panjang dari insersi DNA).

Selanjutnya, *Sequence Alignment* menggunakan STAR atau HISAT2 untuk memetakan sekuens pendek (*short reads*) ke

genom referensi menghasilkan *file* berformat BAM atau SAM. Tingkat ekspesi gen diukur menggunakan TPMCalculator atau HTSeq untuk menghasilkan matriks hitung ekspresi gen. Data ekspresi gen disajikan dalam format *Fragments Per Kilobase of transcript per Million mapped reads* (FPKM) dan *Transcripts Per Million* (TPM). Sementara itu, data transkriptomik sel tunggal tersedia dalam format FASTQ, BAM, dan CSV, yang menyediakan profil ekspresi gen pada tingkat sel individu, heterogenitas seluler, serta proses biologis kompleks.

Penyimpanan atau pengarsipan data dilakukan dengan menempatkan data yang telah diproses ke dalam basis data publik, seperti *Sequence Read Archive* (SRA) di NCBI untuk keperluan analisis lanjutan. Anotasi data dilakukan dengan memanfaatkan basis data fungsi gen, seperti Ensembl dan RefSeq. Hal ini dilakukan untuk memberikan infromasi mengenai fungsi gen, menganalisis pola ekspresi, serta mengidentifikasi gen yang diekspresikan secara berbeda (DEGs).

DAFTAR PUSTAKA

Albert, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *From DNA to RNA*. Garland Science.

Bartel, D. P. (2018). Metazoan microRNAs. *Cell, 173*(1), 20–51. https://doi.org/10.1016/j.cell.2018.03.006

Bruxel, E. M., Bruno, D. C., Canto, A. M. D., Geraldis, J. C., Godoi, A. B., Martin, M., & Lopes-Cendes, I. (2021). Multi-omics in mesial temporal lobe epilepsy with hippocampal sclerosis: Clues into the underlying mechanisms leading to disease. *Seizure, 90*, 34–50. https://doi.org/10.1016/j.seizure.2021.05.017

Dong, Z., & Chen, Y. (2013). Transcriptomics: Advances and approaches. *Science China Life Sciences, 56*(10), 960–967. https://doi.org/10.1007/s11427-013-4551-8

Gehring, N. H., Wahle, E., & Fischer, U. (2017). Deciphering the mRNP code: RNA-bound determinants of post-transcriptional gene regulation. *Trends in Biochemical Sciences, 42*(5), 369–382. https://doi.org/10.1016/j.tibs.2017.02.004

Guo, J. (2014). Transcription: The epicenter of gene expression. *Journal of Zhejiang University Science B, 15*(5), 409–411. https://doi.org/10.1631/jzus.B1400099

Karahalil, B. (2016). Overview of systems biology and omics technologies. *Current Medicinal Chemistry, 23*(37), 4221–4230. https://doi.org/10.2174/0929867323666160920110627

Lee, T. I., & Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell, 152*(6), 1237–1251. https://doi.org/10.1016/j.cell.2013.02.014

Li, J., & Liu, C. (2019). Coding or noncoding, the converging concepts of RNAs. *Frontiers in Genetics, 10*, 496. https://doi.org/10.3389/fgene.2019.00496

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLOS Computational Biology, 13*(5), e1005457. https://doi.org/10.1371/journal.pcbi.1005457

Magar, N. D., Shah, P., Harish, K., Bosamia, T. C., Barbadikar, K. M., Shukla, Y. M., Phule, A., Zala, H. N., Madhav, M. S., Mangrauthia, S. K., Neeraja, C. N., & Sundaram, R. M. (2022). Gene expression and transcriptome sequencing: Basics, analysis, advances. In *Gene expression*. IntechOpen.

Romanov, V., Davidoff, S. N., Miles, A. R., Grainger, D. W., Gale, B. K., & Brooks, B. D. (2014). A critical comparison of protein microarray fabrication technologies. *The Analyst, 139*(6), 1303–1326. https://doi.org/10.1039/C3AN02133K

Verma, R., Savaria-Butler, A., Enguita, F. J., & Meller, R. (2025). Commentary: A review of technical considerations for planning an RNA-sequencing experiment. *BMC Genomics, 26*(1), 918.

Wang, Z., & Burge, C. B. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA, 14*(5), 802–813. https://doi.org/10.1261/rna.876308

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics, 10*(1), 57–63. https://doi.org/10.1038/nrg2484

Young, R. A. (2011). Control of the embryonic stem cell state. *Cell, 144*(6), 940–954. https://doi.org/10.1016/j.cell.2011.01.032

Zhao, W., He, X., Hoadley, K. A., Parker, J. S., Hayes, D. N., & Perou, C. M. (2014). Comparison of RNA-seq by poly(A) capture,

ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics, 15*(1), 419. https://doi.org/10.1186/1471-2164-15-419

BAB 11
Metagenomik dan Mikrobioma

Eko Kusumawati

Definisi Metagenomik dan Mikrobioma

Metagenomik dan mikrobioma merupakan dua konsep kunci dalam bioinformatika modern yang menggeser fokus riset: dari mikroba yang bisa dikultur, menjadi mikroba yang benar-benar ada di suatu ekosistem. Mikrobioma kini didefinisikan sebagai keseluruhan habitat (lingkungan) beserta komunitas mikroba, termasuk *mobile genetic elements* (virus), DNA ekstraseluler, serta interaksi antar-komponen dalam ekosistem tersebut. Definisi ini juga menegaskan pembedaan istilah mikrobiota (organismenya) versus mikrobioma (Berg *et al.*, 2020). Sementara itu, metagenomik merujuk pada kajian berbasis sekuensing dan analisis komputasional terhadap metagenom, yakni kumpulan genom dan gen dari anggota mikrobiota dalam suatu sampel, sehingga peneliti dapat memetakan komposisi dan potensi fungsional komunitas mikroba tanpa harus mengisolasi satu per satu (Berg *et al.*, 2020; N. Kim *et al.*, 2024). Perkembangan *genome-resolved metagenomics* memperkuat arah ini dengan merekonstruksi genom mikroba dari data komunitas untuk memahami variasi dan kapasitas fungsional pada tingkat yang mendekati genom isolat, namun tetap dalam konteks ekologi komunitas (N. Kim *et al.*, 2024).

Perbedaan Metagenomik, Mikrobiologi Klasik, dan Genomik Isolat

Perbedaan utama metagenomik, mikrobiologi klasik, dan genomik isolat dapat dilihat pada objek analisis, cara memperoleh data, dan keluaran ilmiah. Mikrobiologi klasik bertumpu pada kultur untuk mengamati fenotipe, fisiologi, dan uji kausalitas. Bagaimanapun, mikrobiologi klasik secara historis dibatasi oleh *culture bias* banyak mikroba tidak tumbuh pada kondisi laboratorium standar, sehingga keragaman komunitas sering tereduksi. Genomik isolat (sekuensing satu strain murni) memberi resolusi tinggi pada genom organisme tertentu, sangat kuat untuk penetapan gen/varian dan eksperimen validasi, tetapi bergantung pada keberhasilan isolasi. Sebaliknya, metagenomik membaca DNA komunitas secara langsung, sehingga mampu menangkap mikroba yang sulit/tidak dapat dikultur, menggambarkan struktur komunitas dan *repertoar* gen kolektif. Bagaimanapun, interpretasinya memiliki tantangan komputasional seperti kompleksitas campuran spesies, rekonstruksi genom dari komunitas, serta kebutuhan kehati-hatian dalam inferensi kuantitatif berbasis kelimpahan relatif (N. Kim *et al.*, 2024; Plomp *et al.*, 2025). Literatur terbaru juga menekankan bahwa pendekatan kultur modern (*culturomics*) bukan lawan metagenomik, melainkan pelengkap untuk menghubungkan hasil profiling komunitas dengan isolat nyata untuk uji fungsi dan kausalitas (Plomp *et al.*, 2025).

Peran Mikrobioma dalam Sistem Biologis

Dalam sistem biologis, mikrobioma dipahami sebagai komponen fungsional yang memengaruhi homeostasis, metabolisme, dan regulasi imun, sehingga perubahan komposisi

atau fungsi komunitas (disbiosis) dapat berasosiasi dengan berbagai kondisi kesehatan maupun penyakit. *Review* terbaru menunjukkan bahwa mikrobiota di berbagai lokasi tubuh (usus, oral, respiratori, dan kulit) hidup bersimbiosis dengan inang, berkontribusi pada pemeliharaan fungsi imun dan keseimbangan fisiologis, sementara disbiosis berpotensi terkait penyakit kardiovaskular, kanker, gangguan respirasi, dan kondisi inflamasi lainnya (Hou *et al.*, 2022). Selain itu, kajian lain menegaskan peran mikrobioma sebagai pengatur fisiologi inang dan sebagai sumber informasi penting untuk stratifikasi penyakit dan pengembangan intervensi berbasis komunitas (Shehata *et al.*, 2022). Perspektif yang lebih baru bahkan menekankan bahwa mikrobiota dapat membentuk fenotipe inang melalui jalur metabolik dan imun, serta dapat ditransmisikan secara sosial/lingkungan, sehingga mikrobioma tidak hanya dipandang sebagai komponen internal, tetapi juga bagian dari ekologi kesehatan (Sarkar *et al.*, 2024).

Jenis dan Pendekatan Studi Metagenomik

Jenis dan pendekatan studi metagenomik secara umum dapat dibedakan menjadi metagenomik berbasis *amplicon* dan *shotgun* metagenomik, yang masing-masing memiliki tujuan dan resolusi analisis yang berbeda. Sekuensing *amplicon* menargetkan gen penanda konservatif, seperti 16S rRNA untuk bakteri dan arkea, ITS untuk fungi, serta 18S rRNA untuk mikroorganisme eukariotik, sehingga efektif untuk memetakan komposisi dan struktur taksonomi komunitas mikroba dengan biaya relatif rendah dan alur analisis yang lebih sederhana. Pendekatan ini banyak digunakan dalam studi survei ekologi, perbandingan antar-lingkungan, serta pemantauan perubahan

komunitas mikroba akibat perlakuan atau gradien lingkungan. Namun, keterbatasan utamanya adalah resolusi taksonomi yang sering kali hanya sampai tingkat genus, serta minimnya informasi langsung terkait fungsi metabolik (Knight *et al.*, 2018).

Berbeda dari *amplicon*, *shotgun* metagenomik melibatkan sekuensing seluruh DNA tanpa penargetan gen tertentu. Hal ini memungkinkan analisis taksonomi resolusi tinggi (hingga spesies/*strain*) sekaligus kapasitas fungsional komunitas mikroba melalui identifikasi gen dan jalur metabolik. Pendekatan ini mendukung rekonstruksi genom mikroba (*metagenome-assembled genomes/MAGs*), analisis gen resistensi antimikroba, serta eksplorasi potensi biokimia komunitas secara komprehensif. Selain itu, studi metagenomik modern semakin berkembang ke arah pendekatan terintegrasi dengan *meta-omics* lain, seperti metatranskriptomik, metaproteomik, dan metabolomik, untuk menjembatani informasi dari tingkat potensi genetik menuju aktivitas biologis aktual. *Review* mutakhir menekankan bahwa pemilihan pendekatan *amplicon*, *shotgun*, atau integrasi *multi-omics* harus disesuaikan dengan pertanyaan riset, kompleksitas ekosistem, serta sumber daya analitik yang tersedia (Nam *et al.*, 2023).

Pengambilan Sampel dan Preparasi Data Metagenomik

Pengambilan sampel dalam metagenomik harus dipandang sebagai bagian dari desain eksperimen. Hal ini dikarenakan variasi pada tahap awal ini dapat mendominasi sinyal biologis yang ingin ditangkap. Praktik terbaik terbaru menekankan penetapan unit *sampling* yang jelas (seperti individu/lokasi/waktu), replikasi biologis yang memadai, pengacakan urutan pemrosesan, serta pencatatan metadata

(seperti suhu, pH, kondisi inang, paparan lingkungan). Hal ini dilakukan agar hasilnya dapat dibandingkan lintas studi dan dapat direproduksi. Standardisasi prosedur koleksi, penyimpanan, transportasi (seperti kontrol suhu, penggunaan *buffer* preservasi) merupakan hal yang penting untuk mencegah perubahan komposisi komunitas sebelum DNA diekstraksi (Bharti & Grimm, 2021; Gemmell *et al.*, 2024).

Untuk strategi *sampling* lingkungan dan biologis, pendekatan harus disesuaikan dengan heterogenitas ekosistem. Pada sampel lingkungan (tanah, sedimen, air, limbah), variasi spasial sering tinggi. Oleh karena itu, *sampling* perlu mempertimbangkan *grid/stratified sampling*, kedalaman, dan pengulangan lintas titik, sedangkan pada sampel biologis (seperti feses/usus, oral, dan kulit) fluktuasi temporal dan faktor inang (diet, obat, usia, dan kondisi klinis) menjadi sumber variasi utama. Hal ini dikarenakan desain longitudinal atau pengambilan berulang sering lebih informatif dibanding satu kali titik waktu. Dalam konteks aplikasi rutin (seperti *profiling* resistom/AMR pada air limbah), tinjauan sistematis menekankan pentingnya konsistensi volume/massa sampel, homogenisasi, dan kontrol prosedur agar variasi teknis tidak menutupi tren ekologi/epidemiologi yang dipantau (Bharti & Grimm, 2021; Davis *et al.*, 2023).

Tahap isolasi DNA total mikroba bertujuan untuk memperoleh DNA yang mewakili seluruh anggota komunitas. Bagaimanapun, pemilihan metode lisis (mekanik/kimia/enzimatik) dan *kit* ekstraksi dapat menghasilkan bias, karena mikroba memiliki karakter dinding sel yang berbeda (contoh: Gram-positif vs Gram-negatif). Bukti mutakhir pada profil mikrobioma feses menunjukkan bahwa metode ekstraksi

DNA dapat mengubah komposisi taksonomi yang terdeteksi dan bahkan memengaruhi kekuatan asosiasi dengan fenotipe (seperti antropometri/lifestyle), sehingga harmonisasi protokol ekstraksi menjadi syarat penting untuk studi komparatif. Studi optimasi ekstraksi pada matriks limbah/*piggery wastewater* juga menekankan perlunya evaluasi *yield*, kemurnian, dan potensi *selective recovery* antar-metode sebelum dipakai pada analisis metagenomik rutin (Fernández-Pato *et al.*, 2024; Muralidhar *et al.*, 2025).

Setelah DNA diperoleh dan disekuensing, preparasi data metagenomik dilakukan untuk memastikan data yang dianalisis benar-benar merefleksikan sinyal biologis, bukan artefak teknis. Tahap ini umumnya mencakup *quality control* (penilaian kualitas *read*/*base*), *trimming* adaptor dan basa berkualitas rendah, penghapusan duplikasi/eror tertentu, serta *host-genome removal* pada sampel berbasis inang agar pembacaan mikroba tidak tenggelam oleh DNA inang (Bharti & Grimm, 2021; Mak *et al.*, 2025).

Teknologi Sekuensing dalam Studi Metagenomik

Teknologi sekuensing dalam studi metagenomik saat ini didominasi oleh platform *short-read* Illumina dan platform *long-read* generasi ketiga seperti Oxford Nanopore Technologies (ONT) serta PacBio HiFi. Illumina menghasilkan *read* pendek (umumnya ratusan bp) dengan throughput sangat tinggi dan akurasi baik. Karena itulah Illumina dinilai efektif untuk *profiling* komunitas dan kuantifikasi kelimpahan pada studi berskala besar. Bagaimanapun, *read* pendek sering kesulitan merakit (*assembly*) wilayah genom kompleks seperti *repeat*, operon rRNA, atau elemen genetik bergerak, sehingga rekonstruksi genom

lengkap dari komunitas beragam menjadi menantang (Bharti & Grimm, 2021). Sebaliknya, ONT dan PacBio menyediakan *read* panjang (kilobase hingga puluhan kilobase) yang membantu melewati repeat dan meningkatkan *contiguity* assembly, sehingga memperkuat pendekatan *genome-resolved metagenomics* (seperti penyusunan MAG/cMAG). PacBio HiFi lebih unggul karena menggabungkan *read* panjang dengan akurasi sangat tinggi, yang terbukti dapat menghasilkan *assembly* metagenom berkualitas tinggi dan meningkatkan fraksi *contig* yang mendekati/menjadi sirkular pada komunitas tertentu (Benoit *et al.*, 2024; C. Y. Kim *et al.*, 2022). Dalam praktiknya, studi perbandingan juga menunjukkan bahwa penggabungan data *short-read* dan *long-read* (*hybrid*) sering memperluas deteksi variasi genetik dan meningkatkan resolusi genomik dibanding hanya salah satu tipe data (Chen *et al.*, 2022).

Perbandingan read pendek vs read panjang pada metagenomik dapat diringkas sebagai *trade-off* antara kedalaman/biaya/akurasi kuantifikasi (unggul pada *short-read*) versus kemampuan resolusi struktural & perakitan genom (unggul pada *long-read*). Tantangan teknis utama data metagenomik, baik pendek mapupun panjang, meliputi kompleksitas komunitas (banyak spesies/*strain* bercampur), ketidakmerataan cakupan, kontaminasi DNA inang, serta bias preparasi yang memengaruhi interpretasi kelimpahan dan fungsi. Pada *long-read*, tantangan tambahan yang sering dibahas adalah kebutuhan algoritme khusus untuk *binning* dan *assembly*, serta kendala kualitas/karakter eror (terutama pada ONT). Selain itu, faktor keterbatasan biaya untuk skala besar juga menjadi pertimbangan. *Review long-read* metagenomics menekankan bahwa isu-isu ini harus diatasi dengan *pipeline* komputasional

yang tepat dan strategi desain sekuensing yang sesuai tujuan (seperti resolusi genom vs survei populasi) (Bharti & Grimm, 2021; C. Kim *et al.*, 2024).

Analisis Keragaman Mikroba

Analisis keragaman mikroba merupakan pendekatan fundamental dalam metagenomik untuk memahami struktur, stabilitas, dan dinamika komunitas mikroba pada suatu ekosistem. Keragaman komunitas mencerminkan respons mikroba terhadap faktor lingkungan, kondisi inang, maupun tekanan antropogenik, sehingga sering digunakan sebagai indikator awal kesehatan atau gangguan ekosistem.

Dalam praktik bioinformatika modern, analisis keragaman menjadi tahapan eksploratif yang wajib sebelum analisis lanjutan seperti diferensial kelimpahan atau pemetaan fungsi (Knight *et al.*, 2018). Analisis keragaman berperan sebagai jembatan konseptual antara data sekuensing mentah dan interpretasi biologis berbasis ekologi mikroba.

Keragaman Alfa dan Beta

Keragaman alfa (*alpha diversity*) merepresentasikan variasi mikroba di dalam satu sampel, dengan menekankan dua komponen utama, yaitu kekayaan taksa (*richness*) dan pemerataan kelimpahan (*evenness*). Dalam studi metagenomik, keragaman alfa umumnya dihitung menggunakan indeks Shannon, Simpson, dan Chao1, yang masing-masing menangkap dimensi ekologi yang berbeda. Indeks Shannon sensitif terhadap perubahan keseluruhan komunitas, karena mempertimbangkan jumlah taksa dan distribusi kelimpahan. Indeks Simpson lebih menekankan dominansi taksa tertentu, sedangkan Chao1

digunakan untuk mengestimasi kekayaan taksa dengan mempertimbangkan keberadaan mikroba langka. Literatur metodologis menegaskan bahwa interpretasi keragaman alfa harus dilakukan secara hati-hati. Hal ini karena komponennya dipengaruhi oleh kedalaman sekuensing, desain *sampling*, dan metode normalisasi data (Willis, 2019). Pedoman analisis mikrobioma mutakhir juga menempatkan keragaman alfa sebagai indikator awal kesehatan dan stabilitas komunitas. Contohnya seperti pada mikrobioma usus atau ekosistem lingkungan. Bagaimanapun, pedoman tersebut juga menekankan bahwa indeks ini tidak dapat berdiri sendiri tanpa konteks biologis dan statistik yang memadai (Knight *et al.*, 2018).

Keragaman beta (*beta diversity*) digunakan untuk menilai perbedaan struktur komunitas mikroba antar-sampel atau antar-kelompok, sehingga menjadi komponen kunci dalam studi komparatif, seperti perbandingan kontrol versus perlakuan maupun variasi antar-lingkungan (Kers & Saccenti, 2022; Knight *et al.*, 2018). Secara operasional, keragaman beta dihitung dari matriks jarak/*dissimilarity* berbasis kelimpahan atau kehadiran taksa. Contohnya adalah Bray-Curtis (berbasis kelimpahan) dan Jaccard (berbasis kehadiran/ketiadaan), serta metrik filogenetik seperti UniFrac yang memasukkan informasi hubungan evolusioner antar-taksa (Lozupone & Knight, 2005). Hasil perhitungan jarak ini umumnya divisualisasikan melalui metode ordinasi multivariat seperti *Principal Coordinates Analysis* (PCoA) atau *Non-metric Multidimensional Scaling* (NMDS) untuk mengidentifikasi pola pengelompokan sampel dan gradien ekologis (Kers & Saccenti, 2022).

Indeks keragaman (Shannon, Simpson, Chao1)

Indeks keragaman Shannon (H') dan Simpson (D) adalah metrik kerangana alfa yang menangkap tidak hanya jumlah taksa (richness), tetapi juga pemerataan (*evenness*) kelimpahan relatif. Indeks Shannon menimbang kontribusi setiap takson melalui proporsi kelimpahannya (p_i), sehingga relatif sensitif terhadap perubahan komunitas secara menyeluruh, termasuk taksa berkelimpahan menengah. Indeks Simpson lebih menekankan dominansi taksa tertentu (probabilitas dua individu acak berasal dari takson yang sama), sehingga lebih peka terhadap perubahan pada taksa yang sangat melimpah. Dalam konteks data mikrobioma yang bersifat komposisional dan dipengaruhi kedalaman sekuensing, interpretasi Shannon–Simpson harus mempertimbangkan strategi pemrosesan data (misalnya rarefaction/normalisasi) agar perbandingan antar-sampel tidak bias (Willis, 2019).

Berbeda dari Shannon dan Simpson, Chao1 merupakan pengukur kekayaan taksa (*richness*) yang secara khusus memperhitungkan keberadaan taksa langka melalui jumlah *singletons* dan *doubletons*. Secara konsep, Chao1 menambahkan koreksi pada jumlah taksa teramati untuk memperkirakan taksa yang belum tertangkap akibat keterbatasan *sampling*/sekuensing (Chao, 1984). Karena berbasis taksa langka, Chao1 berguna pada ekosistem yang sangat kompleks (misalnya tanah atau sedimen) atau set data dengan banyak taksa berkelimpahan rendah, yang juga sensitif terhadap *noise* (misalnya kesalahan sekuensing yang tampak sebagai taksa langka). Oleh karena itu, praktik yang umum direkomendasikan adalah melaporkan Chao1 bersama Shannon/Simpson serta

menjelaskan langkah kontrol kualitas/*denoising* yang digunakan (Chao *et al.*, 2014; Willis, 2019).

Visualisasi Struktur Komunitas Mikroba

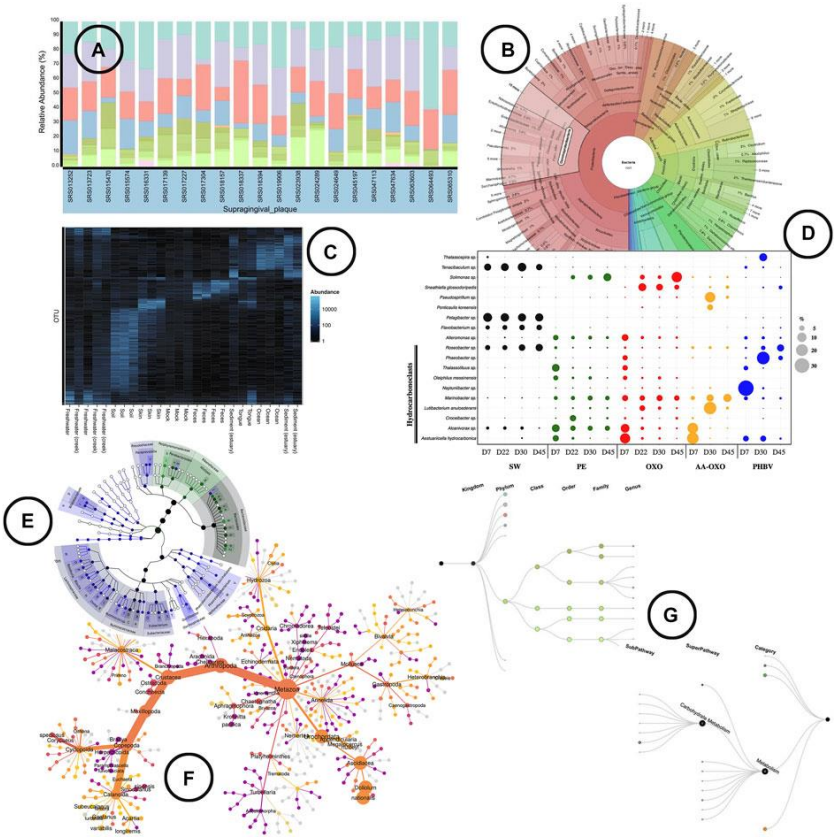
Visualisasi struktur komunitas mikroba merupakan komponen penting dalam analisis metagenomik. Hal ini dikarenakan data mikrobioma bersifat berdimensi tinggi dan sulit diinterpretasikan tanpa representasi grafis yang tepat. Visualisasi yang umum digunakan meliputi *barplot* komposisi taksonomi untuk membandingkan kelimpahan relatif mikroba antar-sampel atau antar-kelompok; *heatmap* yang dikombinasikan dengan *hierarchical clustering* untuk menyoroti pola ko-variasi dan taksa pembeda; serta metode ordinasi berbasis matriks jarak (seperti *Principal Coordinates Analysis* (PCoA) dan *Non-metric Multidimensional Scaling* (NMDS)) untuk menggambarkan perbedaan struktur komunitas antar-sampel. Literatur metodologis menegaskan bahwa ordinasi PCoA/NMDS berbasis metrik keragaman beta (misalnya Bray–Curtis atau UniFrac) sangat efektif dalam mengidentifikasi pengelompokan komunitas dan gradien lingkungan, asalkan interpretasinya disertai uji statistik pendukung (misalnya PERMANOVA) dan pelaporan parameter yang transparan (Knight *et al.*, 2018; Peeters *et al.*, 2021).

Penerapan visualisasi struktur komunitas mikroba juga ditunjukkan dalam penelitian Kusumawati *et al.* (2024). Peneliti mengkaji variasi fisikokimia perairan, keragaman, dan struktur komunitas bakteri pada kolam pascatambang batubara dengan umur berbeda di Samarinda, Kalimantan Timur, Indonesia. Dalam studi tersebut, struktur komunitas bakteri divisualisasikan melalui *barplot* komposisi taksonomi untuk menunjukkan pergeseran

dominansi bakteri seiring bertambahnya umur kolam pascatambang. Ordinasi NMDS/PCoA berbasis metrik jarak juga dilakukan untuk mengungkap pemisahan komunitas antar-kolam yang berkorelasi dengan perbedaan parameter fisikokimia (misalnya pH, konduktivitas, dan nutrien). Visualisasi ini memperlihatkan bahwa kolam dengan umur berbeda membentuk klaster komunitas bakteri yang terpisah. Hasil ini mengindikasikan proses suksesi mikroba dan adaptasi ekologis terhadap kondisi lingkungan pascatambang. Hal ini menunjukkan bahwa visualisasi mikrobioma juga berfungsi sebagai sarana integratif untuk mengaitkan data sekuensing dengan faktor lingkungan dan dinamika ekosistem perairan (Kusumawati *et al.*, 2024).

Gambar 11.1 menampilkan berbagai pendekatan visualisasi kelimpahan relatif komunitas mikroba. Panel (A) menunjukkan diagram batang bertumpuk yang dihasilkan menggunakan BURRITO (McNally et al., 2018), sedangkan panel (B) menampilkan diagram sunburst dari Krona yang menggambarkan hierarki taksonomi bakteri beserta kelimpahan relatifnya (Ondov et al., 2011). Kelimpahan OTU juga kerap divisualisasikan dalam bentuk heatmap menggunakan Phyloseq (McMurdie & Holmes, 2013) (C) atau direpresentasikan sebagai diagram gelembung (*bubble plot*) untuk menyoroti variasi kelimpahan antar-taksa (Dussud et al., 2018) (D). Selain itu, visualisasi berbasis pohon filogenetik banyak digunakan, seperti GraPhlAn yang memungkinkan penambahan anotasi visual pada struktur filogenetik (Asnicar et al., 2015) (E), serta *heat tree* yang menampilkan hierarki taksonomi dalam struktur pohon dengan lebar simpul (*node*) sebagai representasi kelimpahan OTU (Foster et al., 2017) (F). Lebih lanjut, BURRITO juga mampu menampilkan

hierarki taksa dan fungsi secara simultan dalam satu struktur pohon, dengan lebar simpul merepresentasikan tingkat kelimpahan (McNally et al., 2018) (G) (Peeters *et al.*, 2021).



Sumber: Peeters *et al.*, (2021)

Gambar 11.1 Visual untuk menampilkan kelimpahan (relatif) dan struktur hierarkis/relasional dalam analisis mikrobioma

Aplikasi Metagenomik dan Mikrobioma

Aplikasi metagenomik dan mikrobioma yang semakin berkembang pesat berada pada sektor pangan/fermentasi dan jaminan mutu–keamanan. Pada pangan fermentasi, metagenomik memungkinkan pemetaan komunitas mikroba

secara resolusi tinggi (hingga spesies/*strain*) sekaligus inferensi fungsi (misalnya gen metabolisme karbohidrat, biosintesis vitamin, dan pembentukan senyawa volatil). Akibatnya, produsen dan peneliti dapat menautkan "siapa yang ada" dengan "apa yang dikerjakan" dalam pembentukan mutu sensori dan keamanan produk. Tinjauan oleh Walsh *et al.*, (2023) menegaskan bahwa pendekatan molekuler terintegrasi (termasuk metagenomik) mengubah pemahaman tentang komunitas mikroba fermentasi dan implikasinya terhadap kualitas produk, serta interaksinya dengan mikrobioma manusia.

Review Sadurski *et al.*, (2024) menyoroti bahwa *food metagenomics* memperkuat sistem *traceability* dan *monitoring* kontaminan, sekaligus mempercepat identifikasi mikroba fungsional untuk optimasi proses produksi. Contoh penerapan multi-omics yang semakin sering dipakai di pangan fermentasi juga ditunjukkan oleh studi Kharnaior & Tamang, (2022). Penelitiannya mengombinasikan data metagenomik–metabolomik untuk mengaitkan taksa dominan dengan profil metabolit pada produk fermentasi tradisional, sehingga jalur fungsional dapat ditautkan langsung ke keluaran biokimia yang relevan terhadap mutu dan potensi kesehatan.

Di bidang lingkungan dan bioteknologi, metagenomik berperan strategis untuk bioremediasi dan bioprospeksi gen/enzim dari mikroba yang sulit atau tidak dapat dikultur. Dalam bioremediasi, metagenomik dipakai untuk mengidentifikasi konsorsium mikroba dan gen/jalur yang terkait degradasi polutan atau imobilisasi logam berat, sekaligus memantau perubahan struktur komunitas selama intervensi. Contoh studi terapan oleh Zhang *et al.*, (2025) yang menggunakan analisis metagenomik untuk mengevaluasi

pergeseran struktur komunitas dan fungsi pada upaya bioremediasi lahan pascatambang. Hal ini menunjukkan bagaimana pendekatan ini dapat menghubungkan tindakan restorasi dengan respons mikrobioma.

Pada sisi bioprospeksi, review Patel *et al.*, (2022) menegaskan bahwa *enzyme mining* berbasis metagenomik membuka akses ke keragaman genetik mikroba lingkungan untuk menemukan biokatalis baru (misalnya enzim hidrolitik dan enzim industri) yang tidak terjangkau melalui kultur konvensional. Dalam kerangka *One Health*, metagenomik (termasuk viral metagenomik NGS) juga makin penting untuk surveilans patogen. Hal ini sesuai dengan yang dirangkum oleh Russell *et al.*, (2025), yang menekankan kontribusi vmNGS untuk deteksi dan karakterisasi virus zoonotik pada antarmuka ekologi.

DAFTAR PUSTAKA

Benoit, G., Raguideau, S., James, R., Phillippy, A. M., Chikhi, R., & Quince, C. (2024). High-quality metagenome assembly from long accurate reads with metaMDBG. *Nature Biotechnology, 42*(9), 1378–1383. https://doi.org/10.1038/s41587-023-01983-6

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C. C., Charles, T., Chen, X., Cocolin, L., Eversole, K., Corral, G. H., Kazou, M., Kinkel, L., Lange, L., Lima, N., Loy, A., Macklin, J. A., Maguin, E., Mauchline, T., McClure, R., … Schloter, M. (2020). Microbiome definition re-visited: Old concepts and new challenges. *Microbiome, 8*(1), Article 103. https://doi.org/10.1186/s40168-020-00905-x

Bharti, R., & Grimm, D. G. (2021). Current challenges and best-practice protocols for microbiome analysis. *Briefings in Bioinformatics, 22*(1), 178–193. https://doi.org/10.1093/bib/bbz155

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics, 11*(4), 265–270.

Chao, A., Chiu, C. H., & Jost, L. (2014). Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through Hill numbers. *Annual Review of Ecology, Evolution, and Systematics, 45*, 297–324. https://doi.org/10.1146/annurev-ecolsys-120213-091540

Chen, L., Zhao, N., Cao, J., Liu, X., Xu, J., Ma, Y., Yu, Y., Zhang, X., Zhang, W., Guan, X., Yu, X., Liu, Z., Fan, Y., Wang, Y., Liang, F., Wang, D., Zhao, L., Song, M., & Wang, J. (2022). Short- and long-read metagenomics expand individualized

structural variations in gut microbiomes. *Nature Communications, 13*, Article 34149. https://doi.org/10.1038/s41467-022-30857-9

Davis, B. C., Brown, C., Gupta, S., Calarco, J., Liguori, K., Milligan, E., Harwood, V. J., Pruden, A., & Keenum, I. (2023). Recommendations for the use of metagenomics for routine monitoring of antibiotic resistance in wastewater and impacted aquatic environments. *Critical Reviews in Environmental Science and Technology, 53*(19), 1731–1756. https://doi.org/10.1080/10643389.2023.2181620

Fernández-Pato, A., Sinha, T., Gacesa, R., Andreu-Sánchez, S., Gois, M. F. B., Gelderloos-Arends, J., Jansen, D. B. H., Kruk, M., Jaeger, M., Joosten, L. A. B., Netea, M. G., Weersma, R. K., Wijmenga, C., Harmsen, H. J. M., Fu, J., Zhernakova, A., & Kurilshikov, A. (2024). Choice of DNA extraction method affects stool microbiome recovery and subsequent phenotypic association analyses. *Scientific Reports, 14*, Article 54353. https://doi.org/10.1038/s41598-024-54353-w

Gemmell, M. R., Jayawardana, T., Koentgen, S., Brooks, E., Kennedy, N., Berry, S., Lees, C., & Hold, G. L. (2024). Optimised human stool sample collection for multi-omic microbiota analysis. *Scientific Reports, 14*, Article 67499. https://doi.org/10.1038/s41598-024-67499-4

Hou, K., Wu, Z. X., Chen, X. Y., Wang, J. Q., Zhang, D., Xiao, C., Zhu, D., Koya, J. B., Wei, L., Li, J., & Chen, Z. S. (2022). Microbiota in health and diseases. *Signal Transduction and Targeted Therapy, 7*, Article 135. https://doi.org/10.1038/s41392-022-00974-4

Kers, J. G., & Saccenti, E. (2022). The power of microbiome studies: Some considerations on which alpha and beta metrics to use and how to report results. *Frontiers in Microbiology, 12*, Article 796025. https://doi.org/10.3389/fmicb.2021.796025

Kharnaior, P., & Tamang, J. P. (2022). Metagenomic-metabolomic mining of kinema, a naturally fermented soybean food of the eastern Himalayas. *Frontiers in Microbiology, 13*, Article 868383. https://doi.org/10.3389/fmicb.2022.868383

Kim, C., Pongpanich, M., & Porntaveetus, T. (2024). Unraveling metagenomics through long-read sequencing: A comprehensive review. *Journal of Translational Medicine, 22*(1). https://doi.org/10.1186/s12967-024-04917-1

Kim, C. Y., Ma, J., & Lee, I. (2022). HiFi metagenomic sequencing enables assembly of accurate and complete genomes from human gut microbiota. *Nature Communications, 13*, Article 34149. https://doi.org/10.1038/s41467-022-34149-0

Kim, N., Ma, J., Kim, W., Kim, J., Belenky, P., & Lee, I. (2024). Genome-resolved metagenomics: A game changer for microbiome medicine. *Experimental and Molecular Medicine, 56*(7), 1501–1512. https://doi.org/10.1038/s12276-024-01262-7

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L. I., McDonald, D., Melnik, A. V., Morton, J. T., Navas, J., Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R., Tripathi, A., Xu, Z. Z., … Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology, 16*(7), 410–422. https://doi.org/10.1038/s41579-018-0029-9

Lozupone, C., & Knight, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology, 71*(12), 8228–8235. https://doi.org/10.1128/AEM.71.12.8228-8235.2005

Mak, L., Tierney, B., Wei, W., Ronkowski, C., Toscan, R. B., Turhan, B., Toomey, M., Martinez, J. S. A., Fu, C., Lucaci, A. G., Solano, A. H. B., Setubal, J. C., Henriksen, J. R., Zimmerman, S., Kopbayeva, M., Noyvert, A., Iwan, Z., Kar, S., Nakazawa, N., … Hajirasouliha, I. (2025). CAMP: A modular metagenomics analysis system for integrated multi-step data exploration. *bioRxiv*. https://doi.org/10.1101/2023.04.09.536171

BAB 12
Proteomik & Metabolomik Komputasional

Rizki Rahmadi Pratama

Perkembangan cepat teknologi "*omics*" telah merevolusi kapasitas kita dalam mengkarakterisasi sistem biologis secara menyeluruh pada skala molekuler, terutama lewat genomik, proteomik, dan metabolomik (Nurhayati & Lawanda, 2023). Pendekatan multi-omik, yang mengintegrasikan data dari beragam platform, kini esensial untuk mencapai pemahaman holistik mengenai interaksi rumit serta mekanisme regulasi di dalam sistem biologis (Luo *et al.*, 2024). Proteomik dan metabolomik secara khusus menjadi dua fondasi utama studi "*omics*", menyediakan perspektif unik terhadap ekspresi protein, modifikasi pasca-translasi, interaksi protein-protein, profil metabolit, serta jalur biokimia yang mencerminkan kondisi fungsional sel dan organisme secara keseluruhan.

Penggabungan analisis komputasional data proteomik dan metabolomik memfasilitasi identifikasi *biomarker*, penemuan target obat, serta pencerahan jalur penyakit dengan tingkat ketepatan yang tak tertandingi (Ali & Mohammed, 2023). Bagaimanapun, integrasi data masif dari berbagai platform omik masih menghadapi kendala substansial, khususnya terkait standardisasi data, variasi format, dan kerumitan interpretasi

(Wanichthanarak *et al.*, 2015). Oleh karena itu, bab ini akan mengkaji secara mendalam prinsip, metode, serta aplikasi proteomik dan metabolomik komputasional, dengan penekanan pada strategi integrasi multi-omik inovatif, guna mengungkap pola serta interaksi kompleks yang berpotensi terabaikan dalam analisis omik individual (Blum *et al.*, 2018).

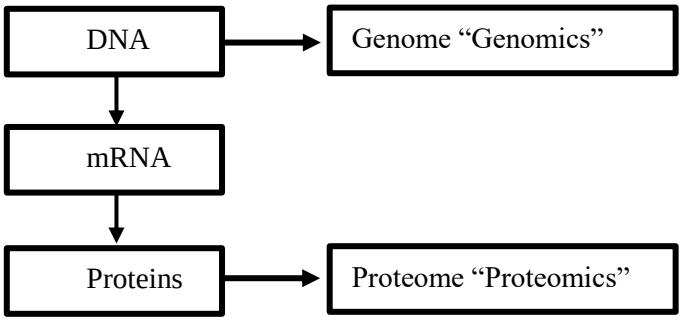
Proteomik Komputasional

Pengantar Proteomik

Proteomik merupakan bidang studi yang mengkaji proteom, yaitu keseluruhan komplemen protein dari genom suatu organisme. Istilah "proteomik" dan "proteom" diperkenalkan oleh Marc Wilkins dan rekan-rekannya pada awal tahun 1990-an, yang mencerminkan konsep "genomik" dan "genom" yang merujuk pada seluruh koleksi gen (Gambar 12.1). Konsep "-omik" ini melambangkan definisi ulang mengenai cara pandang terhadap biologi dan fungsi sistem kehidupan. Sebelum pertengahan 1990-an, para ilmuwan biokimia, biologi molekuler, dan biologi sel umumnya mempelajari gen dan protein individual atau kelompok kecil komponen terkait dalam jalur biokimia spesifik. Teknik yang tersedia saat itu, seperti *Northern blot* untuk ekspresi gen dan *Western blot* untuk tingkat protein, menjadikan pemetaan status lebih dari segelintir gen atau protein sebagai tugas analitis yang sangat menantang.

Dalam konteks ini, proteomik menjadi krusial, didukung oleh kemajuan sekuensing genom dan spektrometri massa, serta *tools* analitis dan perangkat lunak untuk identifikasi protein. Proteomik juga memungkinkan pemetaan modifikasi protein yang vital untuk fungsi seluler, menggunakan teknik seperti 2D-

SDS-PAGE, LC-MS/MS, dan penandaan isotop untuk analisis protein dan profil ekspresi.



Gambar 12.1. Konteks dari genomik dan proteomik (Liebler, 2002)

Definisi dan Ruang Lingkup Proteomik

Proteomik didefinisikan sebagai penelitian sistematis terhadap proteome, yakni seluruh kumpulan protein yang diekspresikan oleh suatu organisme, sistem, atau lingkungan biologis pada waktu spesifik, yang bervariasi antar sel dan waktu akibat regulasi pasca-transkripsi. Cakupan ruang lingkupnya meliputi identifikasi, kuantifikasi, serta karakterisasi modifikasi pasca-translasi protein (seperti fosforilasi dan ubiquitinasi) yang keseluruhannya merupakan penting untuk memahami fungsi seluler, adaptasi terhadap variasi lingkungan, serta jaringan interaksi protein-protein (Murmu *et al.*, 2024; Reis-de-Oliveira *et al.*, 2024). Oleh karena itu, proteomik memberikan representasi fungsional yang lebih presisi ketimbang genomik atau transkriptomik semata. Hal ini dikarenakan protein merupakan pelaku utama dalam hampir seluruh proses biologis, termasuk jalur pensinyalan dan kompleks protein.

Pendekatan tersebut amat vital karena menyajikan pemahaman mendalam mengenai mekanisme patogenesis,

identifikasi *biomarker*, dan inovasi terapi, mengingat kerumitan serta dinamika peran protein dalam sistem hidup, sebagaimana didukung integrasi dengan omik lain seperti metabolomik (Nurhayati & Lawanda, 2023). Penerapan kecerdasan buatan telah merevolusikan proteomik melalui penanganan keterbatasan metode eksperimental konvensional dalam mengolah data biologis masif dari spektrometri massa serta mengungkap pola kompleks yang terselubung, seperti *biomarker* kanker (Tan *et al.*, 2025). Integrasi AI memungkinkan deteksi protein dan peptida dengan akurasi lebih tinggi, kuantifikasi yang tepat, serta interpretasi data proteomik berskala besar secara optimal, termasuk prediksi retensi kromatografi dan fragmentasi peptida (Sun *et al.*, 2025).

Pentingnya Studi Proteomik dalam Biologi Sistem dan Biomedis

Studi proteomik memainkan peran sentral dalam biologi sistem dengan memungkinkan para peneliti untuk menganalisis seluruh protein beserta interaksinya di dalam sel atau organisme, yang menjadi dasar fungsional utama bagi berbagai proses biologis (Rinner, 2016). Aspek ini penting untuk memahami bagaimana protein secara kolektif mengendalikan fungsi seluler, menanggapi perubahan lingkungan, serta berkontribusi terhadap fenotipe kompleks (Ali & Mohammed, 2023). Dalam ranah biomedis, proteomik memberikan pemahaman mendalam tentang mekanisme molekuler penyakit, termasuk pengenalan protein yang berperan dalam patogenesis serta resistensi obat, sehingga mendukung pengembangan strategi diagnostik dan terapeutik yang lebih optimal (Zhou *et al.*, 2024). Perkembangan spektrometri massa dan teknik bioinformatika telah membuka

peluang analisis proteomik berskala besar, sehingga memungkinkan identifikasi ribuan protein dan modifikasinya secara serempak (Sun *et al.*, 2025).

Metode dan Teknik Proteomik Komputasional

Proteomik komputasional memanfaatkan algoritme dan perangkat lunak statistik untuk menganalisis data proteomik berskala besar dari spektrometri massa guna mengidentifikasi protein, mengkuantifikasi ekspresinya, serta mengkarakterisasi modifikasi pasca-translasi. Salah satu fokus utama dalam proteomik komputasional adalah identifikasi protein melalui perbandingan spektrum massa fragmentasi eksperimental dengan basis data sekuens protein teoritis. Tantangan utama muncul dari besarnya set data proteomik modern, variabilitas eksperimental, serta keberadaan modifikasi pasca-translasi, sehingga memerlukan pendekatan bioinformatika dan algoritme pembelajaran mesin untuk memungkinkan analisis data yang lebih efisien dan akurat (Lam *et al.*, 2016; Schmidt *et al.*, 2014; Tan *et al.*, 2025).

Identifikasi Protein

Identifikasi protein merupakan langkah fundamental dalam proteomik komputasional, dengan spektrum massa dari sampel biologis diinterpretasikan untuk menentukan identitas protein penyusunnya (Mohanasundaram *et al.*, 2023). Proses ini melibatkan pencocokan spektrum massa fragmentasi eksperimental dengan spektrum teoretis dari basis data sekuens protein menggunakan perangkat lunak pencarian. Selain metode berbasis basis data, pendekatan *de novo sequencing* digunakan untuk merekonstruksi sekuens peptida langsung dari spektrum

tanpa referensi basis data. Tantangan utama dalam identifikasi protein meliputi protein dengan kelimpahan rendah, modifikasi pasca-translasi yang kompleks, serta sekuens peptida novel yang tidak terdapat dalam basis data. Oleh karena itu, diperlukan pengembangan algoritme khusus dan integrasi dengan data genomik untuk meningkatkan komprehensivitas analisis.

Spektrometri Massa

Spektrometri massa merupakan fondasi utama dalam identifikasi protein, menyediakan informasi krusial mengenai rasio massa-ke-muatan (m/zm/z m/z) peptida dan fragmennya yang diinterpretasikan secara komputasional untuk menentukan urutan asam amino (Mohanasundaram *et al.*, 2023). Metode identifikasi protein berbasis spektrometri massa terbagi menjadi dua strategi utama: (1) berbasis basis data yang melibatkan pencocokan spektrum MS/MS eksperimental dengan spektrum teoretis menggunakan perangkat lunak seperti SEQUEST, Mascot, X!Tandem, dan OMSSA; serta (2) *de novo sequencing* yang menyimpulkan sekuens peptida langsung tanpa referensi basis data, berguna untuk metaproteomik meskipun akurasinya masih rendah (Rathod *et al.*, 2024).

Pengembangan perangkat lunak berbasis pembelajaran mesin dengan metrik penilaian dan model fragmentasi lebih akurat telah meningkatkan sistem penilaian dan mengurangi *false positives* secara signifikan. Kemajuan teknologi spektrometri massa dengan resolusi tinggi memungkinkan analisis ribuan peptida secara simultan serta analisis kuantitatif untuk perbandingan kelimpahan protein antar sampel (Chen *et al.*, 2020; Sun *et al.*, 2025).

Metabolomik Komputasional

Pengantar Metabolomik

Metabolomik didefinisikan sebagai penelitian sistematis terhadap profil metabolit berukuran molekul kecil dalam sistem biologi. Metabolomik menyajikan representasi fungsional langsung atas status fisiologis atau patologis suatu organisme (Kaddurah-Daouk & Weinshilboum, 2015). Berbeda dari genomik, transkriptomik, maupun proteomik yang menekankan potensi genetik atau dinamika ekspresi molekuler, metabolomik secara eksplisit mencerminkan respons biologis yang dinamis akibat variasi lingkungan internal maupun eksternal, sehingga menjadikannya disiplin penting dalam pemahaman patogenesis penyakit dan identifikasi *biomarker* inovatif (Kaddurah-Daouk & Weinshilboum, 2015; Nurhayati & Lawanda, 2023). Pendekatan tersebut memfasilitasi deteksi *biomarker* yang lebih presisi, serta pemetaan jalur metabolik yang terlibat dalam mekanisme penyakit (Kaddurah-Daouk & Weinshilboum, 2015). Perkembangan metode pembelajaran mesin dan kecerdasan buatan telah meningkatkan kapabilitas analisis data multi-omik secara substansial, sehingga mendukung identifikasi target terapeutik yang lebih tepat serta pendekatan pengobatan yang dipersonalisasi (Garg *et al.*, 2024; Mukherjee *et al.*, 2024; Tan *et al.*, 2025). Dengan demikian, metabolomik berfungsi sebagai penghubung vital dalam menjelaskan mengenai fenotipe rumit yang timbul dari interaksi gen, protein, dan faktor lingkungan (Mukherjee *et al.*, 2024). Integrasi metabolomik dengan cabang omik lainnya, seperti genomik dan proteomik, menghasilkan pemahaman holistik mengenai sistem biologis, serta berkontribusi secara integral dalam paradigma kedokteran

presisi (Ivanišević & Sewduth, 2023; Kaddurah-Daouk & Weinshilboum, 2015).

Definisi dan Ruang Lingkup Metabolomik

Metabolomik adalah studi sistematis terhadap metabolit (molekul kecil (<1500 Dalton) yang ditemukan di dalam sel, cairan biologis, jaringan, atau organisme), yang mencerminkan respons fisiologis *real-time* terhadap pengaruh genetik dan lingkungan (Du *et al.*, 2022; Doğan, 2024). Berbeda dengan genomik atau proteomik yang memperkirakan potensi predisposisi, metabolomik berfokus pada analisis kuantitatif dan kualitatif dari metabolom (keseluruhan koleksi metabolit) yang memberikan gambaran fungsional langsung tentang status biologis. Cakupannya terdiri dari berbagai kelas senyawa seperti asam amino, karbohidrat, lipid, nukleotida, dan kofaktor sebagai produk akhir ekspresi gen dan aktivitas protein (Mukherjee *et al.*, 2024; Wörheide *et al.*, 2020). Cakupan metabolomik melampaui identifikasi statis dengan melibatkan pemantauan dinamis perubahan konsentrasi metabolit seiring waktu untuk melacak respons adaptif atau patologis, termasuk analisis metabolit eksogen seperti xenobiotik, diet, dan metabolit mikroba yang membentuk profil metabolik inang (Kosmides *et al.*, 2013). Metabolomik menjadi jembatan vital antara genotipe dan fenotipe, memberikan wawasan mengenai fungsi sel, perubahan metabolisme, dan respons terhadap stres atau penyakit (Nalbantoğlu, 2019).

Keterkaitan Metabolomik dengan "*Omics*" lainnya

Integrasi metabolomik dengan genomik, transkriptomik, dan proteomik sangat krusial untuk membangun pemahaman

holistik tentang sistem biologis yang kompleks (Jendoubi, 2021; Wörheide *et al.*, 2020). Pendekatan multi-omik ini memungkinkan identifikasi *biomarker* penyakit baru dengan akurasi yang lebih tinggi, serta pemetaan jalur penyakit secara komprehensif melalui pengungkapan interaksi lintas lapisan molekuler (Ivanišević & Sewduth, 2023).

Data metabolomik dapat mengungkapkan perubahan metabolik yang tidak terdeteksi oleh analisis genomik atau proteomik saja—seperti fluktuasi konsentrasi metabolit akibat regulasi enzimatik pasca-translasi—sementara data genomik menjelaskan dasar genetik dari variasi metabolik yang diamati. Secara bersamaan, data metabolomik dan data genomic menciptakan sinergi yang saling melengkapi (Mukherjee *et al.*, 2024). Sinergi ini memungkinkan peneliti menguraikan jaringan regulasi yang mengikat genotipe dan fenotipe, memberikan wawasan mendalam mengenai patogenesis penyakit, serta pengembangan strategi terapeutik yang lebih efektif dan personalisasi pengobatan berbasis *biomarker* (Kaddurah-Daouk & Weinshilboum, 2015; Karakitsou *et al.*, 2019).

Metode dan Teknik Metabolomik Komputasional

Tantangan utama integrasi data multi-omika terletak pada dimensi data yang sangat tinggi, volume besar, serta heterogenitasnya, yang memerlukan metode komputasi lanjutan (Ivanišević & Sewduth, 2023; Wörheide *et al.*, 2020). Metode analitik utama untuk akuisisi data metabolomik mencakup spektrometri massa yang dikombinasi dengan kromatografi (LC-MS/MS untuk metabolit polar, GC-MS untuk volatil), resonansi magnetik nuklir (NMR) untuk identifikasi struktural non-destruktif, serta teknik lainnya seperti CE-MS, FT-ICR-MS, dan

ICP-MS. Masing-masing metode memiliki keunggulan dalam identifikasi, kuantifikasi, dan pemantauan dinamis metabolit berukuran kecil (<1500 Da) (Doğan, 2024; Fitriana *et al.*, 2017).

Meskipun rentan supresi ion, LC-MS/MS tetap merupakan pilihan utama untuk metabolit polar dengan sensitivitas tinggi dan *throughput elevated*. GC-MS unggul untuk metabolit volatil dengan resolusi pemisahan tinggi dan reproduktibilitas superior. NMR menawarkan kuantifikasi langsung tanpa destruksi meskipun sensitivitasnya rendah, sementara CE-MS efektif untuk metabolit bermuatan tinggi. FT-ICR-MS memberikan resolusi massa ultra-tinggi, dan ICP-MS difokuskan pada deteksi elemen dan mineral (Euceda *et al.*, 2015; Ivanišević & Sewduth, 2023). Kombinasi teknik ortogonal ini secara sinergis meningkatkan *coverage* metabolomik hingga >50% metabolom, serta mengurangi bias platform tunggal. Strategi integrasi data transkriptomik, proteomik, dan metabolomik mencakup analisis korelasi, model berbasis kendala, analisis pengayaan jalur, dan pembelajaran mesin. Hal ini memungkinkan penelusuran balik dari fenotipe metabolik ke akar genetik dan regulasi protein yang mendasarinya (Karakitsou *et al.*, 2019).

Basis Data Metabolit

Basis data metabolit berfungsi sebagai repositori sentral yang menyimpan informasi terstruktur tentang metabolit, termasuk sifat fisikokimia, jalur biokimia, dan asosiasi penyakit, yang esensial untuk anotasi dan interpretasi data metabolomik skala besar (Nassar *et al.*, 2016). Beberapa basis data metabolit terkemuka mencakup HMDB (*Human Metabolome Database*), yang menyediakan data komprehensif metabolit endogen manusia termasuk data spektral dan konsentrasi, KEGG yang

berfokus pada peta jalur biokimia dan hubungan antarmetabolit, serta METLIN yang unggul dalam spektrum massa dan informasi fragmentasi untuk identifikasi metabolit dari data eksperimen. Kombinasi penggunaan basis data ini krusial untuk validasi silang dan interpretasi data metabolomik yang akurat, termasuk integrasi dengan basis data obat seperti DrugBank untuk menghubungkan metabolomik dengan informasi farmakologi (Liu & Locasale, 2017).

DAFTAR PUSTAKA

Ali, A. M., & Mohammed, M. A. (2023). A comprehensive review of artificial intelligence approaches in omics data processing: Evaluating progress and challenges. *International Journal of Mathematics, Statistics and Computer Science, 2*, Article 114. https://doi.org/10.59543/ijmscs.v2i.8703

Blum, B. C., Mousavi, F., & Emili, A. (2018). Single-platform 'multi-omic' profiling: Unified mass spectrometry and computational workflows for integrative proteomics–metabolomics analysis. *Molecular Omics, 14*(5), 307–319. https://doi.org/10.1039/c8mo00136g

Chen, C., Hou, J., Tanner, J. J., & Cheng, J. (2020). Bioinformatics methods for mass spectrometry-based proteomics data analysis. *International Journal of Molecular Sciences, 21*(8), Article 2873. https://doi.org/10.3390/ijms21082873

Doğan, H. O. (2024). Metabolomics: A review of liquid chromatography mass spectrometry-based methods and clinical applications. *Turkish Journal of Biochemistry, 49*(1), 1–15. https://doi.org/10.1515/tjb-2023-0095

Du, X., Aristizabal-Henao, J. J., Garrett, T. J., Brochhausen, M., Hogan, W. R., & Lemas, D. J. (2022). A checklist for reproducible computational analysis in clinical metabolomics research. *Metabolites, 12*(1), Article 87. https://doi.org/10.3390/metabo12010087

Euceda, L. R., Giskeødegård, G. F., & Bathen, T. F. (2015). Preprocessing of NMR metabolomics data. *Scandinavian Journal of Clinical and Laboratory Investigation, 75*(3), 193–203. https://doi.org/10.3109/00365513.2014.1003593

Fitriana, H. N., Purwadi, R., & Kresnowati, M. T. A. P. (2017). Pemetaan pengaruh proses pengolahan pada kualitas biji kakao menggunakan metode metabolik profiling dengan GC/MS. *REAKTOR, 17*(3), 132–139. https://doi.org/10.14710/reaktor.17.3.132-139

Garg, P., Singhal, G. D., Kulkarni, P., Horne, D., Salgia, R., & Singhal, S. S. (2024). Artificial intelligence–driven computational approaches in the development of anticancer drugs. *Cancers, 16*(22), Article 3884. https://doi.org/10.3390/cancers16223884

Ivanišević, T., & Sewduth, R. N. (2023). Multi-omics integration for the design of novel therapies and the identification of novel biomarkers. *Proteomes, 11*(4), Article 34. https://doi.org/10.3390/proteomes11040034

Jendoubi, T. (2021). Approaches to integrating metabolomics and multi-omics data: A primer. *Metabolites, 11*(3), Article 184. https://doi.org/10.3390/metabo11030184

Kaddurah-Daouk, R., & Weinshilboum, R. M. (2015). Metabolomic signatures for drug response phenotypes: Pharmacometabolomics enables precision medicine. *Clinical Pharmacology & Therapeutics, 98*(1), 71–75. https://doi.org/10.1002/cpt.134

Karakitsou, E., Foguet, C., Atauri, P. de, Kultima, K., Khoonsari, P. E., Santos, V. A. P. M. dos, Saccenti, E., Rosato, A., & Cascante, M. (2019). Metabolomics in systems medicine: An overview of methods and applications. *Current Opinion in Systems Biology, 15*, 91–98. https://doi.org/10.1016/j.coisb.2019.03.009

Kosmides, A. K., Kamisoglu, K., Calvano, S. E., Corbett, S., & Androulakis, I. P. (2013). Metabolomic fingerprinting:

Challenges and opportunities. *Critical Reviews in Biomedical Engineering, 41*(3), 205–221. https://doi.org/10.1615/critrevbiomedeng.2013007736

Lam, M. P. Y., Lau, E., Ng, D. C. M., Wang, D., & Ping, P. (2016). Cardiovascular proteomics in the era of big data: Experimental and computational advances. *Clinical Proteomics, 13*, Article 18. https://doi.org/10.1186/s12014-016-9124-y

Liu, X., & Locasale, J. W. (2017). Metabolomics: A primer. *Trends in Biochemical Sciences, 42*(4), 274–284. https://doi.org/10.1016/j.tibs.2017.01.004

Luo, Y., Zhao, C., & Chen, F. (2024). Multiomics research: Principles and challenges in integrated analysis. *BioDesign Research, 6*, Article 59. https://doi.org/10.34133/bdr.0059

Mohanasundaram, S., Dhatwalia, D., Vijayaraghavan, P., Alzubaidi, L. H., & Makhzuna, K. (2023). Bioinformatics: Computational approaches for genomics and proteomics. *E3S Web of Conferences, 399*, Article 04042. https://doi.org/10.1051/e3sconf/202339904042

Mukherjee, A., Abraham, S., Singh, A., Balaji, S., & Mukunthan, K. S. (2024). From data to cure: A comprehensive exploration of multi-omics data analysis for targeted therapies. *Molecular Biotechnology*. Advance online publication. https://doi.org/10.1007/s12033-024-01133-6

Murmu, S., Sinha, D., Chaurasia, H., Sharma, S., Das, R., Jha, G. K., & Archak, S. (2024). A review of artificial intelligence-assisted omics techniques in plant defense: Current trends and future directions. *Frontiers in Plant Science, 15*, Article 1292054. https://doi.org/10.3389/fpls.2024.1292054

Nalbantoğlu, S. (2019). Metabolomics: Basic principles and

strategies. In *Molecular medicine*. IntechOpen. https://doi.org/10.5772/intechopen.88563

Nassar, A. F., Wu, T., Nassar, S. F., & Wisnewski, A. V. (2017). UPLC–MS for metabolomics: A giant step forward in support of pharmaceutical research. *Drug Discovery Today, 22*(2), 463–470. https://doi.org/10.1016/j.drudis.2016.11.020

Nurhayati, E. S., & Lawanda, I. I. (2023). Perkembangan dan tren penelitian global tentang research data management. *Lentera Pustaka: Jurnal Kajian Ilmu Perpustakaan, Informasi dan Kearsipan, 9*(2), 201–214. https://doi.org/10.14710/lenpust.v9i2.55264

Rathod, S., Patel, N., & Prajapati, B. G. (2024). Bioinformatics in green and sustainable technologies. In *IntechOpen eBooks*. IntechOpen. https://doi.org/10.5772/intechopen.112108

Reis-de-Oliveira, G., Carregari, V. C., Sousa, G. R. dos R. de, & Martins-de-Souza, D. (2024). OmicScope unravels systems-level insights from quantitative proteomics data. *Nature Communications, 15*, Article 50875. https://doi.org/10.1038/s41467-024-50875-z

Rinner, O. (2016). Next-generation proteomics from an industrial perspective. *CHIMIA International Journal for Chemistry, 70*(12), 860–865. https://doi.org/10.2533/chimia.2016.860

Schmidt, A., Forné, I., & Imhof, A. (2014). Bioinformatic analysis of proteomics data. *BMC Systems Biology, 8*(Suppl 2), Article S3. https://doi.org/10.1186/1752-0509-8-s2-s3

Sun, Y., Ai, J., Liu, Z., Sun, R., Qian, L., Payne, S., Bittremieux, W., Ralser, M., Li, C., Chen, Y., Dong, Z., Pérez-Riverol, Y., Khan, A. A., Sander, C., Aebersold, R., Vizcaíno, J. A., Krieger, J. R., Yao, J., Wen, H., & Guo, T. (2025). Strategic priorities for transformative progress in advancing biology with

proteomics and artificial intelligence. *arXiv*. https://doi.org/10.48550/arxiv.2502.15867

Tan, C., Liu, H., Zhang, Z., Liu, X., Ai, Y., Wu, X., Jian, E., Song, Y., & Yang, J. (2025). Progress and trends on machine learning in proteomics during 1997–2024: A bibliometric analysis. *Frontiers in Medicine, 12*, Article 1594442. https://doi.org/10.3389/fmed.2025.1594442

Wanichthanarak, K., Fahrmann, J. F., & Grapov, D. (2015). Genomic, proteomic, and metabolomic data integration strategies. *Biomarker Insights, 10*, 1–6. https://doi.org/10.4137/bmi.s29511

Zhou, Y., Chen, L., & Zhang, W. (2024). Application of proteomics in understanding disease pathogenesis and drug resistance: A systematic review. *International Journal of Molecular Sciences, 25*(3), Article 1450. https://doi.org/10.3390/ijms25031450

BAB 13
Visualisasi Data Biologi

Muhamad Firmanda

Visualisasi data merupakan salah satu bagian penting dalam penyampaian informasi kepada audiens. Informasi dapat disampaikan dalam bentuk tulisan. Penyampaian informasi dalam bentuk tulisan yang panjang dapat disederhanakan dalam bentuk visual. Bentuk visual dapat memperkuat pemahaman audiens terhadap informasi yang ingin disampaikan, sehingga penyampaian informasi dapat lebih efisien dan efektif. Manfaat dari visualisasi data dapat diperoleh apabila bentuk visual merepresentasikan informasi yang ingin disampaikan. Bab ini membahas prinsip visualisasi data ilmiah meliputi tujuan, pemilihan warna, bentuk visualiasi, alat yang dapat digunakan untuk visualisasi, contoh visualsiais dan interpretasinya.

Prinsip Visualisasi Data Ilmiah

Prisip dasar visualisasi data ilmiah antara lain:

1. Informasi dan pesan menjadi prioritas utama.
   Informasi dan pesan yang ingin disampaikan adalah prioritas utama dari visualisasi data. Tujuan visualisasi adalah menyampaikan informasi dan pesan sejelas mungkin, sehingga informasi atau pesan yang ingin disampaikan dapat diterima dengan baik dan tepat. Sebelum menggunakan

perangkat lunak, visualisasi dapat dilakukan di atas kertas terlebih dahulu untuk memperoleh kerangka visual yang paling sesuai dengan data dan tujuan visualisasi. Setelah yakin dengan kerangka visual yang telah dibuat di atas kertas, kemudian dapat memilih perangkat lunak yang paling sesuai untuk membuat visualisasi. Ketika kerangka visualisasi sudah jelas maka proses pembuatan visualisasi pada perangkat lunak dapat langsung ke bentuk yang sudah dirancang sebelumnya, sehingga menghemat waktu untuk "*trial and error*" menentukan bentuk dan revisi bentuk visual.

2. Warna memiliki makna.

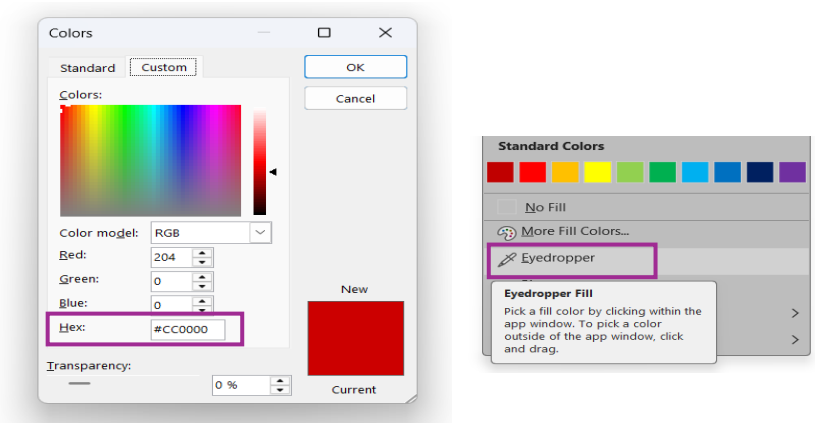
   Warna berfungsi untuk memperkuat pesan yang ingin disampaikan, sehingga pemilihan warna dilakukan dengan alasan yang jelas. Misalnya, warna merah digunakan untuk peningkatan ekspresi gen, warna putih tidak ada perubahan ekspresi, dan warna biru untuk penurunan ekspresi gen. Penggunaan warna umumnya menggunakan tiga skema warna yaitu:

- *Diverging*: dua urutan warna dengan warna netral di tengah untuk melihat kecondongan ke salah satu dari dua ekstrem.

- *Sequential*: urutan warna dari terang ke gelap untuk menyampaikan nilai yang semakin meningkat seiring dengan warna yang semakin gelap.

- *Qualitative*: warna yang kontras satu sama lain untuk menyampaikan perbedaan.

Setiap satu warna pada visualisasi digital memiliki kode *hex* yang tersusun dari enam karakter (Gambar 13.1). Misalnya, 00000 untuk warna hitam, ffffff untuk warna putih, cc000 untuk warna merah, dan sebagainya. *Website ColorBrewer* menyediakan palet warna berdasarkan tiga skema warna *diverging, sequential, dan qualitative*. Selain itu, palet warna dari *website* tersebut disusun supaya *color-blind friendly* dan supaya tetap bisa terbaca dalam cetakan hitam putih. Perangkat lunak pembuat desain umumnya menyediakan kolom untuk memasukkan kode *hex,* sehingga dapat dapat meminimalkan ketidakseragaman warna (warna terlalu muda atau tua). Saat menemukan kombinasi warna yang menarik, kode *hex* bisa digunakan untuk meniru warna tersebut untuk digunakan pada visual yang ingin dibuat. Perangkat lunak pembuat desain umumnya memiliki fitur *color picker* (ikon



Gambar 13.2. Kolom kode hex dan fitur *color picker* pada perangkat lunak PowerPoint

berbentuk pipet) untuk mengambil atau mengetahui kode warna suatu gambar. Selain itu, saat ini tersedia berbagai *website* untuk melihat kode *hex* dari suatu gambar.

Memilih perangkat lunak yang sesuai dengan desain

R dan RStudio

R umum digunakan untuk analisis dan visualisasi data. R menyediakan paket ggplot2 untuk membuat berbagai bentuk visualisasi data. *User interface* paling populer untuk mengoperasikan bahas pemrograman R adalah RStudio. Pengguna perlu menggunakan *command line* untuk melakukan visualisasi data. Pengguna dapat mempelajari visualisasi data pada RStudio melalui tautan berikut: https://psyteachr.github.io/introdataviz/index.html.

SRplot

SRplot merupakan *web server* untuk visualisasi data dengan berbagai modul siap pakai seperti *basic plot, clinical plot,* visualisasi genom, transkriptom, dan epigenom (Tang *et al.,* 2023). *Web server* ini dilengkapi dengan panduan visualisasi data, sehingga pengguna dapat langsung melakukan visualisasi data. Kekurangan *web server* ini adalah pilihan *font* dan elemen visual sangat terbatas. SRplot dapat diakses melalui tautan berikut: https://www.bioinformatics.com.cn/en.

Cytoscape

Cytoscape merupakan perangkat lunak gratis untuk membuat bentuk visualisasi berupa *network,* yang di antaranya adalah miRNA-mRNA, protein-protein, dan gen-protein. Perangkat lunak ini memerlukan *input file excel* berupa *node* dan

edge untuk menyusun *network*. Pengguna dapat mempelajari cara penggunaan Cytoscape melalui tautan berikut: https://manual.cytoscape.org/en/stable/Quick_Tour_of_Cytoscape.html.

Microsoft Excel

Ms. Excel merupakan perangkat lunak pengolahan angka yang dapat digunakan untuk visualisasi data seperti *bar, line, scatter,* dan *boxplot*. Tutorial visualisasi data menggunakan excel banyak dijumpai di berbagai platform salah satunya pada *web* berikut: https://www.excel-easy.com/data-analysis/charts.html

Biovia Discovery Studio dan Molstar

Biovia adalah perangkat lunak untuk membuat visualisasi 2D atau 3D berbagai molekul seperti DNA, RNA, protein, dan senyawa obat. Perangkat lunak ini juga mendukung visualisasi interaksi protein dan ligan hasil *molecular docking*. Di sisi lain, *web server* Molstar berfungsi untuk visualiasi 3D protein dan ligan tanpa perlu *install* perangkat lunak. Molstar dapat diakses pada tautan berikut: https://molstar.org/viewer/.

Menggunakan bentuk visualisasi yang efektif

Data dapat divisualisasi dengan berbagai bentuk seperti histogram, *bar plot*, *heatmap cluster,* dan lain sebagainya. Pemilihan bentuk visualisasi menyesuaikan dengan tujuan visualisasi, contohnya seperti untuk menampikan distribusi, komparasi, hubungan, komposisi, klaster, *network*, dan hierarki. Contoh bentuk visualisasi yang dapat digunakan berdasarkan tujuan visualisasi dapat dilihat pada Tabel 12.1.

Tabel 12.1. Contoh bentuk visualisasi yang dapat digunakan berdasarkan tujuan visualisasi

Tujuan Visualisasi	Bentuk Visualisasi	Contoh Penggunaan
Distribusi	Histogram, *density plot, violin plot*	Melihat distribusi nilai ekspresi gen, read count RNA-seq, atau panjang sekuens DNA/RNA
Komparasi	*Bar plot, dot plot*	Melihat perbandingan ekspresi gen pada dua atau lebih kelompok
Hubungan	*Scatter plot*	Melihat perbedaan ekspresi gen pada kelompok perlakuan dan kontrol
Komposisi	*Stacked bar*	Melihat jumlah pasien pada setiap subtipe kanker payudara stadium I, II, III, dan IV
Klaster	*Heatmap cluster*	Melihat ekspresi miRNA dari beberapa sampel pada kelompok non-kanker dan kanker
Network	*Network graph*	Melihat kumpulan mRNA yang menjadi target miRNA dalam bentuk jaring
Hierarki	Dendogram	Melihat hubungan evolusi antar spesies

Visualisasi data yang efektif dapat dicapai dengan memperhatikan rasio *data-ink*. Prinsip utama rasio *data-ink* adalah menampilkan data dengan meminimalkan elemen visual yang tidak krusial agar audiens bisa fokus sepenuhnya pada data. Dalam sebuah gambar visualisasi data, terdapat tinta yang digunakan untuk menampilkan informasi inti (*data-ink*) dan tinta keseluruhan yang digunakan untuk membuat gambar. Rasio *data-ink* merupakan rasio antara *data-ink* dengan total tinta yang digunakan untuk menyusun gambar (Inbar *et al.,* 2007). Penghitungan rasio *data-ink* dapat menggunakan rumus pada Gambar 13.2. Visualisasi data yang efektif memiliki rasio *data-ink*

yang tinggi (mendekati 1,0). Contoh visualisasi gambar dengan rasio *data-ink* tinggi dan rendah dapat dilihat pada Gambar 13.3.

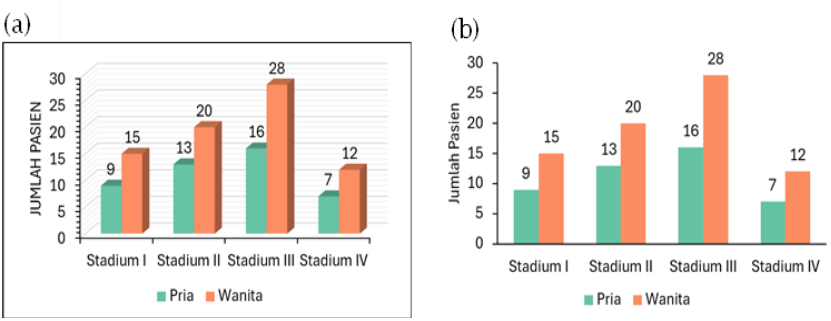
$$\text{Rasio } \textit{data-ink} = \frac{\textit{Data-ink}}{\text{Total } \textit{ink} \text{ yang digunakan untuk mencetak gambar}}$$

Gambar 13.3. Rumus rasio *data-ink*

Keterangan

- *Data-ink*: *Ink* yang langsung menampilkan data. Contoh: batang pada *bar chart*, titik pada *scatter plot*, *error bar*).
- Total *ink*: Seluruh *ink* yang digunakan untuk membuat gambar.
- Bukan *data-ink*: *Ink* dekoratif yang apabila dihapus tidak akan mengubah informasi.

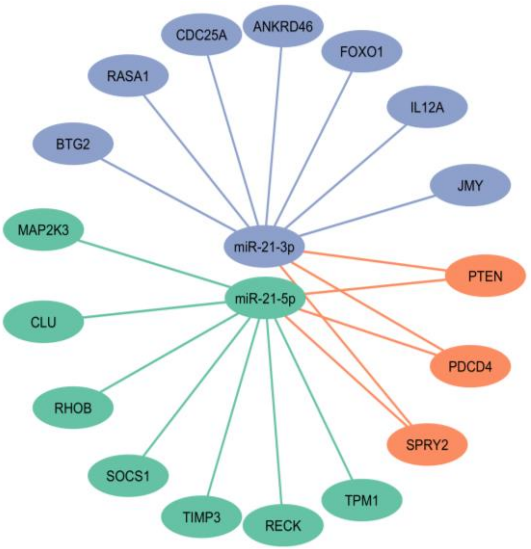
Contoh: *background berwarna*, bingkai, dan *shadow*.



Gambar 13.4. Visual rasio *data-ink* rendah (a) dan rasio *data-ink* tinggi (b)

Contoh visualisasi data dan intepretasinya

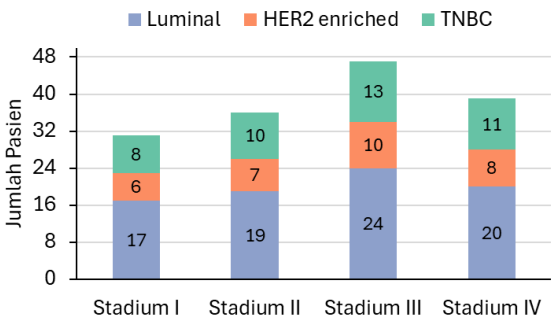
Visualisasi *netowork* miRNA dan gen target menggunakan Cytoscape (Gambar 13.4). Nama miRNA dan gen target merupakan *node,* sedangkan garis yang menghubungkan *node* disebut dengan *edge*. Dari gambar ini, diketahui bahwa miR-21-3p dan miR-21-5p memiliki tiga gen target yang sama, yaitu: PTEN, PDCD4, SPRY2 (warna jingga). Selain itu, terdapat gen yang hanya ditarget oleh miR-21-3p saja (warna ungu) dan oleh miR-21-3p saja (warna hijau).



Gambar 13.5. Gen target miR-21-3p dan miR-21-5p

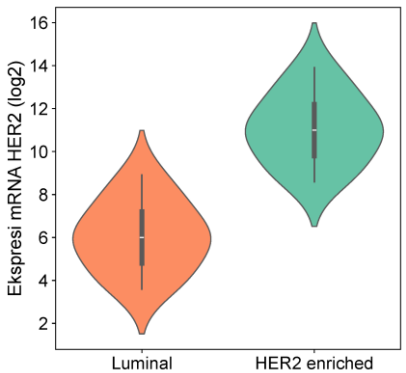
Contoh visualisasi *stacked bar* adalah bisa digunakan untuk menampilkan komposisi pasien kanker payudara menggunakan Ms. Excel. Gambar 13.5 menyajikan informasi jumlah total pasien kanker payudara pada setiap stadium dan jumlah pasien kanker payudara berdasarkan subtipe kanker pada setiap stadium. Setiap warna pada *bar* merepresentasikan subtipe kanker

Luminal (warna ungu), HER2 *enriched* (warna jingga), dan TNBC (warna hijau). Angka pada setiap warna merupakan jumlah pasien tiap subtipe kanker.



Gambar 13.6. Komposisi pasien kanker payudara setiap subtipe dan stadium kanker

Visualisasi *violin plot* ekspresi mRNA HER2 pada sampel kanker payudara subtipe Luminal dan HER2 *enriched* menggunakan *web server* SRplot. Gambar 13.6 menampilkan informasi ekspresi mRNA gen HER2 dalam log2 pada subtipe Luminal (warna jingga) dan HER2 *enriched* (warna hijau). Ekspresi mRNA pada subtipe Luminal berada di kisaran 3-9 dengan median sekitar 6. Sementara itu, eksprsi mRNA pada subtipe

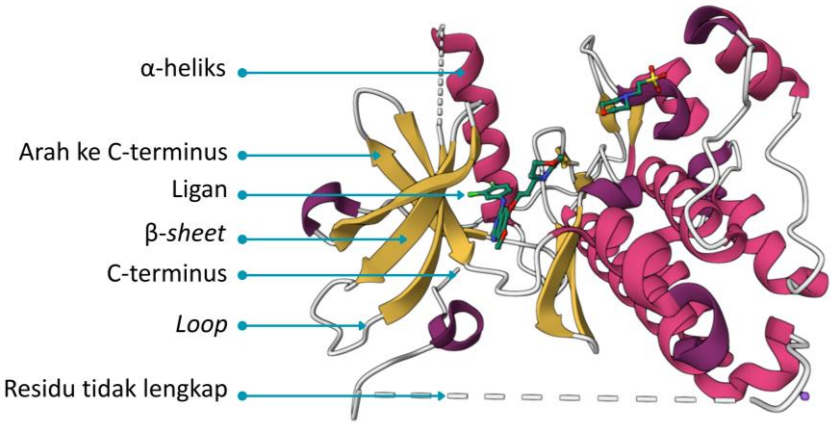


Gambar 13.7. Ekspresi mRNA HER2 pada subtipe Luminal dan HER2 enriched

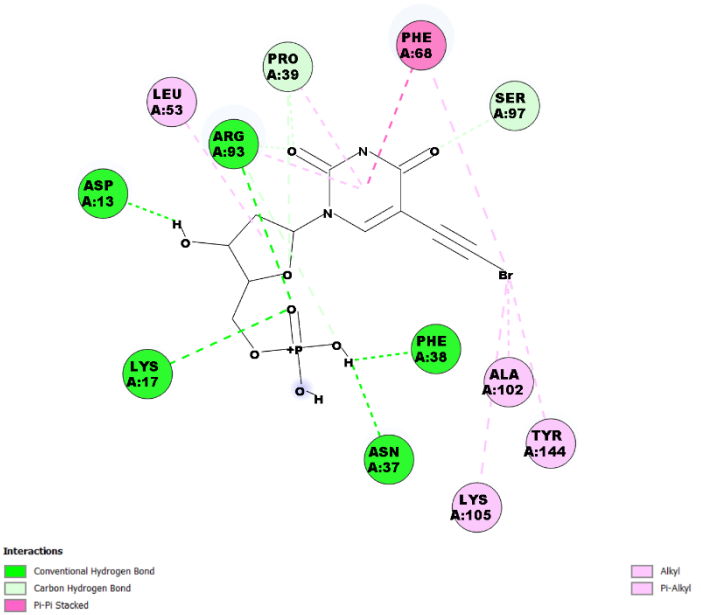
HER2 *enriched* lebih tinggi sekitar 8-14 dengan median sekitar 11. Hal ini menunjukkan bahwa distribusi mRNA HER2 lebih tinggi pada subtipe HER2 *enriched*.

Visualisasi struktur 3D protein dan interaksi protein dengan ligan menggunakan *web server* Molstar dan perangkat lunak Biovia. Gambar 13.7 menampilkan struktur 3D protein EGFR dan ligan gefitinib yang divisualisasikan menggunakan Molstar (PDB ID: 4WKQ). Struktur protein terdiri dari *β-sheet* (warna kuning), *α-helix* (merah muda), dan *loop* (putih) yang menghubungkan keduanya. Sekuens residu asam amino protein memiliki urutan dari N-terminus ke C-terminus sesua arah translasi, sehingga tampilan visual protein yang mirip bentuk anak panah pada *beta sheet* menunjukkan arah dari N-terminus (ujung -NH_2) ke C-terminus (ujung -COOH).

Bagian protein yang terputus-putus mengindikasikan adanya residu asam amino yang tidak lengkap, sehingga tidak tervisualisasikan. Gambar 13.8 menampilkan interaksi ligan *brivudine monophopshate* dengan residu asam amino protein thymidylate kinase (PDB ID: 2V54). Hasil *molecular docking* yang divisualisasikan menggunakan Biovia. Residu asam amino divisualkan dalam bentuk cakram dengan keterangan rantai protein (rantai A) dan nomor urut residu asam amino. Interaksi ligan dengan residu asam amino divisualisasikan dalam bentuk garis putus-putus warna hijau (ikatan hidrogen) dan warna merah muda. Visualisasi 3D berfungsi untuk memperlihatkan struktur protein, lokasi interaksi ligan, serta konformasi dan orientasi ligan. Visualisasi 2D berfungsi untuk menyampaikan informasi residu asam amino yang berperan dalam interaksi dengan ligan dan jenis interaksinya.



Gambar 13.8. Visualisasi protein EGFR dan ligan gefitinib (PDB ID: 4WKQ)



Gambar 13.9. Visualisasi interksi ligan dengan residu asam amino protein thymidylate kinase (PDB ID: 2V54)

DAFTAR PUSTAKA

Inbar, O., Tractinsky, N., & Meyer, J. (2007). Minimalism in information visualization: Attitudes towards maximizing the data-ink ratio. *ACM International Conference Proceeding Series, 250*, 185–188. https://doi.org/10.1145/1362550.1362587

Tang, D., Chen, M., Huang, X., Zhang, G., Zeng, L., Zhang, G., Wu, S., & Wang, Y. (2023). SRplot: A free online platform for data visualization and graphing. *PLoS ONE, 18*(11), e0294236. https://doi.org/10.1371/journal.pone.0294236

BAB 14
Etika, Reproduksibilitas, dan Masa Depan Bioinformatika

Pauline Elisabeth

Etika, reproduksibilitas, dan masa depan bioinformatika merupakan pilar utama dalam pengembangan bioinformatika, karena perannya sebagai penghubung antara data biologis mentah dan aplikasi yang berdampak langsung pada manusia, lingkungan, dan industri. Dari sisi etika, bioinformatika berhadapan dengan data biologis yang sensitif, sehingga praktiknya menuntut perlindungan privasi, persetujuan penggunaan data, keadilan akses teknologi, serta pencegahan penyalahgunaan data dan eksploitasi sumber daya genetik secara tidak adil.

Reproduksibilitas merupakan pilar penting dalam bioinformatika karena hasil analisis komputasi sering menjadi landasan bagi keputusan eksperimental dan pengembangan produk, seperti rekayasa genetik, pemuliaan tanaman berbasis genom, serta penemuan target obat. Analisis yang tidak dapat direproduksi dapat menyebabkan kegagalan eksperimen basah, pemborosan sumber daya, dan kesimpulan biologis yang tidak akurat. Oleh karena itu, reproduksibilitas menuntut dokumentasi *pipeline* yang jelas, penerapan standar data dan metadata, serta keterpaduan antara analisis komputasi dan validasi biologis.

Masa depan bioinformatika bioteknologi akan ditandai oleh integrasi AI, analisis multi-omics, dan kolaborasi lintas disiplin. Bioinformatika berkembang menjadi fondasi strategis bagi bioteknologi berkelanjutan di bidang pangan, kesehatan, dan lingkungan. Tantangan utamanya adalah memastikan kemajuan teknologi tetap selaras dengan prinsip etika dan reproduksibilitas, sehingga inovasi yang dihasilkan tidak hanya unggul secara teknis, tetapi juga tepercaya dan bermanfaat bagi masyarakat.

Etika

Dalam konteks negara Indonesia, penerapan bioteknologi menghadapi tantangan bioetika yang khas karena kuatnya pengaruh nilai agama, sosial, dan budaya, sehingga penerimaan inovasi ilmiah sering bergantung pada otoritas non-sains. Kondisi ini menuntut kerangka regulasi bioetika yang inklusif melalui dialog antara pemerintah, tokoh agama dan adat, maupun ilmuwan. Secara umum, bioteknologi berada di persimpangan antara kemajuan sains dan etika. Meskipun membawa manfaat besar, bioteknologi juga berpotensi menimbulkan dampak negatif, sehingga setiap pengembangannya perlu dikaji secara etis dan hati-hati sebelum diterapkan.

Isu etika representasi data bioinformatika di media digital kian penting ketika informasi ilmiah yang kompleks disederhanakan atau disalahartikan, sehingga menyesatkan persepsi publik. Di Indonesia, rendahnya literasi sains dan maraknya misinformasi di media sosial membuat publikasi hasil genomik atau analisis berbasis AI tanpa konteks yang jelas berisiko memicu *infodemic* dan berdampak negatif pada pemahaman sains dan kesehatan masyarakat.

AI memecahkan masalah dengan cara yang menyerupai manusia melalui proses inferensi, klasifikasi, dan prediksi (Russell & Norvig, 2021). Pada penelitian Greenhill & Edmunds (2020), kemampuan tersebut memungkinkan AI berperan dalam bidang medis, seperti membantu diagnosis, memprediksi prognosis, dan merekomendasikan pengobatan. AI modern umumnya berbasis pembelajaran mesin yang memanfaatkan kumpulan data besar untuk melatih algoritme agar mampu menghasilkan prediksi dari pola data.

AI membawa potensi risiko dan persoalan etis yang perlu dikaji sebelum diterapkan secara luas, termasuk dalam kedokteran genomik pada fase awal kehidupan manusia. Dalam beberapa tahun terakhir, perspektif etika juga menekankan pentingnya *explainable* AI, tata kelola data yang kuat, mitigasi bias algoritmik, serta perlindungan data genomik yang berkelanjutan. Menurut Coghlan *et al.* (2023), pengembangan AI genomik disertai regulasi adaptif, keterlibatan multidisipliner (klinisi, etikus, dan pasien), serta mekanisme pengawasan yang memastikan manfaat klinis tanpa mengorbankan hak dan keselamatan individu.

Perspektif yang disampaikan pada penelitian Shaw *et al.* (2024), dalam bidang medis, kontroversi utama muncul ketika hasil analisis genom atau prediksi AI dipublikasikan ke publik atau media ilmiah seolah-olah bersifat deterministik dan siap klinis, padahal masih bersifat probabilistik dan eksperimental. Praktik ini sering disebut *overclaiming*, terutama dalam studi genomik penyakit kompleks dan AI diagnosis. Selain itu, ada isu serius terkait privasi data genomik, dengan data pasien dapat dire-identifikasi meskipun telah dianonimkan, tetapi tetap dibagikan secara terbuka atau dikomunikasikan secara tidak etis.

Media sering menyederhanakan temuan bioinformatika medis, sehingga memperkuat misinformasi kesehatan (*health misinformation*).

Bioteknologi bidang industri menghadapi beberapa kontroversi utama terkait *sustainability*, "*naturalness*", dan keadilan ekonomi. Menurut Asveld (2019), salah satu isu yang dibahas adalah persepsi publik terhadap teknologi bioteknologi yang dipandang sebagai "tidak alami" atau mengubah tatanan alam secara signifikan, yang dapat memicu resistensi konsumen dan konflik sosial. Selain itu, pertanyaan tentang distribusi manfaat ekonomi muncul ketika perusahaan besar mengeksploitasi teknologi dengan efek yang tidak merata terhadap pekerja, komunitas lokal, atau negara berkembang.

Risiko lainnya mencakup ketergantungan industri pada teknologi tertentu, dampak lingkungan dari proses produksi bioteknologi, dan ketidakpastian jangka panjang terhadap ekosistem. Studi *Societal and Ethical Issues in Industrial Biotechnology* mengidentifikasi lima tema etika utama seperti *sustainability*, *naturalness*, *innovation pathways*, *risk management*, dan *economic justice*, yang sering menjadi pusat perdebatan dalam industri bioteknologi modern.

Dari sisi keunggulan, bioteknologi industri menawarkan potensi besar dalam hal produktivitas, efisiensi sumber daya, dan inovasi ramah lingkungan. Contohnya, penggunaan mikroorganisme dalam produksi bahan bakar *bio-based*. Bioplastik dapat mengurangi ketergantungan pada bahan fosil serta menurunkan jejak karbon industri berat. Pendekatan bioteknologi yang etis dan bertanggung jawab dapat memperkuat nilai-nilai keberlanjutan dan inovasi yang adil, terutama bila disertai praktik *responsible research and innovation*

(RRI) yang melibatkan pemangku kepentingan luas dalam proses pengambilan keputusan. Selain itu, industri bioteknologi mampu mempercepat solusi untuk tantangan global seperti perubahan iklim, limbah industri, dan ketahanan energi. Hal ini mampu terwujud jika diatur dengan kebijakan etika yang tepat dan transparan. Kerangka bioetika modern mendukung pendekatan yang seimbang antara inovasi dan tanggung jawab sosial untuk memastikan bahwa manfaat teknologi dapat dinikmati secara aman dan adil oleh masyarakat.

Dalam bioinformatika bidang pertanian, isu paling kontroversial adalah *biopiracy* dan data *colonialism*, yaitu eksploitasi data genom tanaman lokal atau plasma nutfah dari negara berkembang, yang kemudian dianalisis dan dikomersialisasi oleh institusi atau perusahaan luar tanpa keterlibatan adil masyarakat lokal. Media dan publikasi ilmiah sering mengabaikan konteks sosial dan ekologis dari data yang dianalisis, sehingga bioinformatika tampil netral secara teknis tetapi problematis secara etis. Selain itu, algoritme pemuliaan berbasis data besar sering bias terhadap varietas elit, mempersempit keanekaragaman hayati, serta mengabaikan sistem pertanian tradisional.

Bioetika tanaman, khususnya pada pengembangan tanaman transgenik (GMO), masih menghadapi berbagai kekurangan yang memicu kontroversi ilmiah dan sosial, terutama terkait ketidakpastian dampak jangka panjang terhadap kesehatan manusia, lingkungan, dan keanekaragaman hayati. Isu etika muncul karena pelepasan GMO ke ekosistem berpotensi menyebabkan aliran gen ke tanaman liar, ketergantungan petani pada benih berpaten, serta ketimpangan distribusi manfaat bioteknologi yang lebih menguntungkan korporasi besar

dibandingkan petani kecil. Selain itu, persepsi publik mengenai "ketidakalamian" rekayasa genetika memperkuat resistensi sosial akibat kurangnya komunikasi ilmiah yang transparan dan berbasis bukti. Oleh karena itu, regulasi ke depan perlu menekankan prinsip kehati-hatian, evaluasi risiko yang terbuka dan *reproducible*, perlindungan keanekaragaman hayati, serta pelibatan publik dalam pengambilan keputusan agar inovasi bioteknologi tanaman dapat berkembang secara etis, berkelanjutan, dan adil.

Tujuan mulia bioteknologi yang ditulis perspektif Jonas (1984), seperti mengurangi penderitaan manusia, tidak otomatis membenarkan semua bentuk penerapannya. Sejak lama, kemampuan manusia untuk berinovasi dipahami memiliki dua sisi, yaitu membawa manfaat besar sekaligus berpotensi menimbulkan dampak moral yang serius. Pemikiran Sophocles, Hans Jonas, dan Habermas menunjukkan bahwa teknologi modern terutama bioteknologi menghadirkan skala, konsekuensi, dan jenis persoalan etika yang sama sekali baru, sehingga kerangka etika lama tidak lagi memadai. Oleh karena itu, bioteknologi menuntut evaluasi etika yang lebih mendalam, hati-hati, dan kontekstual terhadap tanggung jawab manusia atas dampak jangka panjang dari tindakannya.

Reproduksibilitas

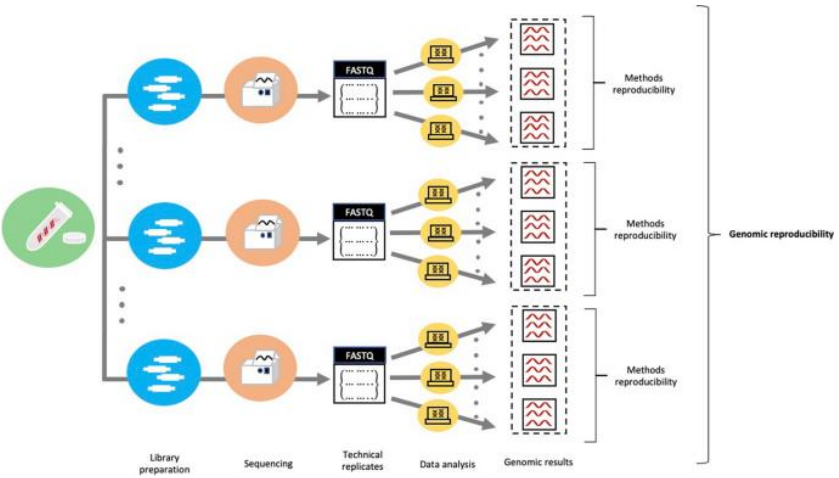
Reproduksibilitas pada bioinformatika adalah kemampuan suatu analisis komputasi untuk menghasilkan kembali hasil yang sama atau setara ketika data, metode, dan parameter yang digunakan diulang oleh peneliti lain. Dalam bioinformatika modern, reproduksibilitas menjadi isu krusial karena hasil penelitian sangat bergantung pada *pipeline* komputasi,

algoritme, basis data, dan konfigurasi perangkat lunak, sehingga ketidakjelasan metode dapat menurunkan kredibilitas ilmiah dan memunculkan persoalan etika. Oleh karena itu, pembahasan reproduksibilitas tidak hanya berkaitan dengan aspek teknis analisis data, tetapi juga menyangkut integritas ilmiah, transparansi, serta tanggung jawab peneliti dalam memastikan bahwa temuan bioinformatika dapat diverifikasi dan digunakan secara aman di berbagai sektor.

Reproduksibilitas dalam penelitian genomik sangat bergantung pada prosedur eksperimental dan metode komputasi yang digunakan. Banyaknya tahapan, mulai dari persiapan sampel dan pengurutan hingga analisis bioinformatika, serta adanya variasi eksperimental, menjadi tantangan utama dalam menghasilkan hasil genomik yang konsisten dan andal (Baykal *et al.*, 2024).

Penelitian Baykal *et al.* (2024), Gambar 14.1 menunjukkan bahwa sampel biologis yang sama diproses berulang mulai dari persiapan pustaka → sekuensing → analisis data. Ketika satu sampel disekuensing beberapa kali (bahkan di laboratorium berbeda) dengan protokol yang sama, hasil FASTQ yang dihasilkan disebut *technical replicates*. Variasi di tahap ini mencerminkan *noise* teknis, bukan perbedaan biologis. Selanjutnya, FASTQ yang sama dianalisis berulang menggunakan *pipeline* komputasi (*alignment*, *filtering*, *variant calling*, dsb.). Konsistensi hasil dari analisis berulang inilah yang disebut metode yang reprodusibel (mampu direproduksi). Konsep ini mengacu pada apakah *pipeline* bioinformatika menghasilkan *output* yang stabil ketika dijalankan ulang. Terakhir, reproduksibilitas genomik menilai apakah hasil genomik akhir (misalnya gen diferensial, SNP, atau pola ekspresi) tetap

konsisten antar *technical replicates*, sehingga benar-benar merepresentasikan sinyal biologis yang nyata.



Sumber: Baykal *et al.* (2024)

Gambar 14.1. Skema konsep ulangan teknis, reproduksibilitas metode, dan reproduksibilitas genomik dalam analisis bioinformatika

Kaitannya dengan reproduksibilitas bioinformatika, gambar ini menegaskan bahwa masalah reproduksibilitas tidak hanya berasal dari data basah (*wet-lab*), tetapi sangat kuat dipengaruhi oleh analisis komputasi. Dua peneliti bisa menggunakan data FASTQ yang sama, tetapi menghasilkan kesimpulan biologis berbeda hanya karena perbedaan parameter, versi *software*, atau workflow yang tidak terdokumentasi dengan baik. Oleh karena itu, reproduksibilitas bioinformatika menuntut transparansi pipeline, standardisasi metode, dan dokumentasi lengkap, agar perbedaan hasil dapat ditelusuri apakah berasal dari variasi teknis, metode analisis, atau benar-benar perbedaan biologis. Inilah alasan mengapa reprodusibel *pipelines* dan pelaporan metode yang rinci menjadi fondasi etika dan kredibilitas

bioinformatika modern, terutama saat hasilnya digunakan untuk keputusan medis, industri, maupun pertanian.

- Biomedis

  Dalam riset klinis dan genomik medis yang sensitif, pipeline reprodusibel memastikan bahwa hasil analisis (seperti varian genetik atau ekspresi gen yang terkait dengan penyakit) bukan sekadar artefak dari versi *software* atau parameter yang berbeda, tetapi dapat diverifikasi ulang di laboratorium lain penting untuk kepercayaan diagnosis dan rekomendasi terapi.

- Bioindustri

  Dalam industri biofarmasi atau bioteknologi, *reproducible workflows* memungkinkan audit internal dan eksternal atas proses analisis data besar (*big data*), sekaligus mendukung standardisasi rantai produksi dan validasi temuan, sehingga produk dapat memenuhi standar kualitas dan regulasi.

- Pertanian

  Untuk analisis genom tanaman atau mikroba pertanian, reproduksibilitas memastikan interpretasi data (misalnya *genomic selection* atau QTL *mapping*) konsisten di berbagai lingkungan komputasi, sehingga varietas unggul yang dihasilkan dapat diandalkan dan diuji ulang oleh peneliti lain di lokasi berbeda.

Etika integritas data dalam bioinformatika menuntut peneliti untuk tidak memanipulasi data maupun memilih hasil analisis semata-mata demi mencapai signifikansi statistik, karena praktik

semacam ini sering tersembunyi di balik kompleksitas *pipeline* komputasi dan dapat luput dari pengawasan. Oleh sebab itu, akademisi dan praktisi bioinformatika memiliki tanggung jawab etis untuk menerapkan transparansi penuh dalam pelaporan metode, parameter, dan sumber data, serta memastikan penggunaan set data yang relevan secara biologis, agar hasil analisis dapat direproduksi secara ilmiah dan dipertanggungjawabkan. Khususnya ketika temuan tersebut digunakan sebagai dasar pengambilan keputusan di bidang medis, industri, dan pertanian.

Kredibilitas hasil bioinformatika sangat bergantung pada reproduksibilitasnya, karena hasil yang tidak dapat direplikasi berisiko menimbulkan dampak serius. Dalam medis dapat memicu diagnosis keliru dan klaim AI yang *overclaim*, dalam industri menghambat validasi dan kepatuhan regulasi, serta dalam pertanian berpotensi menghasilkan penerapan genomik yang salah di lapangan. Oleh karena itu, reproduksibilitas merupakan prasyarat etis agar bioinformatika dapat diterapkan secara aman, andal, dan bertanggung jawab.

Masa Depan Bioinformatika Bidang Medis, Industri, dan Pertanian

Bidang Biomedis

Menurut penelitian Vidanagamachchi *et al.* (2024), penyakit tropis yang disebabkan oleh virus, bakteri, parasit, dan jamur memerlukan pendekatan analisis yang komprehensif untuk memahami mekanisme penyakit dan pola penyebarannya. Pemanfaatan data *omics* yang dikombinasikan dengan alat bioinformatika dan teknik kecerdasan buatan (AI) telah menjadi strategi penting dalam identifikasi dan prediksi penyakit tropis.

Tinjauan literatur periode 2015–2024 menunjukkan bahwa integrasi berbagai data *omics* dengan metode bioinformatika dan model AI (termasuk pendekatan *explainable* AI dan model berbasis *transformer*), mampu meningkatkan akurasi identifikasi *biomarker*, prediksi wabah, serta mendukung pengembangan terapi dan pengeditan gen yang lebih efisien.

Studi Grigoriadis *et al.* (2022), menegaskan peran penting metode komputasi berbasis AI dalam meningkatkan kualitas data genomik, bukan hanya untuk analisis lanjutan, tetapi juga untuk validitas biologis hasilnya. Ke depan, integrasi model *deep learning* yang transparan dan tervalidasi secara biologis akan sangat krusial agar interpretasi TSS dan regulasi gen tidak bias oleh *noise* teknis, sekaligus meningkatkan keandalan data transkriptomik dalam riset dan aplikasi klinis.

Bidang Bioindustri

Menurut Gargalo *et al.* (2020), bioinformatika diproyeksikan menjadi komponen kunci dalam bioteknologi industri dengan mendukung biomanufaktur cerdas yang berbasis data. Melalui pemanfaatan *big data omics* dan integrasi dengan AI serta sistem *Industry 4.0,* bioinformatika memungkinkan pemodelan, pemantauan *real-time*, dan optimasi otomatis proses bioproduksi. Pendekatan ini meningkatkan efisiensi, menurunkan biaya dan limbah, serta mendukung produksi yang lebih berkelanjutan di sektor biofuel, bahan kimia, enzim industri, dan pangan.

Bidang Pertanian

Penelitian Prajval *et al.* (2024) menjelaskan bahwa kemajuan bioinformatika telah mentransformasi pertanian modern melalui

integrasi komputasi dan genomika tanaman. Analisis data genomik berkecepatan tinggi memungkinkan pemuliaan presisi, identifikasi penanda genetik, serta percepatan pengembangan varietas tanaman yang lebih produktif, tahan penyakit, dan adaptif terhadap cekaman lingkungan. Pendekatan ini tidak hanya meningkatkan ketahanan pangan, tetapi juga mendukung pertanian berkelanjutan dengan mengurangi kebutuhan input dan dampak lingkungan.

Penelitian Zhang *et al.* (2022) memaparkan mengenai perkembangan NGS yang mendorong munculnya pendekatan multi-omik sebagai strategi efektif dalam pemuliaan tanaman. Integrasi data genomik, transkriptomik, proteomik, epigenomik, metabolomik, dan mikrobiom (melalui bioinformatika dan biologi sistem) memungkinkan pemahaman hubungan gen–fenotipe serta prediksi sifat tanaman yang kompleks. Pendekatan multi-omik yang dipadukan dengan kecerdasan buatan berpotensi mendukung pemuliaan presisi yang lebih cerdas dan mempercepat pengembangan varietas unggul, meskipun masih menghadapi tantangan dalam integrasi dan interpretasi data.

Bioinformatika memiliki peran strategis dalam mendukung pemuliaan presisi, pertanian cerdas, dan pembangunan berkelanjutan melalui integrasi AI dan *big data*. Namun, kemajuan ini juga dihadapkan pada isu *biopiracy*, ketimpangan akses teknologi, serta keadilan bagi petani kecil dan negara berkembang. Keberhasilan penerapannya bergantung pada validitas analisis genomik, reproduktibilitas, dan keterkaitannya dengan uji lapang agar tidak menimbulkan risiko terhadap produksi dan ketahanan pangan. Ke depan, bioinformatika perlu diposisikan sebagai infrastruktur digital lintas sektor yang diatur secara etis dan adaptif, dengan kebijakan yang melindungi

sumber daya genetik lokal, menjamin transparansi algoritme, dan memastikan keberlanjutan.

Tantangan dan Peluang Implementasi

Meskipun bioinformatika medis semakin maju, tantangan ke depan justru semakin kompleks. Isu yang saat ini banyak dibicarakan adalah keamanan dan privasi data genomik pasien, bias algoritme AI dalam diagnosis, serta rendahnya *explainability* model yang digunakan dalam pengambilan keputusan klinis. Ketergantungan pada infrastruktur komputasi besar dan vendor teknologi global juga berisiko menciptakan ketimpangan layanan kesehatan. Ke depan, tantangan lain mencakup validasi klinis hasil analisis genomik lintas populasi, tanggung jawab hukum atas kesalahan prediksi AI, serta kesenjangan kompetensi tenaga medis dalam memahami hasil bioinformatika.

Di sektor industri, kemajuan bioinformatika memunculkan tantangan baru berupa ketergantungan pada otomatisasi dan AI, risiko kegagalan sistem berbasis data, serta isu kepemilikan dan keamanan *industrial biological data*. Topik yang banyak dibahas saat ini adalah ketahanan rantai pasok berbasis data, *vendor lock-in*, dan transparansi algoritme dalam proses produksi. Ke depan, industri juga akan menghadapi tekanan regulasi, tuntutan keberlanjutan, serta kebutuhan SDM yang mampu menggabungkan pemahaman biologis, komputasi, dan bisnis secara simultan.

Dalam pertanian, meskipun teknologi genomik dan pertanian presisi berkembang pesat, tantangan yang mengemuka adalah kesenjangan adopsi teknologi, ketergantungan pada data dan platform digital, serta potensi marginalisasi petani kecil. Isu yang tengah ramai dibicarakan

mencakup *biopiracy*, kepemilikan data genetik tanaman lokal, dan ketidakpastian hasil prediksi genomik ketika dihadapkan pada kondisi lapang yang dinamis akibat perubahan iklim. Ke depan, tantangan utama adalah menjamin reproduksibilitas seleksi genom, integrasi data bioinformatika dengan uji lapang jangka panjang, serta memastikan bahwa inovasi benar-benar meningkatkan ketahanan pangan, bukan hanya efisiensi teknologi.

Secara keseluruhan, kemajuan bioinformatika tidak otomatis menghilangkan tantangan, melainkan menggesernya ke isu yang lebih struktural dan etis. Oleh karena itu, kolaborasi lintas sektor, penguatan regulasi adaptif, serta pembangunan kapasitas SDM menjadi kunci agar bioinformatika dapat berkembang secara bertanggung jawab dan inklusif.

DAFTAR PUSTAKA

Asveld, L., Osseweijer, P., & Posada, J. A. (2019). Societal and ethical issues in industrial biotechnology. *Advances in Biochemical Engineering/Biotechnology*, 121–141. https://doi.org/10.1007/10_2019_100

Baykal, P. I., Łabaj, P. P., Markowetz, F., Schriml, L. M., Stekhoven, D. J., Mangul, S., & Beerenwinkel, N. (2024). Genomic reproducibility in the bioinformatics era. *Genome Biology, 25*(1). https://doi.org/10.1186/s13059-024-03343-2

Coghlan, S., Gyngell, C., & Vears, D. F. (2023). Ethics of artificial intelligence in prenatal and pediatric genomic medicine. *Journal of Community Genetics, 15*(1), 13–24. https://doi.org/10.1007/s12687-023-00678-4

Gargalo, C. L., Udugama, I., Pontius, K., Lopez, P. C., Nielsen, R. F., Hasanzadeh, A., Mansouri, S. S., Bayer, C., Junicke, H., & Gernaey, K. V. (2020). Towards smart biomanufacturing: A perspective on recent developments in industrial measurement and monitoring technologies for bio-based production processes. *Journal of Industrial Microbiology & Biotechnology, 47*(11), 947–964. https://doi.org/10.1007/s10295-020-02308-1

Greenhill, A. T., & Edmunds, B. R. (2020). A primer of artificial intelligence in medicine. *Techniques and Innovations in Gastrointestinal Endoscopy, 22*(2), 85–89. https://doi.org/10.1016/j.tgie.2019.150642

Grigoriadis, D., Perdikopanis, N., Georgakilas, G. K., & Hatzigeorgiou, A. G. (2022). DeepTSS: Multi-branch convolutional neural network for transcription start site identification from CAGE data. *BMC Bioinformatics, 23*,

Article 356. https://doi.org/10.1186/s12859-022-04945-y

Jonas, H. (1984). *The imperative of responsibility: In search of an ethics for the technological age*. University of Chicago Press.

Prajval, V., Krishnamoorthi, A., Thakur, R., Ruchitha, T., Kapoor, M., Singh, A., & Chittibomma, K. (2024). From code to crop: How bioinformatics is transforming crop genomics in modern agriculture and bettering environment. *Journal of Advances in Biology & Biotechnology, 27*(4), 27–49. https://doi.org/10.9734/jabb/2024/v27i4737

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Shaw, J., Ali, J., Atuire, C. A., Cheah, P. Y., Español, A. G., Gichoya, J. W., Hunt, A., Jjingo, D., Littler, K., Paolotti, D., & Vayena, E. (2024). Research ethics and artificial intelligence for global health: Perspectives from the global forum on bioethics in research. *BMC Medical Ethics, 25*(1). https://doi.org/10.1186/s12910-024-01044-w

Vidanagamachchi, S. M., & Waidyarathna, K. M. G. T. R. (2024). Opportunities, challenges and future perspectives of using bioinformatics and artificial intelligence techniques on tropical disease identification using omics data. *Frontiers in Digital Health, 6*. https://doi.org/10.3389/fdgth.2024.1471200

Zhang, R., Zhang, C., Yu, C., Dong, J., & Hu, J. (2022). Integration of multi-omics technologies for crop improvement: Status and prospects. *Frontiers in Bioinformatics, 2*. https://doi.org/10.3389/fbinf.2022.1027457

PROFIL PENULIS

Yusril Ihza Farhan Wijaya, M.Sc. adalah peneliti di BRC-INBIO Indonesia bidang keahliannya yaitu biologi molekuler, bioteknologi, omics, dan bioinformatika. Saat ini, ia aktif melakukan penelitian tentang genomik dan transkriptomik.
Email korespondensi: yusrilihzafarhanwijaya@gmail.com

Sonia Meta Angraini, S.Si., M.Biotech adalah dosen pada Program Studi Pendidikan Kimia, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Sembilanbelas November Kolaka. Bidang keahliannya meliputi biologi molekuler dan bioteknologi. Saat ini, ia aktif melakukan penelitian tentang fitofarmaka bahan alam sumber daya lokal sebagai agen antikanker.
Email korespondensi: soniameta65@gmail.com

Armansyah Maulana Harahap, S.Si, M.Biomed adalah seorang dosen Universitas Negeri Medan dengan Bidang keahlian Farmakologi dan Toksikologi. Saat ini aktif melakukan riset dengan Badan Riset Inovasi Nasional (BRIN) dengan topik penelitian Kajian dan Potensi Bahan Alam terhadap penyakit Tropis dan Sindrom Metabolik dengan menggunakan pendekatan *in silico* (Komputasi), *in vitro* meliputi Bioassay dan Enzimatis dan *in vivo*
Email korespondensi: armansyahhrp@unimed.ac.id

Kiky Mey Putranty S.Si., M.Si adalah peneliti di BRC-INBIO Indonesia. Bidang keahliannya meliputi Mirobiologi, Bioteknologi dan Bioinformatika Tanaman. Saat ini, ia aktif melakukan penelitian tentang study *in silico* senyawa aktif mikroalga.
Email korespondensi: kikymey@inbio-indonesia.org

Rini Hafzari, S.Si., M.Si lahir pada tahun 1991 di Tanjung Jabung Barat, Jambi. Penulis menyelesaikan studi sarjana pada tahun 2015 di Universitas Padjadjaran dan menyelesaikan studi S-2 pada tahun 2018 di Universitas Padjadjaran pada bidang Genetika, dan Biologi molekuler. Penulis memulai karir sebagai Asisten Peneliti di Laboratorium Genetika dan Biologi Molekuler, Departemen Biologi, FMIPA UNPAD. Kemudian pada tahun 2019 penulis menjadi Dosen Luar Biasa di Departemen Biologi, FMIPA UNPAD. Pada tahun 2020-2023 penulis bekerja sebagai aplikasi spesialis di salah satu Perusahaan Bioteknologi dengan fokus produk biologi molekuler, PCR, NGS dan Bioinformatics Solutions. Sekarang penulis bekerja sebagai Dosen di Program Studi Biologi, FMIPA UNIMED. Mata kuliah yang diampu adalah biologi molekuler, biologi sel, bioteknologi, biokimia dan bioinformatik. Fokus penelitian penulis pada bidang pengembangan penanda molekuler, desain primer spesifik, sekuensing, dan ekspresi gen serta pencarian potensi bahan alam untuk obat herbal.

Email korespondensi: rinihafzari@unimed.ac.id

M. Khoiron Ferdiansyah, STP, MSc., PhD adalah dosen dan peneliti di Program Studi Teknologi Pangan, Universitas PGRI Semarang sekaligus auditor halal di Korea Testing Certification (KTC). Fokus risetnya selama program doktoral di Jeonbuk National University, Republic of Korea, mencakup eksplorasi enzim mikroba, khususnya purine nucleosidase dari *Levilactobacillus brevis*, dan senyawa inhibitor xanthine oxidase dari bahan alami untuk pengendalian hiperurisemia. Minat risetnya saat ini mencakup pengembangan enzim dan inhibitor enzim dari mikroorganisme dan bahan alami dan aplikasinya dalam pangan fungsional, dengan pendekatan interdisipliner yang menggabungkan biokimia, biologi molekuler, dan bioteknologi. Email korespondensi: khoironstp@gmail.com

Rolina Anna Erica Sihombing, S.Si., M.Si. adalah staf riset di institusi swasta. Bidang keahliannya meliputi mikrobiologi, mikrobiologi molekuler, dan memiliki ketertarikan pada topik mikrobiomik. Saat ini, ia aktif melakukan penelitian tentang pengembangan produk berbasis mikroba.
Email korespondensi: rolinaerica@gmail.com

dr. Lustyafa Inassani Alifia, M.Biomed.

Lahir di Malang, 16 Mei 1992. Menyelesaikan studi pendidikan dokter dan profesi dokter di Fakultas Kedokteran Universitas Muhammadiyah Malang pada tahun 2017, dan melanjutkan studi Magister Ilmu Biomedik Universitas Brawijaya pada tahun 2021, dengan minat imunologi-parasitologi. Saat ini menjadi dosen aktif di Fakultas Kedokteran Universitas Muhammadiyah Malang. Beberapa publikasi yang dihasilkan adalah terkait imunologi pada malaria, terutama pada *gut immunity*.

Email korespondensi: inassani@umm.ac.id

apt. Lolita,M.Sc., Ph.D adalah adalah Dosen di Fakultas Farmasi Universitas Ahmad Dahlan, Yogyakarta. Saat ini bertugas sebagai Ketua Program Studi Sarjana Farmasi pada tahun 2023 hingga 2026. Gelar Sarjana Farmasi dari Fakultas Farmasi Universitas Ahmad Dahlan Yogyakarta pada tahun 2005, dan gelar Magister Sains (M.Sc) Farmasi dari Universitas Gadjah Mada pada tahun 2013. Gelar Ph.D. Clinical Pharmacy diperoleh dari Nanjing Medical University, Republik

Rakyat Tiongkok, pada tahun 2022. Bidang keahliannya meliputi Farmakoepidemiologi, Farmakogenetik, Farmakoterapi Kimia Obat Infeksi dan Penyakit Kronis, Perilaku dan Luaran Kesehatan. Saat ini, ia aktif melakukan penelitian tentang Farmakoepidemiologi dan Farmakogenetik Terapi Imunosupresan Pada Resipien Transplantasi Ginjal.
Email korespondensi: lolita@pharm.uad.ac.id

Chindy Nur Rosmeita, M.Si. adalah seorang peneliti di Bioinformatics Research Center – Indonesian Institute of Bioinformatics (INBIO) dan salah satu industri kimia di Indonesia. Bidang keahliannya meliputi pengembangan vaksin dengan bahan baku vaksin yang berasal dari protein rekombinan. Saat ini, ia aktif melakukan penelitian tentang pengembangan vaksin secara *in silico*.
Email korespondensi: chindynr@gmail.com

Dr. Eko Kusumawati, S.Si, MP adalah Dosen di Jurusan Biologi, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Mulawarman, Samarinda. Bidang keahliannya adalah Mikrobiologi. Saat ini, aktif melakukan penelitian mengenai (1) Analisis Potensi dan Filogenetik Bakteri Pengoksidasi

Sulfur dari Air Kolam Pascatambang Batubara di Samarinda, Kalimantan Timur; (2) Eksplorasi Bioaktivitas Daun Kokang (*Lepisanthes amoena* (Hassk) Leenh.): Aktivitas Antioksidan dan Antibakteri pada *Cutibacterium acnes* secara *In Vitro;* (3) Profil Metabolit Sekunder Daun Kokang (*Lepisanthes amoena*) menggunakan LC-MS/MS dan Interaksinya terhadap Protein Target *Cutibacterium acnes* melalui Pendekatan *In silico;* (4) Eksplorasi Bakteri Patogen pada Talas (*Colocasia esculenta*); (5) Potensi Bakteri Endofit dan Rhizosfer Talas (*Colocasia esculenta*) sebagai Agen Biokontrol terhadap Patogen Bakteri Penyebab Busuk Daun; (6) Analisis Yeast Potensial Biodegradable Plastik.
Email korespondensi: eko.kusumawati11@fmipa.unmul.ac.id

apt. Rizki Rahmadi Pratama, M. Farm mengemban jabatan sebagai Asisten Ahli di Universitas Islam Kalimantan Muhammad Arsyad Al Banjari Banjarmasin, di mana beliau secara aktif berkontribusi dalam pengembangan akademik dan riset. Bidang kepakarannya meliputi biologi farmasi dan kimia medisinal tanaman, yang menjadi fondasi utama dalam eksplorasi potensi senyawa alami. Saat ini, fokus penelitiannya terarah pada studi metabolomik, khususnya melalui pemanfaatan kromatografi cair-spektrometri massa tandem resolusi tinggi (LC-MS/MS-QToF) untuk menganalisis profil metabolit dari berbagai spesies tanaman. Pendekatan ini bertujuan untuk mengidentifikasi dan mengkarakterisasi senyawa bioaktif yang berpotensi memiliki aplikasi terapeutik.

Selain itu, beliau juga aktif dalam melakukan uji *in-silico*, sebuah metode komputasi yang digunakan untuk memprediksi interaksi molekuler dan skrining awal potensi farmakologis senyawa, sehingga mempercepat proses penemuan obat.
Email korespondensi: rizkirahmadi_p@uniska-bjm.ac.id

Muhamad Firmanda, S.Si. adalah mahasiswa Magister Ilmu Biomedik di Universitas Gadjah Mada. Bidang keahliannya meliputi pencarian *biomarker* prognostik dan prediktif untuk kanker payudara. Saat ini, ia aktif melakukan penelitian tentang *plasma circulating microRNA* sebagai biomarker prediktor rekurensi kanker payudara subtipe luminal ($HR^+/HER2^-$).
Email korespondensi: muhamadfirmanda01@gmail.com

Pauline Elisabeth, S.Agr., M.Si. adalah seorang peneliti muda di bidang sains bioteknologi tumbuhan yang memiliki ketertarikan kuat pada integrasi riset laboratorium dan penerapannya di lapangan. Telah menyelesaikan pendidikan Sarjana Agroteknologi (S.Agr.) di Universitas Padjadjaran dan saat ini sedang menempuh program Magister Bioteknologi (M.Si.) di

Institut Teknologi Bandung. Saat ini berfokus pada penelitian ekspresi gen terkait regulasi dan optimasi jalur biosintesis steviol glikosida dan jalur antioksidan melalui elisitasi metil jasmonat pada tanaman *Stevia rebaudiana* (Bert.) secara *in vitro*. Penelitian ini bertujuan untuk memahami mekanisme molekuler peningkatan metabolit bernilai tinggi dan respons stres tanaman sebagai dasar pengembangan teknologi produksi stevia yang lebih efisien dan berkelanjutan. Dalam kegiatan akademik dan pengabdian masyarakat telah banyak terlibat meneliti di kawasan transmigrasi. Ia meyakini bahwa desa merupakan sumber ide kreatif dan inovatif, yang mampu menjembatani wawasan laboratorium dengan kebutuhan nyata di lapangan. Ketertarikan ini juga mendorongnya untuk menuangkan gagasan ilmiah ke dalam buku dan karya tulis populer berbasis sains terapan.
Email korespondensi: pauline.elisabeth00@gmail.com

FUTURE SCIENCE

PENGANTAR BIOINFORMATIKA

ISBN 978-621-8438-20-0