La-Net

BIO 413

BIOINFORMATICS

Downoad on the App Store

GET IT ON Google Play

Course video is available on our app

lanetonline.com

Course code: BIO 413

Course title: Bioinformatics

Course credit load: 2 Units

Series 1: History of Bioinformatics

- **Part 1.1: Origins of Bioinformatics**
    - Sub Part 1.1.1: Defining bioinformatics
    - Sub Part 1.1.2: Early foundations and the first biological databases
    - Sub Part 1.1.3: The role of Margaret Dayhoff and early sequence analysis
- **Part 1.2: The Emergence of Modern Bioinformatics**
    - Sub Part 1.2.1: The genomics revolution and the Human Genome Project
    - Sub Part 1.2.2: Key algorithms and computational breakthroughs
    - Sub Part 1.2.3: The rise of public biological databases
- **Part 1.3: Bioinformatics in the Twenty-First Century**
    - Sub Part 1.3.1: Next-generation sequencing and big data in biology
    - Sub Part 1.3.2: Bioinformatics in medicine, agriculture, and industry
    - Sub Part 1.3.3: Bioinformatics in Nigeria and Africa
- **Part 1.4: Scope and Subdisciplines of Bioinformatics**
    - Sub Part 1.4.1: Core subdisciplines of bioinformatics
    - Sub Part 1.4.2: Bioinformatics and related disciplines
    - Sub Part 1.4.3: Career pathways in bioinformatics

Introduction

Bioinformatics has transformed biology from a largely descriptive science into a quantitative, data-intensive discipline capable of addressing questions about the molecular basis of life at a scale and speed unimaginable just decades ago. It sits at the intersection of biology, computer science, mathematics, and statistics, providing the computational and analytical tools needed to store, retrieve, analyse, and interpret the vast quantities of molecular biological data generated by modern genomic and proteomic technologies.

This Series traces the historical development of bioinformatics from its earliest roots in the compilation of protein sequence data in the 1960s through the pivotal genomics revolution and the completion of the Human Genome Project, to the era of next-generation sequencing and big data biology in the twenty-first century. It also establishes the scope and subdisciplines of bioinformatics and situates the discipline in the Nigerian and African context, providing the conceptual foundation on which the technical content of the remaining Series of this course builds.

Series Learning Outcomes

Upon completing this Series, you should be able to:

- **SLO 1.1** define bioinformatics and describe its early origins, including the contributions of Margaret Dayhoff and the first biological databases,
- **SLO 1.2** describe the genomics revolution, the Human Genome Project, and the key computational breakthroughs that shaped modern bioinformatics,
- **SLO 1.3** describe the applications of bioinformatics in the twenty-first century, including in medicine, agriculture, and the Nigerian and African context, and
- **SLO 1.4** identify the core subdisciplines of bioinformatics and explain its relationship with related disciplines and career pathways.

1.1 Origins of Bioinformatics

1.1.1 Defining bioinformatics

Bioinformatics is the interdisciplinary science that develops and applies computational methods, mathematical models, and statistical approaches to store, retrieve, analyse, and interpret biological data, particularly molecular biological data such as nucleotide and amino acid sequences, protein structures, and gene expression profiles; the term was coined by Paulien Hogeweg and Ben Hesper in 1970 to describe the study of information processes in biotic systems, though the practical application of computing to biological sequence data predates the coinage of the term by nearly a decade.

The scope of bioinformatics extends from the management of large biological databases through the development of algorithms for sequence alignment and structure prediction, to the systems-level analysis of entire genomes, transcriptomes, and proteomes; it is distinct from computational biology (which is more broadly concerned with mathematical modelling of biological systems) and from biomedical informatics (which focuses primarily on clinical and health data), though these disciplines overlap substantially and the boundaries between them are not universally agreed.

Fill in the blank: The term bioinformatics was coined by Paulien Hogeweg and Ben Hesper in _____ to describe the study of information processes in biotic systems.

1.1.2 Early foundations and the first biological databases

The foundations of bioinformatics were laid in the early 1960s with the first systematic efforts to collect and organise biological macromolecule sequence data; the development of protein sequencing methods (principally by Frederick Sanger, who determined the first complete protein sequence, that of bovine insulin, in 1951 using his end-group analysis and partial hydrolysis methods) generated the data that needed organising, and the simultaneous development of electronic computers created the tools with which to organise it; the combination of biological data generation and computational storage capacity created the conditions from which bioinformatics as a discipline emerged.

Early biological databases were entirely paper-based before their transfer to machine-readable formats: the Atlas of Protein Sequence and Structure (initiated by Margaret Dayhoff in 1965, later known as the Protein Information Resource, PIR) was the first systematically organised collection of protein sequence data, initially compiled and distributed as a book before being transferred to computer-readable magnetic tape; this early database established many of the principles of biological database design that persist to the present, including the importance of standardised formats, comprehensive annotation, and free availability of data to the scientific community.

Fill in the blank: The Atlas of Protein Sequence and Structure, initiated by Margaret Dayhoff in _____ , was the first systematically organised collection of protein sequence data.

1.1.3 The role of Margaret Dayhoff and early sequence analysis

Margaret Dayhoff (1925 to 1983) is widely regarded as the founder of bioinformatics, having pioneered the creation of protein sequence databases, developed computational methods for their analysis, and established many of the conceptual frameworks that still underpin the field; her most significant scientific contributions include the development of the first amino acid substitution matrices (the PAM matrices, Percent Accepted Mutations, constructed from aligned protein sequences and quantifying the relative rates at which one amino acid replaces another over evolutionary time) which became the foundation for all subsequent work in sequence alignment and database searching.

Dayhoff also developed the single-letter amino acid code (which replaced the three-letter code for computational convenience), contributed to the first computational phylogenetic analyses, and promoted the principle that biological sequence data should be collected in a central repository and made freely available to the entire scientific community; her Atlas was accompanied by early computer programs for sequence comparison, making her work the direct ancestor of modern bioinformatics tools and databases including BLAST, GenBank, UniProt, and the NCBI suite of resources.

Fill in the blank: Margaret Dayhoff developed the PAM (Percent Accepted _____) matrices, which quantify the rates at which amino acids replace each other over evolutionary time and underpin modern sequence alignment.

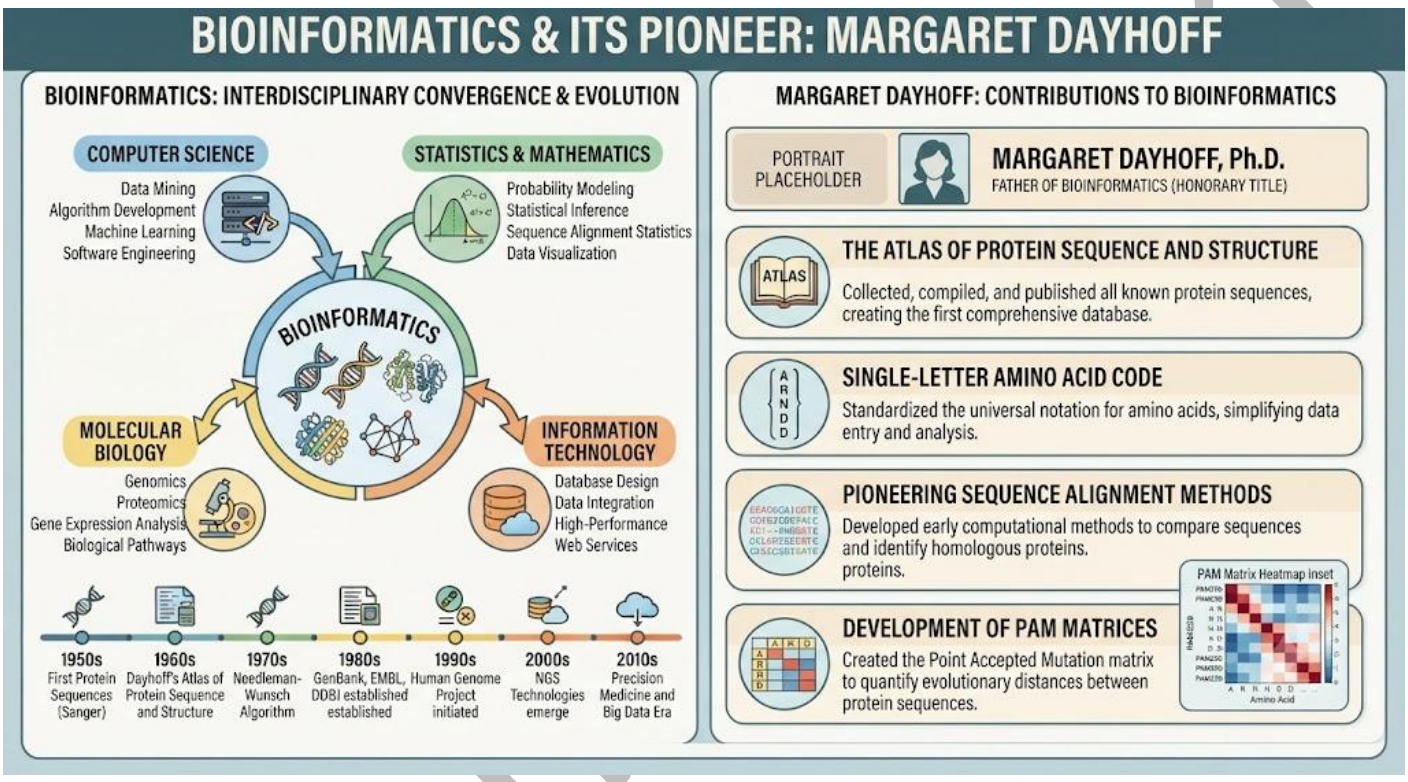

Figure 1.1: Origins of bioinformatics

Pilot questions 1.1

1. Define bioinformatics and state who coined the term and when.
2. Distinguish between bioinformatics and computational biology.
3. Name the first systematically organised protein sequence database and state who created it.
4. Explain what PAM matrices are and why they are important.
5. State two principles of biological database design established by Margaret Dayhoff.

1.2 The Emergence of Modern Bioinformatics

1.2.1 The genomics revolution and the Human Genome Project

The development of automated DNA sequencing, based on Frederick Sanger dideoxy chain-termination method (1977) and later automated using fluorescent dye-labelled terminators and capillary electrophoresis, transformed the scale at which biological sequence data could be generated; the first complete genome sequences of living organisms (the bacteriophage phiX174 in 1977, the bacterium Haemophilus influenzae in 1995, and the yeast Saccharomyces cerevisiae in 1996) demonstrated the feasibility of whole-genome sequencing and established the computational infrastructure for managing and analysing complete genomes.

The Human Genome Project (HGP), initiated in 1990 as an international collaborative effort involving research centres in the United States, United Kingdom, France, Germany, Japan, and China, with the goal of determining the complete sequence of all approximately 3.2 billion base pairs of the human genome, was completed in its draft form in 2001 (announced simultaneously in Nature and Science) and declared essentially complete in 2003; the HGP was transformative not only for its scientific output but for the bioinformatics infrastructure it created, including GenBank, the European Molecular Biology Laboratory nucleotide database (EMBL-Bank), and the DNA Data Bank of Japan (DDBJ), which together form the International Nucleotide Sequence Database Collaboration (INSDC).

Fill in the blank: The Human Genome Project was initiated in _____ as an international collaborative effort to determine the complete sequence of the approximately 3.2 billion base pairs of the human genome.

1.2.2 Key algorithms and computational breakthroughs

Several computational breakthroughs were essential for the development of modern bioinformatics: the Needleman-Wunsch algorithm (1970), the first dynamic programming algorithm for global pairwise sequence alignment, made systematic comparison of biological sequences possible; the Smith-Waterman algorithm (1981), a local alignment variant, identified regions of similarity within longer sequences; and the development of BLAST (Basic Local

Alignment Search Tool) by Altschul and colleagues in 1990 provided a heuristic algorithm capable of searching large sequence databases rapidly while maintaining acceptable sensitivity, becoming the most widely used bioinformatics tool in history.

Other critical algorithmic developments include FASTA (a predecessor to BLAST for rapid database searching), the ClustalW and MUSCLE programs for multiple sequence alignment (essential for phylogenetic analysis and conserved region identification), the development of hidden Markov models (HMMs) for profile-based sequence searching (implemented in the HMMER software suite), and the construction of phylogenetic tree inference algorithms including neighbour-joining, maximum likelihood, and Bayesian methods; together these tools constitute the computational toolkit that made the analysis of genomic data produced by the HGP scientifically productive.

Fill in the blank: _____ (Basic Local Alignment Search Tool), developed by Altschul and colleagues in 1990, became the most widely used bioinformatics tool, enabling rapid database searching with statistical scoring.

1.2.3 The rise of public biological databases

The growth of biological sequence data was accompanied by the establishment of major public databases that remain the core infrastructure of bioinformatics: GenBank (established by the National Center for Biotechnology Information, NCBI, in 1982) is the primary nucleotide sequence repository for the United States and a founding member of the INSDC; the Protein Data Bank (PDB, established 1971) is the single worldwide repository for three-dimensional protein and nucleic acid structure data; and UniProt (Universal Protein Resource, established 2002 through the merger of Swiss-Prot and TrEMBL) is the primary annotated protein sequence database.

These databases are interconnected through cross-referencing systems, so that a nucleotide sequence record in GenBank links to the corresponding protein record in UniProt, which in turn links to the three-dimensional structure in the PDB if available; the databases are updated continuously as new data are submitted by researchers worldwide, and are freely accessible through web interfaces and programmatic application programming interfaces (APIs) that allow

automated retrieval and analysis of large datasets; the principle of mandatory data deposition (requiring submission of sequence data to a public database as a condition of publication) was established in the 1990s and has been essential for making biological data publicly available.

Fill in the blank: UniProt was established in 2002 through the merger of _____ -Prot (a manually curated protein sequence database) and TrEMBL (a computationally annotated database).

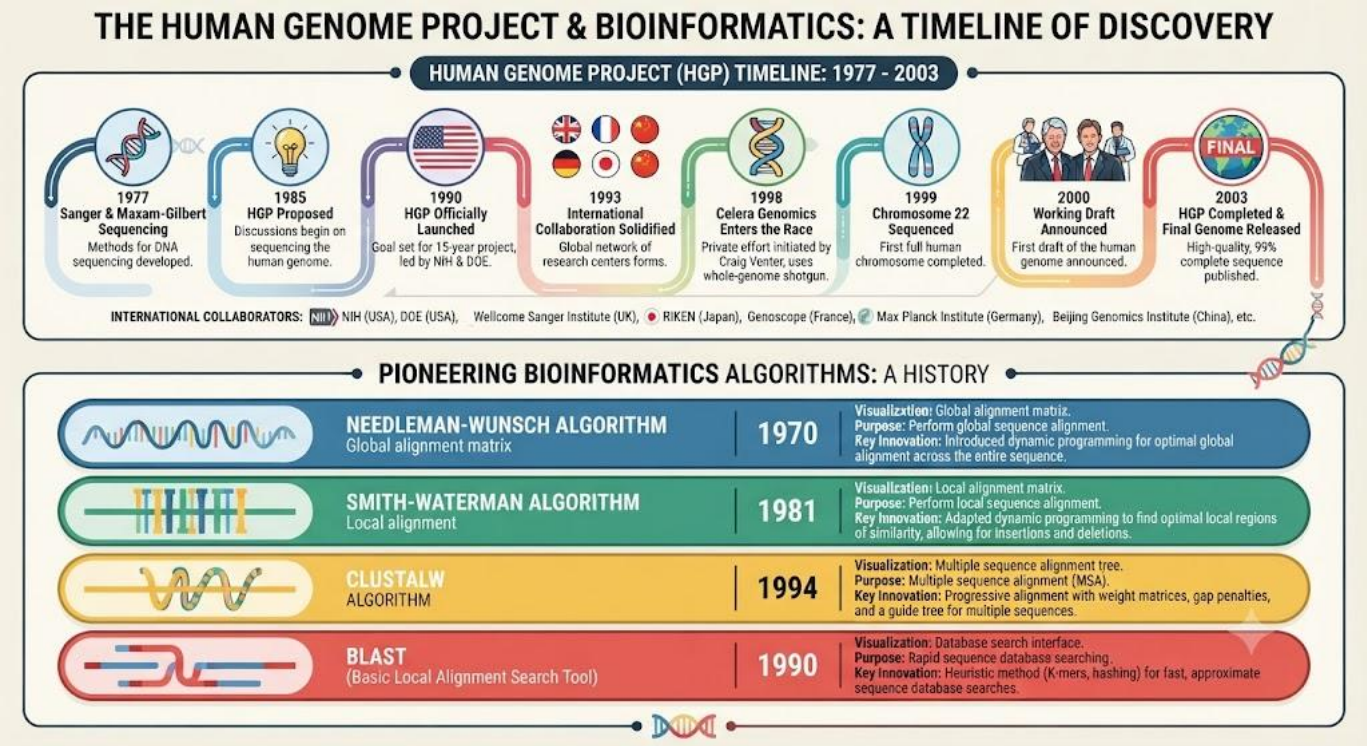

Figure 1.2: The emergence of modern bioinformatics

Pilot questions 1.2

1. State the year the Human Genome Project was initiated and when it was declared essentially complete.
2. Name the three founding members of the International Nucleotide Sequence Database Collaboration.
3. Explain what the Needleman-Wunsch algorithm does and when it was developed.
4. Explain what BLAST is and why it became the most widely used bioinformatics tool.
5. Name the three major public biological databases and state what type of data each holds.

1.3 Bioinformatics in the Twenty-First Century

1.3.1 Next-generation sequencing and big data in biology

The introduction of next-generation sequencing (NGS) platforms in the mid-2000s (beginning with the 454 pyrosequencing platform in 2005, followed by the Illumina sequencing-by-synthesis platform, which rapidly became the dominant technology) reduced the cost and time of DNA sequencing by orders of magnitude compared to Sanger capillary sequencing, generating sequence data at a scale that fundamentally changed the nature of bioinformatics challenges; while a complete human genome required over a decade and approximately three billion US dollars to sequence using first-generation technology, NGS reduced this to a matter of days and thousands of dollars, and continues to become faster and cheaper.

The consequence of NGS for bioinformatics has been a shift from data scarcity (where the limiting factor in biological research was the amount of sequence data available) to data abundance (where the limiting factor is computational capacity and analytical expertise to process and interpret the data generated); this big data challenge has driven the development of new algorithms for sequence assembly, variant calling, and expression analysis, the adoption of cloud computing and high-performance computing clusters for routine bioinformatics analysis, and the application of machine learning approaches to pattern recognition in large biological datasets.

Fill in the blank: Next-generation _____ platforms, introduced from 2005, reduced the cost and time of DNA sequencing by orders of magnitude, fundamentally shifting bioinformatics from data scarcity to data abundance.

1.3.2 Bioinformatics in medicine, agriculture, and industry

Bioinformatics has become essential across multiple sectors beyond academic research: in medicine, clinical bioinformatics encompasses the analysis of patient genome sequences to identify disease-causing variants (whole-genome and whole-exome sequencing in clinical diagnosis), pharmacogenomics (predicting individual patient drug responses from genomic data), infectious disease genomics (sequencing pathogen genomes during outbreaks to track

transmission and identify antimicrobial resistance genes, as demonstrated dramatically during the COVID-19 pandemic with SARS-CoV-2 genomic surveillance), and cancer genomics (identifying somatic mutations driving tumour development to guide targeted therapy).

In agriculture, bioinformatics supports crop improvement through comparative genomics (identifying gene function by comparison with related species), marker-assisted selection (using molecular markers linked to desirable traits to accelerate plant and animal breeding), and the development of disease-resistant or drought-tolerant crop varieties by identifying and manipulating genes responsible for relevant traits; in industry, bioinformatics tools are used in the development of new pharmaceuticals and biotherapeutics (identifying drug targets through structural bioinformatics and protein-protein interaction analysis), biotechnology applications (engineering microbial metabolic pathways for biofuel, food ingredient, or pharmaceutical production), and environmental monitoring (using metagenomic sequencing to characterise microbial communities in soil, water, and industrial processes).

Fill in the blank: _____ genomics uses the sequencing and comparison of pathogen genomes during outbreaks to track transmission, identify resistance genes, and inform public health response.

1.3.3 Bioinformatics in Nigeria and Africa

Africa is increasingly engaged in bioinformatics research and training, driven by recognition that African genomic diversity (Africa contains more human genetic variation than any other continent, reflecting its status as the cradle of anatomically modern humans) is critically underrepresented in global databases and research programmes, and that this underrepresentation has direct consequences for health equity (many medically important genetic variants have only been characterised in populations of European ancestry, limiting the applicability of genomic medicine to other populations); major initiatives to address this include the Human Heredity and Health in Africa (H3Africa) programme and the African BioGenome Project.

In Nigeria specifically, bioinformatics capacity is being developed through university programmes, the growth of genome sequencing infrastructure at institutions including the African Centre of Excellence for Genomics of Infectious Diseases (ACEGID) at Redeemer

University (which played a critical role in African SARS-CoV-2 genomic surveillance and Lassa fever virus genomics), and participation in international training programmes such as those of the Bioinformatics and Next-Generation Sequencing (BNGS) community; Nigerian bioinformaticians and computational biologists are contributing to malaria genomics, HIV genomics, neglected tropical disease research, and agricultural crop improvement, demonstrating the direct relevance of bioinformatics training to Nigerian national health and economic development priorities.

Fill in the blank: Africa contains more human genetic _____ than any other continent, but this variation is critically underrepresented in global databases, limiting the applicability of genomic medicine to African populations.

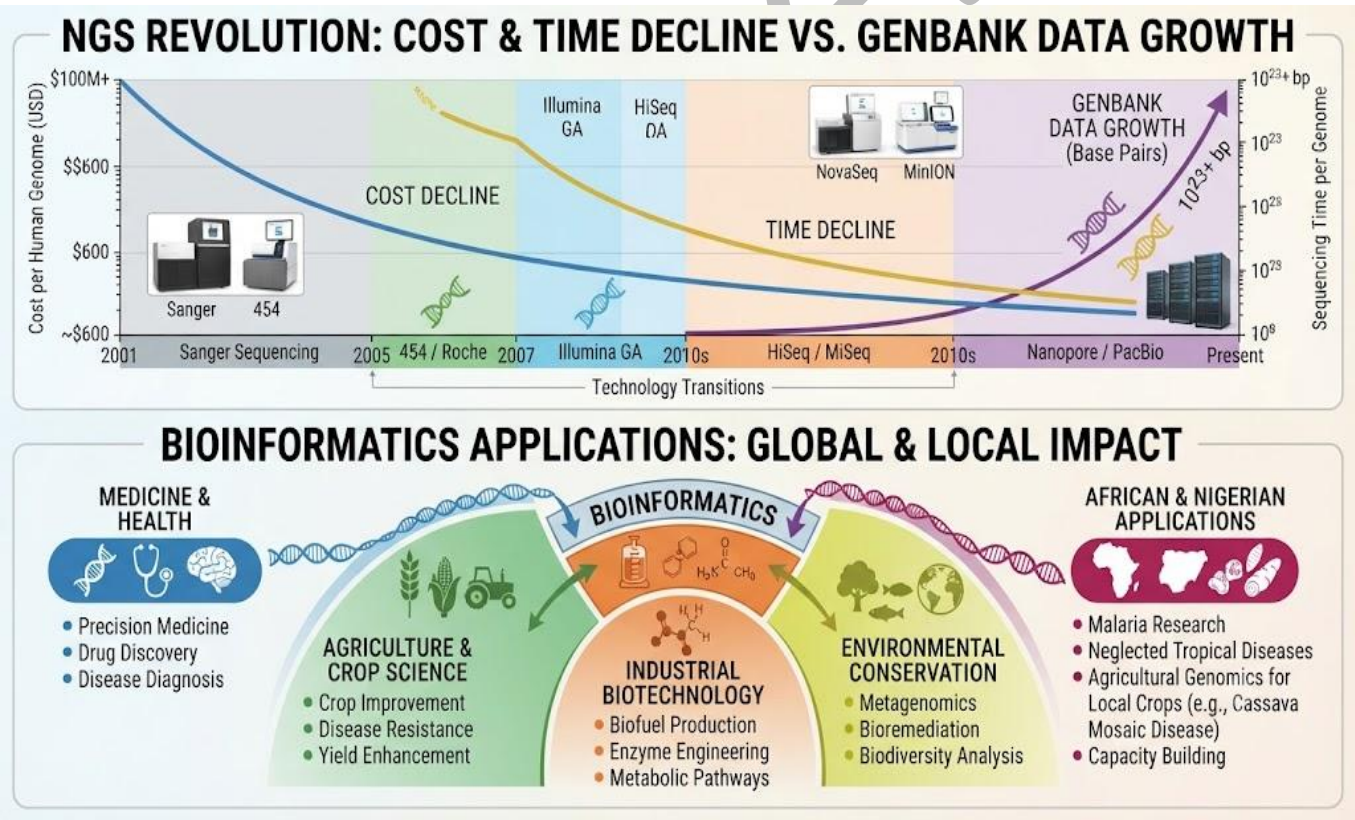

Figure 1.3: Bioinformatics in the twenty-first century

Pilot questions 1.3

1. Explain what next-generation sequencing is and how it changed bioinformatics.
2. Name two medical applications of bioinformatics in the twenty-first century.

3. Explain how bioinformatics is used in crop improvement.
4. Explain why African genomic diversity is underrepresented in global databases and why this matters.
5. Name one Nigerian institution contributing to bioinformatics and state its area of research.

1.4 Scope and Subdisciplines of Bioinformatics

1.4.1 Core subdisciplines of bioinformatics

Bioinformatics encompasses several core subdisciplines that address distinct types of biological data and analytical problems: sequence bioinformatics (the storage, alignment, comparison, and analysis of nucleotide and amino acid sequences, the oldest and still most central subdiscipline); structural bioinformatics (the prediction and analysis of the three-dimensional structures of biological macromolecules, including protein structure prediction from amino acid sequences using approaches such as homology modelling and the revolutionary AlphaFold deep learning system); and functional genomics (the large-scale analysis of gene expression, regulation, and function using techniques such as RNA-seq transcriptomics, ChIP-seq for chromatin immunoprecipitation, and ATAC-seq for chromatin accessibility).

Additional major subdisciplines include phylogenomics (the use of genome-scale sequence data to reconstruct evolutionary relationships among organisms), metagenomics (the sequencing and analysis of DNA extracted directly from environmental samples without prior cultivation, revealing the diversity and function of microbial communities), proteomics bioinformatics (the analysis of mass spectrometry data to identify and quantify proteins in biological samples), and systems biology (the computational modelling of interactions among biological components including genes, proteins, and metabolites to understand emergent system-level properties); each subdiscipline has its own specialised databases, algorithms, and software tools, though all share the common foundation of computational analysis of biological data.

Fill in the blank: _____ bioinformatics involves the prediction and analysis of three-dimensional structures of biological macromolecules from their sequences, including through approaches such as homology modelling and AlphaFold.

1.4.2 Bioinformatics and related disciplines

Bioinformatics is closely related to and overlaps with several other disciplines: computational biology (which overlaps substantially with bioinformatics but has historically emphasised mathematical modelling of biological processes, such as population genetics models, developmental patterning models, and network models of gene regulation, rather than database-oriented sequence analysis); systems biology (which applies engineering and network science approaches to understand biological systems as integrated networks of interacting components, using bioinformatics tools to generate the data and models it analyses); and biomedical informatics (which applies informatics methods to clinical and health data, including electronic health records, medical imaging data, and patient genomics).

Bioinformatics also draws heavily on machine learning and artificial intelligence (in particular deep learning approaches that have revolutionised protein structure prediction with AlphaFold, image analysis in pathology, and variant interpretation in clinical genomics), statistics (for the design and analysis of genomic experiments, including methods for multiple testing correction, differential expression analysis, and population genetic analysis), and database management and software engineering (for the design, maintenance, and querying of the large biological databases that form the data infrastructure of bioinformatics).

Fill in the blank: _____ biology applies engineering and network science approaches to biological systems as integrated networks, using bioinformatics tools to generate and model system-level data.

1.4.3 Career pathways in bioinformatics

Bioinformatics offers a diverse range of career pathways across academic research, healthcare, agriculture, industry, and government; in academic research, bioinformaticians work as principal investigators leading their own research groups, as postdoctoral researchers and research

scientists contributing to collaborative genomics and computational biology projects, and as bioinformatics core facility staff providing analytical services to experimental biologists; in healthcare, clinical bioinformaticians work in hospital genomics laboratories, analysing patient genome data for diagnostic purposes, and in public health agencies conducting genomic epidemiology of infectious disease outbreaks.

In industry, bioinformaticians work in pharmaceutical and biotechnology companies contributing to drug discovery, clinical trial design, and the interpretation of regulatory genomics data; in agriculture and food science, in seed companies, animal genetics organisations, and food safety agencies; and in government, in national public health institutes, research funding agencies, and regulatory bodies; the skills most valued in all these contexts combine domain biological knowledge with programming ability (particularly in Python and R), statistical understanding, and experience with standard bioinformatics tools and databases; for Nigerian graduates, bioinformatics skills open pathways in national health institutes (NIMR, NCDC), agricultural research (IAR, IITA), and international health and development organisations.

Fill in the blank: The programming languages most valued in bioinformatics careers are _____ and R, combined with statistical understanding and domain biological knowledge.

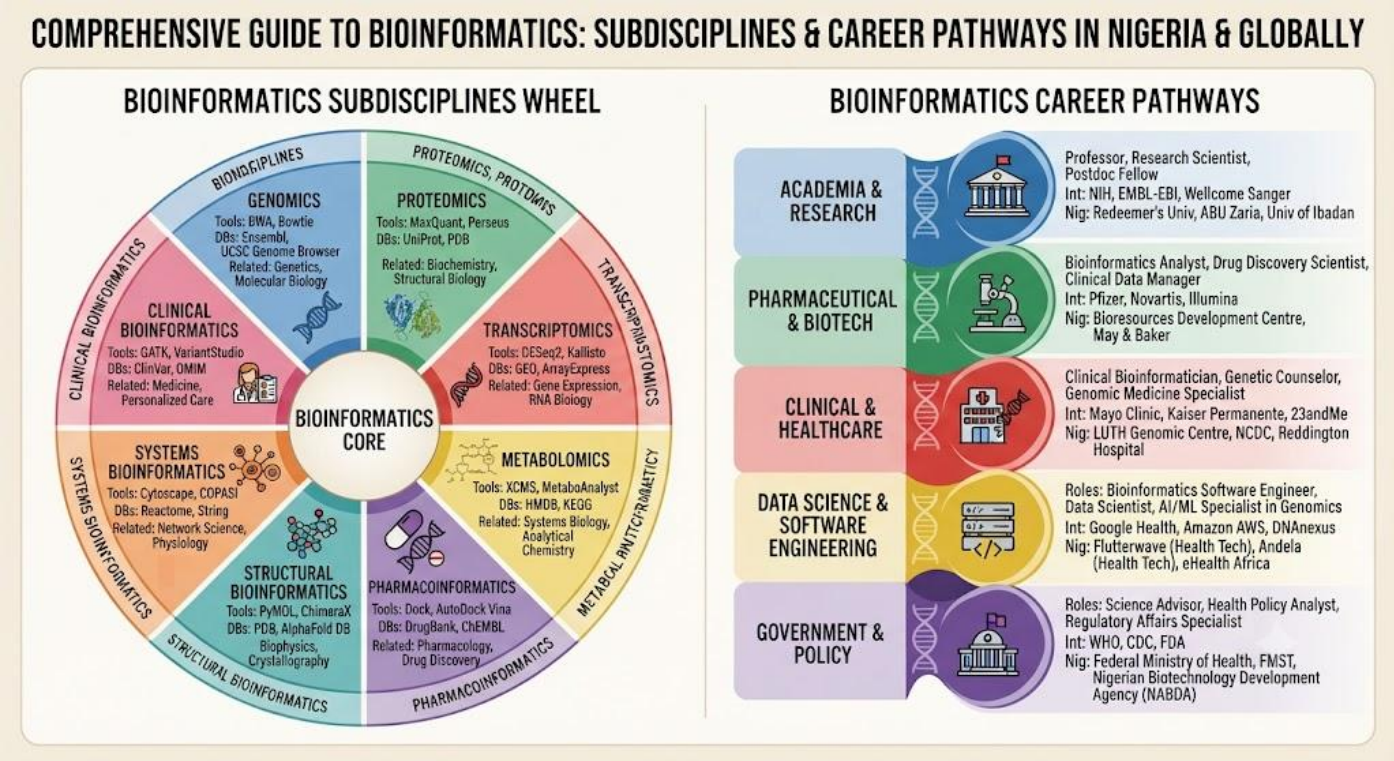

Figure 1.4: Scope and career pathways in bioinformatics

Pilot questions 1.4

1. Name four core subdisciplines of bioinformatics and state the type of data or analysis each addresses.
2. Explain what metagenomics is and what type of question it can answer.
3. Distinguish between bioinformatics and computational biology.
4. Explain what AlphaFold is and why it was considered revolutionary.
5. Name two career sectors for bioinformatics graduates in Nigeria and give one example role in each.

Series Summary

This Series traced the history and scope of bioinformatics from its origins in the 1960s to its present-day status as an indispensable discipline across science, medicine, agriculture, and industry. The field was founded on the twin imperatives of managing growing volumes of protein sequence data and developing computational tools to analyse them, exemplified by the pioneering contributions of Margaret Dayhoff, whose Atlas of Protein Sequence and Structure and PAM substitution matrices established enduring principles of database design and sequence comparison. The Human Genome Project, completed in 2003, created the global database infrastructure and algorithmic toolkit that defines modern bioinformatics.

The next-generation sequencing revolution has since shifted bioinformatics from data scarcity to data abundance, driving applications in clinical genomics, crop improvement, infectious disease surveillance, and environmental metagenomics. Africa and Nigeria are increasingly central to this global endeavour through institutions such as ACEGID and programmes such as H3Africa, addressing the critical underrepresentation of African genomic diversity in global databases. Bioinformatics encompasses a rich set of subdisciplines from structural and sequence analysis through phylogenomics and systems biology, and offers diverse career pathways across academic, healthcare, industrial, and governmental sectors that are directly relevant to Nigerian development priorities.

Glossary of Terms

- **Bioinformatics:** the interdisciplinary science that develops and applies computational methods to store, retrieve, analyse, and interpret biological data, particularly molecular biological data such as nucleotide and amino acid sequences.
- **BLAST (Basic Local Alignment Search Tool):** a heuristic algorithm for rapid similarity searching of large sequence databases, developed by Altschul and colleagues in 1990, and the most widely used bioinformatics tool.

- **GenBank:** the primary nucleotide sequence repository of the National Center for Biotechnology Information (NCBI), a founding member of the International Nucleotide Sequence Database Collaboration.
- **Human Genome Project (HGP):** the international collaborative effort initiated in 1990 to determine the complete sequence of the human genome, declared essentially complete in 2003.
- **Metagenomics:** the sequencing and analysis of total DNA extracted directly from environmental samples, revealing the diversity and function of microbial communities without prior cultivation.
- **Next-generation sequencing (NGS):** high-throughput DNA sequencing platforms introduced from 2005, capable of generating millions to billions of sequence reads per run at dramatically lower cost than Sanger sequencing.
- **PAM matrix:** a substitution scoring matrix developed by Margaret Dayhoff quantifying the relative rates at which amino acids replace each other over evolutionary time, foundational to sequence alignment scoring.
- **Protein Data Bank (PDB):** the worldwide repository for three-dimensional structures of proteins and nucleic acids determined by X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy.
- **Structural bioinformatics:** the subdiscipline of bioinformatics concerned with the prediction and analysis of three-dimensional structures of biological macromolecules from their sequences.
- **UniProt:** the universal protein resource, the primary annotated protein sequence database, formed by the merger of Swiss-Prot and TrEMBL in 2002.

Concept Check Questions [Series 1]

Objective Questions

1. The term bioinformatics was coined by Paulien Hogeweg and Ben Hesper in _____ . (Linked to SLO 1.1)
2. The Human Genome Project was initiated in _____ and declared essentially complete in 2003. (Linked to SLO 1.2)
3. _____ (Basic Local Alignment Search Tool) became the most widely used bioinformatics tool after its publication in 1990. (Linked to SLO 1.2)
4. Africa contains more human genetic _____ than any other continent, reflecting its status as the origin of anatomically modern humans. (Linked to SLO 1.3)
5. _____ bioinformatics predicts and analyses three-dimensional macromolecular structures from sequence data. (Linked to SLO 1.4)

Short-Answer Questions

6. Define bioinformatics and distinguish it from computational biology. (Linked to SLO 1.1)
7. Name three key contributions of Margaret Dayhoff to bioinformatics. (Linked to SLO 1.1)
8. Explain what BLAST is and why it became the most widely used bioinformatics tool. (Linked to SLO 1.2)
9. Explain how NGS changed the nature of bioinformatics challenges. (Linked to SLO 1.3)
10. Name four core subdisciplines of bioinformatics. (Linked to SLO 1.4)

Long-Answer Questions

11. Describe the early origins of bioinformatics and the contributions of Margaret Dayhoff. (Linked to SLO 1.1)
12. Describe the Human Genome Project and the key computational breakthroughs that shaped modern bioinformatics. (Linked to SLO 1.2)
13. Describe the applications of bioinformatics in the twenty-first century and explain its scope and career pathways. (Linked to SLOs 1.3, 1.4)

Answers to Concept Check Questions [Series 1]

Objective Questions

1. 1970.
2. 1990.
3. BLAST.
4. Diversity.
5. Structural.

Short-Answer Questions

6. Bioinformatics develops and applies computational methods to store, retrieve, analyse, and interpret biological data; computational biology more broadly focuses on mathematical modelling of biological processes such as population genetics and network models.
7. The Atlas of Protein Sequence and Structure (first protein sequence database), PAM substitution matrices (quantifying amino acid replacement rates), and the single-letter amino acid code (or free data-sharing principle).
8. BLAST is a heuristic algorithm for rapid similarity searching of large sequence databases with statistical scoring; it became most widely used because it balances speed and sensitivity, enabling database searches that would be computationally impractical with exhaustive dynamic programming methods.
9. NGS shifted bioinformatics from data scarcity (sequence data was the limiting factor) to data abundance (computational capacity and analytical expertise became the limiting factors), requiring new algorithms for assembly, variant calling, and expression analysis, and driving adoption of high-performance and cloud computing.
10. Sequence bioinformatics, structural bioinformatics, functional genomics, and phylogenomics (or metagenomics, proteomics bioinformatics, systems biology, clinical bioinformatics).

Long-Answer Questions

11. Bioinformatics emerged from the need to manage growing protein sequence data generated by Sanger sequencing (insulin, 1951). Margaret Dayhoff (1925-1983) created the Atlas of Protein Sequence and Structure (1965), the first protein sequence database, later the PIR. She developed PAM matrices (amino acid substitution rates over evolutionary time), the single-letter amino acid code, and promoted free data sharing. These established the conceptual foundations of database design and sequence analysis that persist to the present.

12. The HGP (1990-2003) sequenced the approximately 3.2 billion base pairs of the human genome internationally (USA, UK, France, Germany, Japan, China), creating GenBank, EMBL-Bank, and DDBJ as the INSDC. Key algorithms: Needleman-Wunsch (1970, global alignment, dynamic programming), Smith-Waterman (1981, local alignment), BLAST (1990, rapid heuristic database search with statistical scoring), ClustalW/MUSCLE (multiple sequence alignment), HMMs (profile searching). These tools made genomic data scientifically productive.
13. NGS (from 2005, Illumina dominant) reduced cost from billions to hundreds of dollars per genome, creating big data challenges driving cloud computing and machine learning. Applications: medicine (clinical genomics, pharmacogenomics, infectious disease, cancer), agriculture (marker-assisted selection, crop genomics), industry (drug discovery, biotechnology), environmental (metagenomics). Africa: genetic diversity underrepresented; H3Africa, African BioGenome Project, ACEGID (Nigeria) addressing this. Subdisciplines: sequence, structural, functional genomics, phylogenomics, metagenomics, proteomics, systems biology. Careers: academic, healthcare, industry, government, international organisations; Python and R key skills.

Answers to Pilot Questions [Series 1]

Part 1.1

1. Bioinformatics develops and applies computational methods to store, retrieve, analyse, and interpret biological data; coined by Paulien Hogeweg and Ben Hesper in 1970.
2. Bioinformatics focuses on computational analysis of biological sequence and structure data, especially database-oriented approaches; computational biology more broadly encompasses mathematical modelling of biological processes such as population genetics, developmental patterning, and network models.
3. The Atlas of Protein Sequence and Structure, created by Margaret Dayhoff in 1965; later known as the Protein Information Resource (PIR).
4. PAM (Percent Accepted Mutations) matrices quantify the relative rates at which amino acids replace each other over evolutionary time, derived from aligned protein sequences; they are important because they provide the scoring system for sequence alignment algorithms, allowing biologically meaningful comparisons.
5. Standardised data formats and comprehensive annotation; free availability of data to the entire scientific community.

Part 1.2

1. The HGP was initiated in 1990 and declared essentially complete in 2003.
2. GenBank (USA/NCBI), EMBL-Bank (European Molecular Biology Laboratory), and DDBJ (DNA Data Bank of Japan).
3. The Needleman-Wunsch algorithm (1970) is the first dynamic programming algorithm for global pairwise sequence alignment, making systematic comparison of two complete biological sequences possible.
4. BLAST is a heuristic algorithm for rapid similarity searching of large sequence databases with statistical scoring of results; it became most widely used because it is much faster than exact dynamic programming methods while maintaining acceptable sensitivity.
5. GenBank (nucleotide sequences), PDB/Protein Data Bank (three-dimensional structures), UniProt (annotated protein sequences).

Part 1.3

1. NGS platforms (introduced from 2005) generate sequence data at dramatically lower cost and higher throughput than Sanger sequencing; this shifted bioinformatics from data scarcity to data abundance, requiring new algorithms, high-performance computing, and machine learning to process and interpret the volume of data generated.
2. Clinical genomics (sequencing patient genomes to identify disease-causing variants) and infectious disease genomics (pathogen genome sequencing for outbreak tracking and resistance detection).
3. Bioinformatics supports crop improvement through comparative genomics (identifying gene functions), marker-assisted selection (using molecular markers linked to desirable traits to accelerate breeding programmes), and identifying genes responsible for drought tolerance or disease resistance.
4. Africa contains more human genetic variation than any other continent because anatomically modern humans originated there and have had more time to accumulate variation; underrepresentation in databases means many medically important variants found in African populations are not characterised, limiting the applicability of genomic medicine.
5. ACEGID (African Centre of Excellence for Genomics of Infectious Diseases, Redeemer University, Nigeria); areas of research include SARS-CoV-2 genomic surveillance and Lassa fever virus genomics.

Part 1.4

1. Sequence bioinformatics (nucleotide and amino acid sequence storage, alignment, and analysis), structural bioinformatics (three-dimensional macromolecular structure prediction and analysis), functional genomics (large-scale gene expression and regulation analysis), and metagenomics (environmental DNA sequencing and microbial community analysis).
2. Metagenomics sequences and analyses total DNA extracted directly from environmental samples without prior cultivation; it can answer questions about which microorganisms are present in a community, what metabolic functions they encode, and how the community composition changes with environmental conditions.
3. Bioinformatics focuses on computational storage and analysis of molecular biological data, particularly sequences and structures; computational biology more broadly encompasses mathematical modelling of biological processes, including population genetics models, developmental models, and gene regulatory network models.
4. AlphaFold is a deep learning system developed by DeepMind that predicts protein three-dimensional structure from amino acid sequence with accuracy comparable to experimental methods; it was considered revolutionary because protein structure prediction had been an unsolved problem for over 50 years, and AlphaFold solved it at scale, enabling structure prediction for essentially all known proteins.
5. Healthcare: clinical bioinformatician in a hospital genomics laboratory; government/public health: genomic epidemiologist at the NCDC (Nigerian Centre for Disease Control) tracking infectious disease outbreaks.

References and Attributions

Textual References (Books and External Sources)

- Lesk, A. M. (2019). Introduction to bioinformatics (4th ed.). Oxford University Press.
- Baxevanis, A. D., Bader, G. D., & Wishart, D. S. (Eds.). (2020). Bioinformatics (4th ed.). Wiley-Blackwell.
- Hogeweg, P. (2011). The roots of bioinformatics in theoretical biology. PLoS Computational Biology, 7(3), e1002021.
- National Open University of Nigeria (NOUN). (2023). Bioinformatics course material. NOUN Press.

Figures, Plates, Tables, and Boxes

- Figure 1.1: Origins of bioinformatics Attribution: AI generated. Suggested OER search string: "bioinformatics definition history Margaret Dayhoff PAM matrix Atlas protein sequence database origins OER".
- Figure 1.2: The emergence of modern bioinformatics Attribution: AI generated. Suggested OER search string: "Human Genome Project bioinformatics history BLAST Needleman Wunsch GenBank sequence alignment algorithms OER".
- Figure 1.3: Bioinformatics in the twenty-first century Attribution: AI generated. Suggested OER search string: "next generation sequencing bioinformatics applications Nigeria Africa ACEGID H3Africa genomic medicine agriculture OER".
- Figure 1.4: Scope and career pathways in bioinformatics Attribution: AI generated. Suggested OER search string: "bioinformatics subdisciplines career pathways structural sequence functional genomics proteomics metagenomics Nigeria OER".

Course code: BIO 413

Course title: Bioinformatics

Course credit load: 2 Units

Series 2: Bioinformatics Instrumentation

- **Part 2.1: Computing Hardware for Bioinformatics**
  - Sub Part 2.1.1: Personal computers and workstations
  - Sub Part 2.1.2: High-performance computing and computing clusters
  - Sub Part 2.1.3: Cloud computing in bioinformatics
- **Part 2.2: Operating Systems and Programming Environments**
  - Sub Part 2.2.1: Linux and the Unix command line
  - Sub Part 2.2.2: Python for bioinformatics
  - Sub Part 2.2.3: R and Bioconductor for biological data analysis
- **Part 2.3: Bioinformatics Software and Tools**
  - Sub Part 2.3.1: Sequence analysis tools
  - Sub Part 2.3.2: Genome browsers and visualisation tools
  - Sub Part 2.3.3: Workflow management and reproducibility tools
- **Part 2.4: Laboratory Instruments Generating Bioinformatics Data**
  - Sub Part 2.4.1: DNA sequencing instruments
  - Sub Part 2.4.2: Microarray and gene expression platforms
  - Sub Part 2.4.3: Mass spectrometry for proteomics

Introduction

Bioinformatics is a computationally intensive discipline, and the hardware, software, and laboratory instruments that generate and process biological data are as foundational to the field as its algorithms and databases. Understanding the instrumentation of bioinformatics, from the personal computer and its operating system through high-performance computing clusters and cloud services to the sequencing machines and mass spectrometers that produce the raw data, equips the student to make informed choices about the tools appropriate for a given analytical task.

This Series surveys the computing hardware used in bioinformatics from workstations to cloud computing; the operating systems and programming environments that form the practical working environment of the bioinformatician, particularly the Linux command line, Python, and R; the major categories of bioinformatics software and visualisation tools; and the key laboratory instruments whose outputs constitute the primary data of bioinformatics, including DNA sequencers, microarrays, and mass spectrometers for proteomics.

Series Learning Outcomes

Upon completing this Series, you should be able to:

- **SLO 2.1** describe the computing hardware used in bioinformatics, from personal computers to high-performance computing and cloud computing,
- **SLO 2.2** describe the Linux operating system, Python, and R as the primary computing environments of bioinformatics,
- **SLO 2.3** identify and describe the major categories of bioinformatics software, genome browsers, and workflow tools, and
- **SLO 2.4** describe the DNA sequencing instruments, microarray platforms, and mass spectrometers that generate primary bioinformatics data.

2.1 Computing Hardware for Bioinformatics

2.1.1 Personal computers and workstations

A personal computer (PC) running a modern operating system is sufficient for many routine bioinformatics tasks, including web-based database searching, sequence alignment using BLAST against remote servers, basic R and Python scripting for data analysis, and training in bioinformatics concepts; the key hardware specifications relevant to bioinformatics on a personal computer are: processor speed and core count (more cores allow parallel execution of independent analysis tasks), random-access memory (RAM, with 16 to 32 GB recommended for moderate bioinformatics work, as many tools load entire datasets into memory), and storage capacity (bioinformatics datasets, particularly raw sequencing data, can be hundreds of gigabytes to terabytes in size, requiring substantial local or networked storage).

A workstation is a more powerful, single-user computer designed for computationally intensive tasks; bioinformatics workstations typically have 32 to 128 GB or more of RAM, multi-core processors (16 to 64 cores), large fast storage arrays using solid-state drives (SSDs), and may run Linux rather than a consumer operating system to provide better performance and access to bioinformatics software ecosystems; workstations are appropriate for moderate-scale analyses such as RNA-seq analysis of a small number of samples, whole-exome sequencing analysis, or development and testing of bioinformatics pipelines before deployment on a computing cluster.

Fill in the blank: The key hardware specifications relevant to bioinformatics on a personal computer include processor core count, _____ -access memory (RAM), and storage capacity.

2.1.2 High-performance computing and computing clusters

High-performance computing (HPC) refers to the use of multiple interconnected computers (nodes) working together as a computing cluster to perform computationally intensive analyses that exceed the capacity of any single machine; a typical HPC cluster consists of a login node (through which users connect and submit jobs), compute nodes (high-core-count, high-memory servers that perform the actual computation), a high-speed interconnect (a fast network linking the nodes, essential for parallel computations that require communication between nodes), and a

shared file system (providing all nodes with access to a common storage area where data and results are stored).

Jobs on an HPC cluster are submitted and managed through a workload manager or job scheduler (such as SLURM, PBS, or LSF), which queues submitted jobs and allocates them to available compute resources according to defined priority and resource request rules; bioinformatics analyses well-suited to HPC include whole-genome sequencing analysis of large cohorts, large-scale sequence database searches, phylogenomic analyses of thousands of genomes, and training of machine learning models on genomic data; many Nigerian universities and research institutions have access to HPC resources through national facilities or international collaborations such as the African Research Cloud.

Fill in the blank: A job _____ such as SLURM or PBS manages the queue of submitted jobs on an HPC cluster, allocating them to available compute nodes according to resource requests and priority.

2.1.3 Cloud computing in bioinformatics

Cloud computing provides on-demand access to computing resources (virtual machines, storage, and specialised computing services such as GPU instances for machine learning) through the internet, without requiring the user to own or maintain physical hardware; the major commercial cloud platforms used in bioinformatics include Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure, all of which provide bioinformatics-specific services including managed genomics analysis pipelines, scalable storage for large datasets, and pre-configured virtual machine images with bioinformatics software installed.

The advantages of cloud computing for bioinformatics include scalability (resources can be scaled up or down on demand, paying only for what is used), accessibility (analyses can be run from anywhere with internet access, without the need for local HPC infrastructure), and access to cutting-edge hardware including GPU clusters for deep learning; disadvantages include cost (large-scale cloud analyses can be expensive if not carefully managed), data transfer costs and latency (moving large bioinformatics datasets to and from cloud storage is slow and may have

significant cost), and concerns about data privacy and governance for sensitive clinical genomics data.

Fill in the blank: The major commercial cloud platforms used in bioinformatics include Amazon Web Services, Google Cloud _____ , and Microsoft Azure.

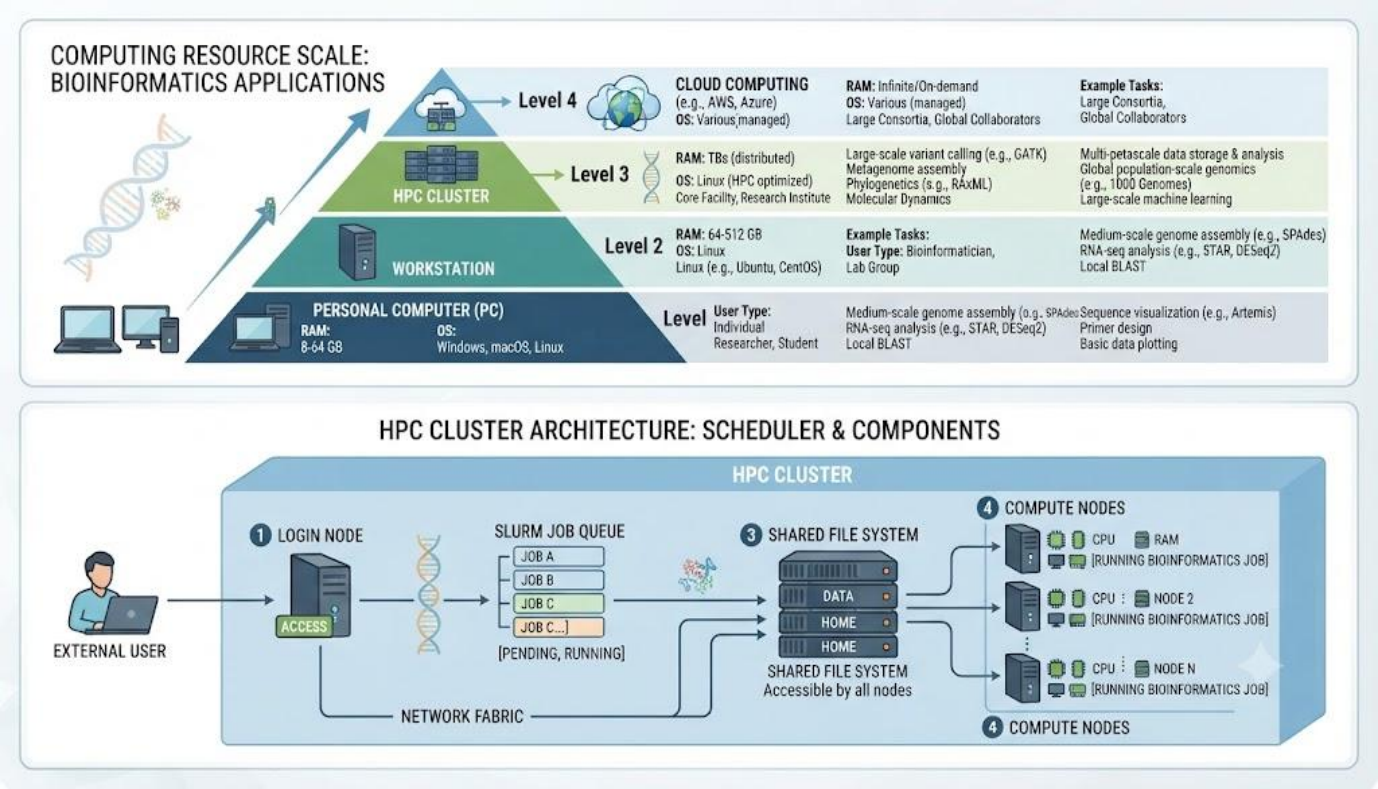

Figure 2.1: Computing hardware for bioinformatics

Pilot questions 2.1

1. Name three hardware specifications relevant to bioinformatics performance on a personal computer.
2. Explain what a computing cluster is and name its main components.
3. What is a job scheduler and why is it needed on an HPC cluster?
4. Name three commercial cloud platforms used in bioinformatics.
5. State two advantages and two disadvantages of cloud computing for bioinformatics.

2.2 Operating Systems and Programming Environments

2.2.1 Linux and the Unix command line

Linux is the operating system of choice for bioinformatics, used on the majority of HPC clusters, most cloud computing instances, and a large proportion of bioinformatics workstations, because the vast majority of bioinformatics software is developed and tested on Linux, many tools are available only for Linux, and the Linux command-line environment provides powerful text processing and automation capabilities essential for bioinformatics workflows; Linux is a free, open-source operating system based on the Unix architecture, available in many distributions (distros) including Ubuntu, CentOS, and Debian, all of which share the same core command-line interface.

The Unix command line (or shell, typically Bash in Linux) provides a text-based interface for interacting with the computer, navigating the file system, executing programs, and writing scripts that automate sequences of commands; essential Unix commands for bioinformatics include: ls (list files in a directory), cd (change directory), pwd (print working directory), mkdir (create a new directory), cp and mv (copy and move files), rm (delete files), grep (search for text patterns in files), awk and sed (text processing tools for reformatting data files), and pipe (|) and redirect (>) operators for chaining commands and directing output; the ability to write Bash shell scripts that chain these commands into automated workflows is a foundational bioinformatics skill.

Fill in the blank: The Unix shell command _____ searches for text patterns within files, a tool essential for processing bioinformatics data files at the command line.

2.2.2 Python for bioinformatics

Python is the most widely used programming language in bioinformatics, valued for its readable syntax, large standard library, extensive ecosystem of scientific computing packages, and the availability of specialised bioinformatics libraries; the key Python libraries for bioinformatics include: Biopython (providing modules for reading and writing common bioinformatics file formats such as FASTA, FASTQ, GenBank, and PDB, interfacing with NCBI Entrez databases, and performing sequence manipulation and alignment); NumPy and pandas (for numerical array

operations and tabular data management, respectively); matplotlib and seaborn (for data visualisation); and scikit-learn (for machine learning applications in genomics and proteomics).

Python is also extensively used for writing bioinformatics pipeline scripts that chain multiple tools in sequence (for example, a typical RNA-seq analysis pipeline written in Python or Bash might run quality control with FastQC, adapter trimming with Trimmomatic, alignment with STAR or HISAT2, and differential expression analysis with DESeq2, with each step reading the output of the previous step); the Jupyter Notebook environment (which allows mixing of executable code, output, plots, and narrative text in a single document) has become widely used for bioinformatics data analysis documentation and teaching.

Fill in the blank: _____ is the primary Python library for bioinformatics, providing modules for reading and writing FASTA, GenBank, and other biological file formats and interfacing with NCBI databases.

2.2.3 R and Bioconductor for biological data analysis

R is a statistical programming language and environment widely used in bioinformatics for statistical data analysis, particularly for the analysis of high-throughput genomic data; its strengths include a vast ecosystem of statistical packages, excellent data visualisation capabilities (particularly through the ggplot2 package, which implements the grammar of graphics and produces publication-quality figures), and the Bioconductor project, an open-source repository of R packages specifically designed for the analysis of genomic data.

Bioconductor (bioconductor.org) hosts over 2,000 R packages for bioinformatics, covering RNA-seq differential expression analysis (DESeq2, edgeR, limma), microarray data normalisation and analysis (affy, limma), ChIP-seq analysis (ChIPseeker, DiffBind), variant annotation and population genomics (VariantAnnotation, SNPRelate), single-cell RNA-seq analysis (Seurat, SingleCellExperiment), and many other applications; the consistent object-oriented data structures used across Bioconductor packages (such as SummarizedExperiment, GenomicRanges, and Biostrings) allow complex multi-step analyses to be assembled efficiently from modular components.

Fill in the blank: _____ is the open-source repository of R packages specifically designed for genomic data analysis, hosting over 2,000 packages covering RNA-seq, ChIP-seq, variant analysis, and single-cell genomics.

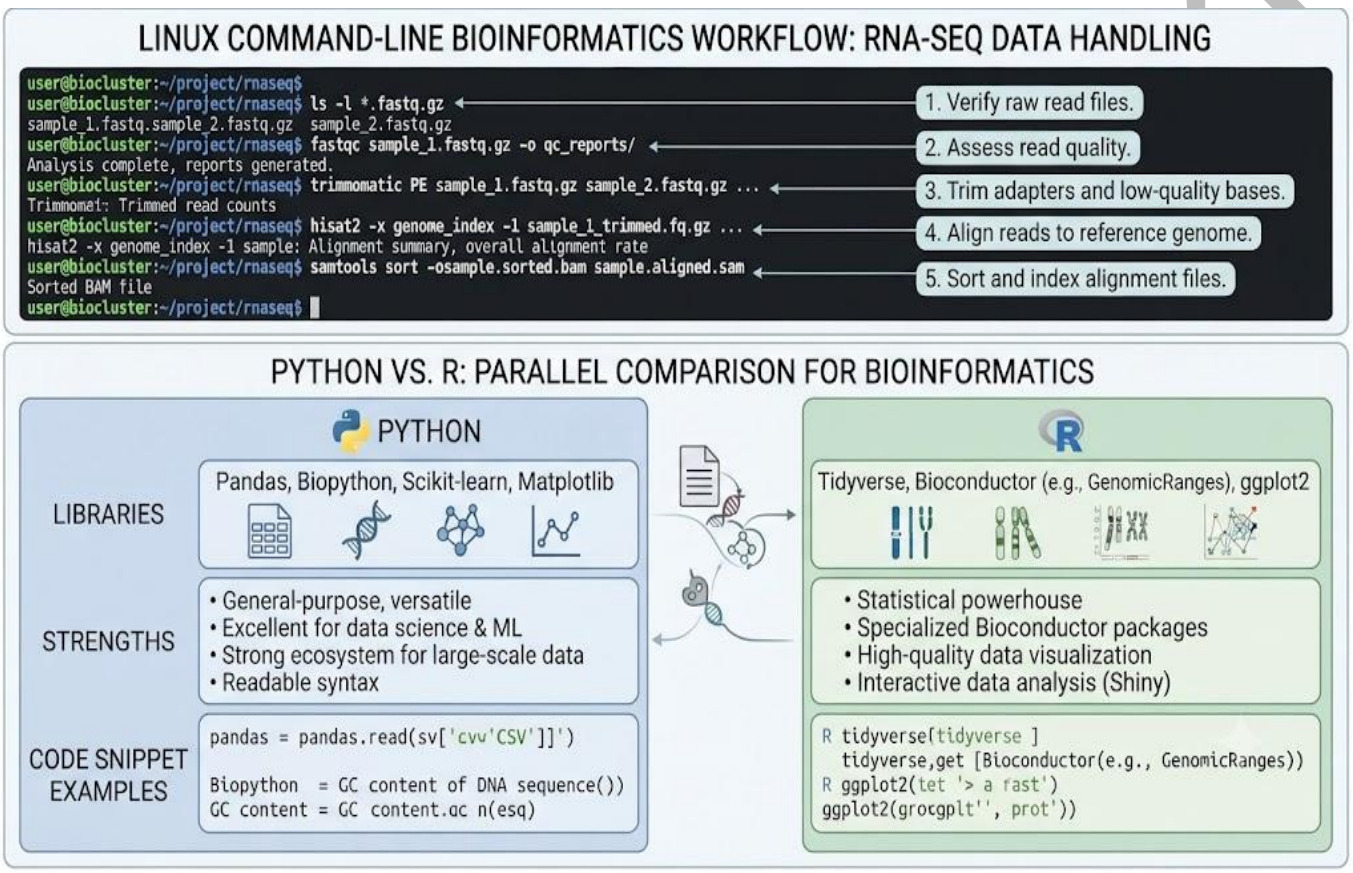

Figure 2.2: Programming environments for bioinformatics

Pilot questions 2.2

1. Explain why Linux is the preferred operating system for bioinformatics.
2. Name five essential Unix command-line commands for bioinformatics and state what each does.
3. Name four Python libraries used in bioinformatics and state the purpose of each.
4. Explain what Bioconductor is and name two R packages it contains.
5. Distinguish between Python and R in terms of their primary strengths in bioinformatics.

2.3 Bioinformatics Software and Tools

2.3.1 Sequence analysis tools

Sequence analysis tools are the most widely used category of bioinformatics software, addressing tasks from quality control of raw sequencing data through alignment, assembly, and annotation to database searching; key tools include: FastQC (quality control assessment of raw FASTQ sequencing reads, producing reports of per-base quality scores, adapter content, and sequence duplication levels); Trimmomatic and Cutadapt (adapter trimming and quality filtering of raw reads before alignment); STAR and HISAT2 (splice-aware aligners for mapping RNA-seq reads to reference genomes, essential for transcriptomics); BWA-MEM and Bowtie2 (aligners for DNA sequencing reads, used in whole-genome and whole-exome sequencing analysis).

Additional widely used sequence analysis tools include: SAMtools (manipulation and indexing of BAM and SAM format alignment files, the standard intermediate file format in sequencing analysis pipelines); GATK (Genome Analysis Toolkit, the standard tool for variant calling from whole-genome and whole-exome sequencing data, developed by the Broad Institute); SPAdes and MEGAHIT (de novo genome assemblers for bacterial and metagenomic data respectively); Prokka (rapid annotation of bacterial genome assemblies); and Trinity (de novo assembly of transcriptomes from RNA-seq data without a reference genome); the appropriate choice of tools depends on the organism, data type, and analytical goal.

Fill in the blank: _____ is the quality control tool for raw sequencing reads, producing reports of per-base quality scores, adapter content, and sequence duplication levels for FASTQ files.

2.3.2 Genome browsers and visualisation tools

Genome browsers are software applications or web services that allow users to visualise biological sequence data and genome annotations in an integrated graphical interface organised along the linear coordinates of a reference genome; the most widely used genome browsers are the UCSC Genome Browser (University of California Santa Cruz, providing web-based access to

hundreds of genomes with extensive pre-computed annotation tracks) and Ensembl (developed by the European Bioinformatics Institute and Wellcome Sanger Institute, also web-based with extensive annotation and comparative genomics data for vertebrates and many other organisms).

For visualisation of local data (such as the output of a student own sequencing analysis), the Integrative Genomics Viewer (IGV) is the most widely used desktop application, allowing users to load BAM alignment files, VCF variant files, BED annotation files, and many other formats and view them alongside a reference genome with zoom and pan navigation; other important visualisation tools include Cytoscape (for network visualisation of protein-protein interaction and gene regulatory networks), R packages such as ggplot2, GenomicRanges, and Gviz (for genomics data plotting in R), and the CIRCOS tool (for circular genome visualisation and synteny comparison).

Fill in the blank: The Integrative Genomics Viewer (_____) is a desktop application for visualising BAM alignment files, VCF variant files, and other genomics data formats alongside a reference genome.

2.3.3 Workflow management and reproducibility tools

Workflow management systems (WMS) allow complex multi-step bioinformatics analyses to be defined, automated, and run reproducibly across different computing environments; the most widely used WMS in bioinformatics are Snakemake (a Python-based WMS that defines workflows as a set of rules specifying input files, output files, and the commands to generate outputs from inputs, automatically determining the order of rule execution and enabling parallel execution of independent steps) and Nextflow (a Groovy/DSL-based WMS with built-in support for containerised execution using Docker or Singularity, enabling workflows to run identically on local machines, HPC clusters, or cloud computing platforms).

Reproducibility tools are essential in bioinformatics because analyses involving many software packages and their specific versions must be precisely documented to be reproducible: conda (and its bioinformatics-specific channel Bioconda) is a package manager that installs and manages exact versions of bioinformatics software and their dependencies in isolated environments; Docker and Singularity are container systems that package a complete software

environment (operating system, libraries, and tools) into a portable image that runs identically on any compatible host; version control systems such as Git and the GitHub platform allow analysis code to be tracked, shared, and reproduced, and are considered essential practice in modern bioinformatics.

Fill in the blank: _____ is a package manager with the Bioconda channel that installs and manages exact versions of bioinformatics software in isolated environments, supporting reproducible analysis.

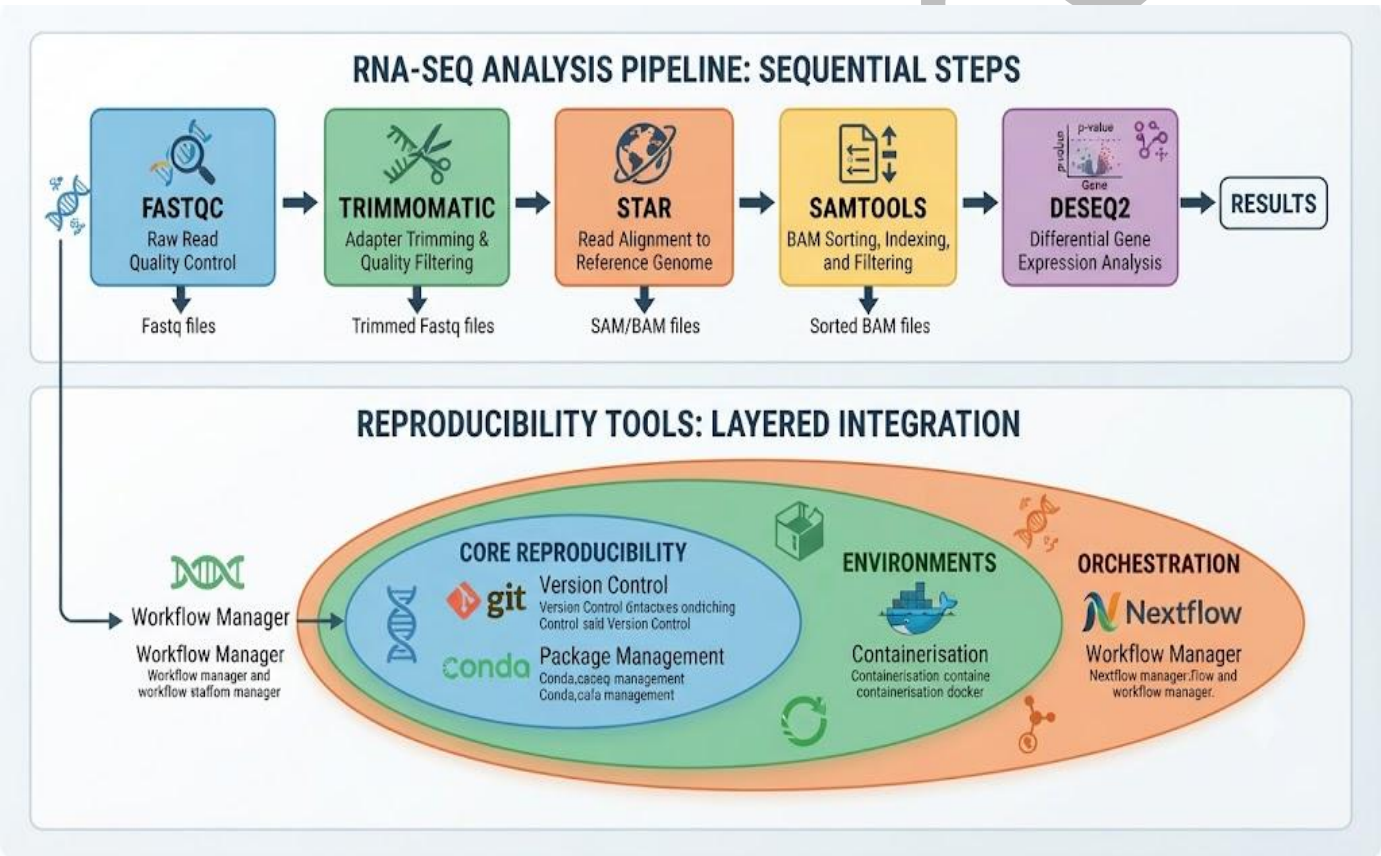

Figure 2.3: Bioinformatics software and workflow tools

Pilot questions 2.3

1. Name the tool used for quality control of raw FASTQ sequencing reads and state what it reports.
2. Name two splice-aware aligners used in RNA-seq analysis.
3. Distinguish between the UCSC Genome Browser and IGV.
4. Explain what a workflow management system is and name two examples.

6. Explain why Docker or Singularity containers are important for bioinformatics reproducibility.

2.4 Laboratory Instruments Generating Bioinformatics Data

2.4.1 DNA sequencing instruments

DNA sequencing instruments generate the nucleotide sequence data that is the primary input to the majority of bioinformatics analyses; the major sequencing platforms and their outputs are: Illumina sequencing instruments (the dominant short-read NGS platform, generating paired-end reads of 150 to 300 bp with very high accuracy, in quantities from millions to billions of reads per run depending on the instrument model, covering Illumina MiSeq for small-scale applications, NextSeq for moderate throughput, and NovaSeq for high-throughput production sequencing); Oxford Nanopore Technologies (ONT) sequencers (generating long reads of tens of thousands of base pairs by measuring the electrical current disruption as single DNA molecules pass through protein nanopores, enabling direct sequencing of DNA methylation and resolving repetitive regions that short reads cannot).

PacBio (Pacific Biosciences) SMRT sequencing instruments (generating long, highly accurate reads using single-molecule real-time sequencing chemistry, with the HiFi/CCS mode producing reads averaging 10 to 25 kb with Illumina-comparable accuracy, ideal for high-quality genome assembly); and Sanger capillary sequencers (still widely used for sequencing of individual PCR products, cloned inserts, or small amplicons where the high throughput of NGS is not needed and the lower cost per sample of capillary sequencing remains advantageous for targeted applications); the raw data output format of all NGS platforms is the FASTQ file, containing the sequence and per-base quality scores for each read.

Fill in the blank: The FASTQ file format contains the nucleotide _____ and per-base quality scores for each sequencing read, and is the standard raw data output of all next-generation sequencing platforms.

2.4.2 Microarray and gene expression platforms

Microarrays are glass slides or chips containing thousands to millions of short oligonucleotide probes immobilised at defined positions; when a labelled nucleic acid sample (such as cDNA prepared from mRNA, or genomic DNA) is hybridised to the array, sequences in the sample complementary to the probes bind and are detected by fluorescence scanning; gene expression microarrays (such as Affymetrix GeneChip arrays, used extensively from the late 1990s through the 2010s) measure the relative abundance of thousands of mRNA transcripts simultaneously by hybridising cDNA samples from two conditions (labelled with different fluorescent dyes) to the same array.

SNP genotyping arrays (such as Illumina BeadChip arrays) hybridise genomic DNA to probes designed at known single nucleotide polymorphism (SNP) positions across the genome, generating genotype calls at hundreds of thousands to millions of SNP positions in a single experiment at lower cost than whole-genome sequencing; these arrays are widely used in genome-wide association studies (GWAS) to identify genetic variants associated with disease or trait variation in large human populations; DNA methylation arrays (such as Illumina EPIC array) measure the methylation status at hundreds of thousands of CpG sites across the human genome, used in epigenomics research and clinical diagnostics for cancer and imprinting disorders.

Fill in the blank: SNP genotyping _____ hybridise genomic DNA to probes at known variant positions, generating genotype calls at up to millions of SNP positions in a single experiment, used widely in genome-wide association studies.

2.4.3 Mass spectrometry for proteomics

Mass spectrometry (MS) is the primary instrument for large-scale protein identification and quantification in proteomics; the general principle involves three steps: ionisation of the sample (converting proteins or peptides into gas-phase ions, most commonly by electrospray ionisation, ESI, or matrix-assisted laser desorption/ionisation, MALDI), mass analysis (separating ions by their mass-to-charge ratio, m/z, in a mass analyser such as a quadrupole, time-of-flight, or

orbitrap instrument), and detection (measuring the abundance of ions at each m/z value to generate a mass spectrum).

In shotgun proteomics (the most widely used approach for global protein identification), a complex protein mixture is digested with trypsin to generate a mixture of peptides, which are separated by liquid chromatography (LC), ionised by ESI, and analysed by tandem mass spectrometry (MS/MS or LC-MS/MS); the MS/MS spectrum of each peptide (showing the masses of fragment ions generated by fragmentation of the peptide backbone) is matched computationally against a protein sequence database using search engines such as Mascot, Sequest, or MaxQuant to identify the originating protein; the output of a proteomics experiment is a list of identified proteins with quantitative abundance information, which is then analysed using bioinformatics tools to address biological questions.

Fill in the blank: In shotgun proteomics, proteins are digested with _____ to generate peptides, which are separated by liquid chromatography and analysed by tandem mass spectrometry for protein identification.

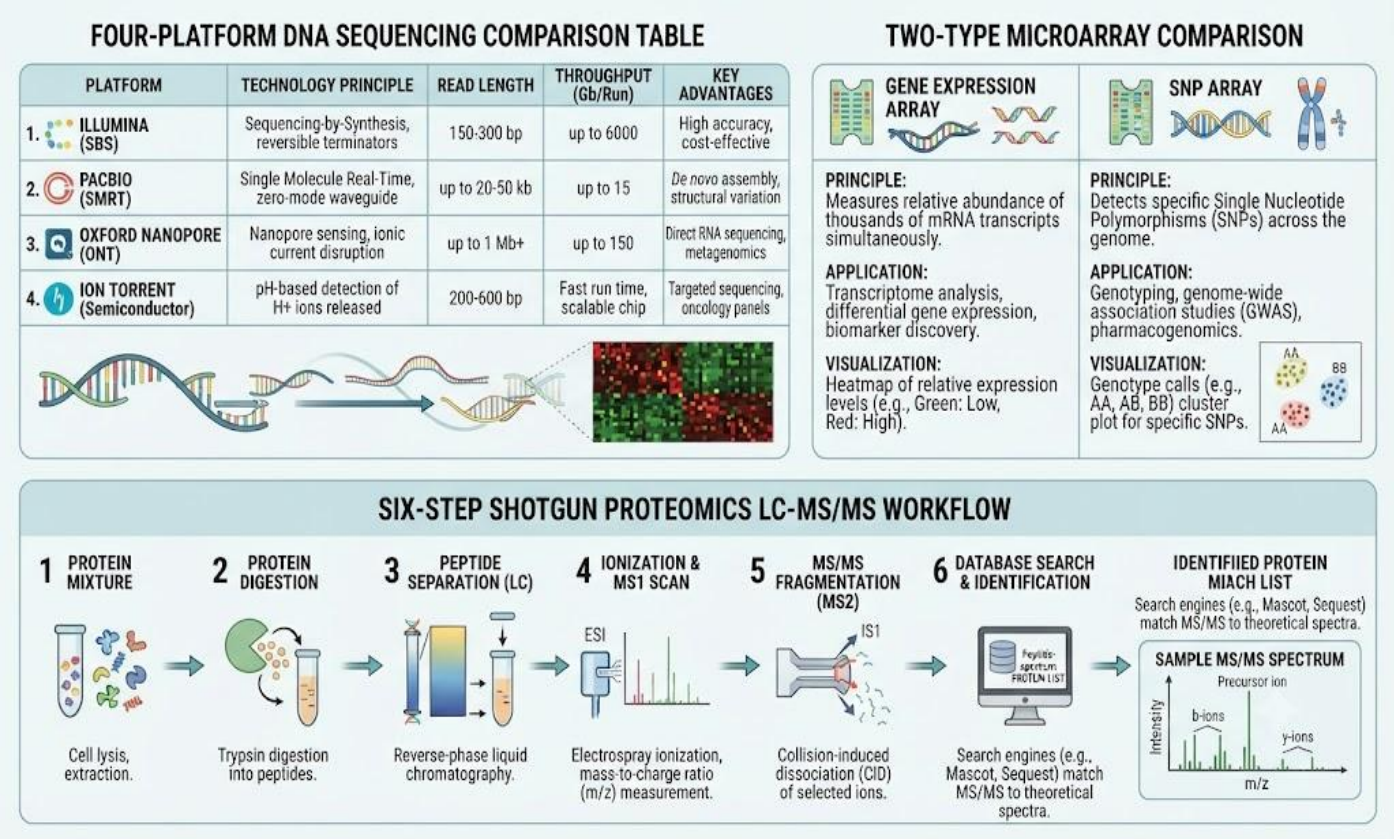

Figure 2.4: Laboratory instruments generating bioinformatics data

Pilot questions 2.4

7. Name four DNA sequencing platforms and state the read length characteristic of each.
8. Explain what the FASTQ file format contains.
9. Explain the principle of a gene expression microarray.
10. Explain what SNP genotyping arrays are used for.
11. Describe the basic steps of shotgun proteomics by mass spectrometry.

Series Summary

This Series surveyed the instrumentation that makes bioinformatics possible. Computing hardware ranges from personal computers suitable for routine tasks through workstations for moderate analyses and HPC clusters for large-scale parallel computation, to cloud platforms that provide scalable on-demand resources globally. Linux is the operating system underpinning essentially all bioinformatics computing infrastructure, and the Unix command line combined with Python and R programming form the practical working environment in which most bioinformatics analysis is conducted.

The bioinformatics software ecosystem spans quality control and alignment tools for sequencing data, genome browsers and visualisation applications for exploring results, and workflow management and reproducibility tools that ensure analyses are automated and can be re-run identically by others. The primary laboratory instruments generating bioinformatics data are DNA sequencers (producing FASTQ reads across Illumina, Oxford Nanopore, PacBio, and Sanger platforms), microarrays (measuring gene expression and SNP genotypes in parallel across thousands of loci), and mass spectrometers (identifying and quantifying proteins in complex mixtures through the LC-MS/MS shotgun proteomics workflow). Together these instruments and tools form the integrated infrastructure of modern bioinformatics.

Glossary of Terms

- **BAM/SAM format:** the standard file format for storing aligned sequencing reads; SAM is human-readable text while BAM is its compressed binary equivalent, both managed by SAMtools.
- **Bioconductor:** an open-source project providing over 2,000 R packages specifically for genomic data analysis, covering RNA-seq, ChIP-seq, variant analysis, and single-cell genomics.
- **Conda:** a cross-platform package manager that installs and manages exact versions of bioinformatics software and their dependencies in isolated environments.

- **FASTQ format:** the standard file format for raw next-generation sequencing data, storing nucleotide sequences and per-base quality scores for each sequencing read.
- **High-performance computing (HPC):** the use of interconnected computer clusters to perform computationally intensive analyses beyond the capacity of any single machine.
- **IGV (Integrative Genomics Viewer):** a desktop application for visualising genomics data including BAM alignments, VCF variants, and annotation tracks alongside a reference genome.
- **Mass spectrometry (MS):** an analytical technique that identifies and quantifies molecules by separating ions according to their mass-to-charge ratio.
- **Microarray:** a glass slide or chip bearing thousands of oligonucleotide probes used to measure gene expression levels or genotype SNPs by hybridisation.
- **Shotgun proteomics:** a mass spectrometry approach for global protein identification in which complex mixtures are digested to peptides, separated by LC, and analysed by MS/MS.
- **Workflow management system (WMS):** software that defines, automates, and runs multi-step bioinformatics analyses reproducibly across different computing environments.

Concept Check Questions [Series 2]

Objective Questions

1. The three key hardware specifications for bioinformatics on a personal computer are processor core count, _____ -access memory, and storage capacity. (Linked to SLO 2.1)
2. A job _____ such as SLURM manages job queues on HPC clusters, allocating jobs to available compute nodes. (Linked to SLO 2.1)
3. The Unix command _____ searches for text patterns within files, essential for processing bioinformatics data. (Linked to SLO 2.2)
4. _____ is the primary Python library for bioinformatics, supporting FASTA/GenBank file reading and NCBI database access. (Linked to SLO 2.2)
5. The FASTQ file format stores nucleotide _____ and per-base quality scores for each sequencing read. (Linked to SLO 2.4)

Short-Answer Questions

1. Explain two advantages and two disadvantages of cloud computing for bioinformatics. (Linked to SLO 2.1)
2. Name four Python libraries used in bioinformatics and state the purpose of each. (Linked to SLO 2.2)
3. Explain what Bioconductor is and name two R packages it contains. (Linked to SLO 2.2)
4. Name four sequence analysis tools and state what each does. (Linked to SLO 2.3)
5. Describe the basic steps of shotgun proteomics. (Linked to SLO 2.4)

Long-Answer Questions

1. Describe the computing hardware and operating system environments used in bioinformatics. (Linked to SLOs 2.1, 2.2)
2. Describe the major categories of bioinformatics software tools and the laboratory instruments that generate primary bioinformatics data. (Linked to SLOs 2.3, 2.4)

Answers to Concept Check Questions [Series 2]

Objective Questions

1. Random.
2. Scheduler.
3. grep.
4. Biopython.
5. Sequence.

Short-Answer Questions

1. Advantages: scalability (resources scaled on demand, pay only for use) and accessibility (analyses run from anywhere without local HPC). Disadvantages: cost (large-scale analyses can be expensive) and data privacy/transfer concerns for sensitive clinical data.
2. Biopython (read/write biological file formats, NCBI access), NumPy (numerical array operations), pandas (tabular data management), matplotlib (data visualisation), scikit-learn (machine learning).
3. Bioconductor is an open-source repository of R packages for genomic data analysis; examples: DESeq2 (RNA-seq differential expression) and edgeR (differential expression) or limma.
4. FastQC (quality control of raw FASTQ reads), STAR or HISAT2 (splice-aware RNA-seq alignment), BWA-MEM (DNA read alignment), GATK (variant calling from WGS/WES).
5. Protein mixture is digested with trypsin to generate peptides; peptides are separated by liquid chromatography (LC); ionised by ESI; analysed by tandem MS/MS; spectra are searched against a protein database with Mascot or MaxQuant to identify proteins.

Long-Answer Questions

1. Computing hardware: personal computers (16-32 GB RAM, routine tasks), workstations (32-128 GB, moderate analyses), HPC clusters (interconnected nodes with job scheduler, large-scale analyses), cloud (AWS/GCP/Azure, scalable, on-demand). Linux is the OS of choice on HPC and cloud; Unix command line (grep, awk, pipes) enables text processing and automation. Python (Biopython, NumPy, pandas, matplotlib, scikit-learn) for general programming and pipelines; R and Bioconductor (DESeq2, edgeR, ggplot2) for statistical genomic analysis.

2. Software tools: FastQC (QC), Trimmomatic (adapter trimming), STAR/HISAT2 (RNA-seq alignment), BWA-MEM (DNA alignment), GATK (variant calling), SPAdes (assembly). Genome browsers: UCSC and Ensembl (web, pre-computed tracks), IGV (local data desktop viewer). Reproducibility: Snakemake/Nextflow (workflow managers), conda/Bioconda (package management), Docker/Singularity (containers). Instruments: Illumina (short reads, high throughput), ONT (long reads, real-time), PacBio (long HiFi reads), Sanger (targeted); microarrays (gene expression, SNP genotyping); mass spectrometry/LC-MS/MS (shotgun proteomics).

Answers to Pilot Questions [Series 2]

Part 2.1

1. Processor core count (multiple cores allow parallel tasks), random-access memory (RAM, many tools load datasets into memory), and storage capacity (sequencing datasets can be hundreds of GB to terabytes).
2. A computing cluster is multiple interconnected computers working together; main components: login node (user entry), compute nodes (perform computation), high-speed interconnect (fast network between nodes), shared file system (common storage), and job scheduler.
3. A job scheduler manages the queue of submitted jobs, allocating them to available compute nodes according to resource requests and priority; it is needed because many users share the cluster resources and jobs must be ordered and allocated fairly and efficiently.
4. Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure.
5. Advantages: scalability (scale up/down on demand, pay for use) and accessibility (run analyses from anywhere). Disadvantages: cost (large analyses can be expensive) and data privacy concerns for sensitive clinical genomics data (or data transfer costs and latency).

Part 2.2

1. Linux is the OS of choice because the vast majority of bioinformatics software is developed and tested on Linux, many tools are available only for Linux, and the command-line environment provides powerful text processing and automation capabilities.

2. ls (list files), cd (change directory), grep (search for text patterns), awk (text processing and reformatting), and pipe | (chain command output to input of next command).
3. Biopython (read/write biological file formats, NCBI access), NumPy (numerical array operations), pandas (tabular data management), matplotlib (data visualisation).
4. Bioconductor is an open-source repository of over 2,000 R packages for genomic data analysis; examples: DESeq2 (RNA-seq differential expression) and edgeR (differential expression).
5. Python is stronger for general programming, pipeline scripting, and machine learning; R is stronger for statistical analysis of genomic data and publication-quality visualisation through ggplot2 and Bioconductor packages.

Part 2.3

1. FastQC; it reports per-base quality scores, adapter content, sequence duplication levels, and other quality metrics for raw FASTQ sequencing reads.
2. STAR and HISAT2.
3. The UCSC Genome Browser is a web-based service with hundreds of pre-computed genome annotation tracks; IGV (Integrative Genomics Viewer) is a desktop application for visualising the user own local data files such as BAM alignments and VCF variants alongside a reference genome.
4. A workflow management system defines, automates, and runs multi-step bioinformatics analyses reproducibly across computing environments; examples: Snakemake and Nextflow.
5. Docker and Singularity containers package the complete software environment (OS, libraries, and tools) into a portable image that runs identically on any compatible host, ensuring that analyses produce identical results regardless of the underlying system.

Part 2.4

1. Illumina (150-300 bp short reads), Oxford Nanopore (tens of thousands of bp long reads), PacBio SMRT/HiFi (10-25 kb long reads), Sanger capillary (400-1000 bp).
2. FASTQ format contains the nucleotide sequence and per-base quality score (encoded as ASCII characters) for each sequencing read, along with a unique read identifier.
3. A gene expression microarray contains thousands of oligonucleotide probes on a glass slide; labelled cDNA prepared from mRNA is hybridised to the probes; the fluorescence intensity at each probe position indicates the relative abundance of the corresponding mRNA in the sample.

4. SNP genotyping arrays hybridise genomic DNA to probes at known SNP positions across the genome, generating genotype calls at hundreds of thousands to millions of positions; they are used in genome-wide association studies (GWAS) to identify genetic variants associated with disease or traits in large populations.
5. Protein mixture is digested with trypsin to peptides; peptides are separated by liquid chromatography (LC); ionised by electrospray ionisation (ESI); analysed by tandem mass spectrometry (MS/MS); mass spectra are searched against a protein sequence database with search engines such as Mascot or MaxQuant to identify the proteins.

References and Attributions

Textual References (Books and External Sources)

- Lesk, A. M. (2019). Introduction to bioinformatics (4th ed.). Oxford University Press.
- Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., ... & de Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422-1423.
- Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., ... & Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods, 12(2), 115-121.
- National Open University of Nigeria (NOUN). (2023). Bioinformatics course material. NOUN Press.

Figures, Plates, Tables, and Boxes

- Figure 2.1: Computing hardware for bioinformatics Attribution: AI generated. Suggested OER search string: "bioinformatics computing hardware HPC cluster cloud computing workstation Linux job scheduler OER".
- Figure 2.2: Programming environments for bioinformatics Attribution: AI generated. Suggested OER search string: "Linux command line Unix bioinformatics Python Biopython R Bioconductor DESeq2 programming environments OER".
- Figure 2.3: Bioinformatics software and workflow tools Attribution: AI generated. Suggested OER search string: "bioinformatics pipeline RNA-seq FastQC STAR DESeq2 Snakemake conda Docker reproducibility workflow OER".
- Figure 2.4: Laboratory instruments generating bioinformatics data Attribution: AI generated. Suggested OER search string: "DNA sequencing Illumina Oxford Nanopore PacBio mass spectrometry proteomics microarray SNP array bioinformatics OER".

Course code: BIO 413

Course title: Bioinformatics

Course credit load: 2 Units

Series 3: DNA and Protein Databases

- **Part 3.1: Principles of Biological Database Design**
    - Sub Part 3.1.1: Types of biological databases
    - Sub Part 3.1.2: Database architecture and annotation
    - Sub Part 3.1.3: Data standards and interoperability
- **Part 3.2: Nucleotide Sequence Databases**
    - Sub Part 3.2.1: GenBank, EMBL-Bank, and DDBJ
    - Sub Part 3.2.2: Genome databases and reference sequences
    - Sub Part 3.2.3: Specialised nucleotide databases
- **Part 3.3: Protein Sequence and Structure Databases**
    - Sub Part 3.3.1: UniProt: Swiss-Prot and TrEMBL
    - Sub Part 3.3.2: The Protein Data Bank
    - Sub Part 3.3.3: Protein family and domain databases
- **Part 3.4: Database Searching and Access**
    - Sub Part 3.4.1: BLAST and sequence similarity searching
    - Sub Part 3.4.2: Entrez and programmatic database access
    - Sub Part 3.4.3: Interpreting database search results

Introduction

Biological databases are the foundational infrastructure of bioinformatics, storing the accumulated molecular biological knowledge of the scientific community in structured, searchable, and interconnected repositories. The quality, comprehensiveness, and accessibility of these databases determine what questions can be asked and how reliably they can be answered; a bioinformatician who does not understand what the major databases contain, how they are organised, and how to search them effectively is as limited as a scientist without access to the scientific literature.

This Series covers the principles of biological database design and annotation; the major nucleotide sequence databases including the International Nucleotide Sequence Database Collaboration members and specialised genome databases; the primary protein sequence databases UniProt and its components Swiss-Prot and TrEMBL, the Protein Data Bank, and protein family databases; and the principal methods for searching biological databases, including BLAST, the Entrez system, and programmatic access, with guidance on interpreting the results of database searches.

Series Learning Outcomes

Upon completing this Series, you should be able to:

- **SLO 3.1** describe the types, architecture, and annotation principles of biological databases and explain data standards and interoperability,
- **SLO 3.2** describe the major nucleotide sequence databases and genome databases and explain the types of data each contains,
- **SLO 3.3** describe UniProt, the Protein Data Bank, and protein family databases and explain the types of protein data each provides, and
- **SLO 3.4** describe BLAST and Entrez database searching and explain how to interpret database search results.

3.1 Principles of Biological Database Design

3.1.1 Types of biological databases

Biological databases can be classified in several ways: by data type (sequence databases storing nucleotide or amino acid sequences; structure databases storing three-dimensional atomic coordinates; expression databases storing gene expression measurements; literature databases storing scientific publications and their curated biological content); by scope (primary databases that store original experimental data submitted directly by researchers versus secondary or derived databases that store data processed, curated, or derived from primary databases); and by curation approach (manually curated databases where biological experts review and annotate each entry versus automatically annotated databases where computational methods assign annotations, typically with lower reliability but higher coverage).

The distinction between primary and secondary databases is important for understanding the reliability of database content: GenBank is a primary nucleotide sequence database (sequences are deposited by submitters with minimal curation beyond basic format checking), while Swiss-Prot (the manually curated component of UniProt) is a secondary database in which expert curators review and annotate protein sequences, adding functional information from the literature and cross-references to other databases; the higher curation effort in Swiss-Prot makes its annotations more reliable but means it contains fewer entries than the automatically annotated TrEMBL component of UniProt.

Fill in the blank: _____ databases store original experimental data submitted directly by researchers, while secondary or derived databases store data processed or curated from primary sources.

3.1.2 Database architecture and annotation

A biological database record (or entry) consists of several components: a unique accession number (a stable identifier assigned to the entry when it is first submitted, which does not change even if the sequence or annotation is updated), a sequence identifier or locus name (a more human-readable but potentially variable identifier), the biological sequence itself (nucleotide or

amino acid), and annotation (structured information describing the biological properties of the sequence, including its source organism, gene name, protein function, subcellular localisation, known variants, cross-references to other databases, and literature citations from which the annotation was derived).

Annotation in biological databases is captured using controlled vocabularies (defined sets of standardised terms used consistently across entries to allow computational queries) and ontologies (hierarchically organised systems of controlled vocabulary terms with defined relationships between terms); the most widely used biological ontology is the Gene Ontology (GO), which organises gene and protein functions into three hierarchical namespaces: molecular function (the biochemical activity of the protein), biological process (the broader biological context in which the molecular function operates), and cellular component (where in the cell the protein is active); GO annotations are assigned to protein entries in databases such as UniProt and used extensively in functional genomics analysis.

Fill in the blank: The Gene _____ (GO) organises gene and protein functions into three hierarchical namespaces: molecular function, biological process, and cellular component.

3.1.3 Data standards and interoperability

Biological databases use standardised data formats to ensure that data can be exchanged between databases and read by different software tools: the FASTA format (a simple plain-text format with a header line beginning with ">", followed by sequence data, used universally for nucleotide and amino acid sequences); the GenBank flat file format (a richly annotated format used by NCBI databases with structured fields for sequence metadata and biological annotation); FASTQ format (for raw sequencing reads with quality scores); and for structural data, the PDB format (for three-dimensional atomic coordinates) and mmCIF/PDBx format (the more modern structured data exchange format for structural data).

Interoperability between biological databases is achieved through cross-referencing (each database entry carries identifiers from related databases, allowing automated linking), common ontologies (using GO terms and other shared vocabularies that mean the same thing across databases), and programmatic access through standardised application programming interfaces

(APIs) that allow software tools to query multiple databases using consistent protocols; the INSDC (International Nucleotide Sequence Database Collaboration) represents the highest level of database interoperability, where GenBank, EMBL-Bank, and DDBJ exchange data daily so that a sequence submitted to any one of them is immediately available through all three.

Fill in the blank: The _____ format is a simple plain-text sequence format with a header line beginning with ">" followed by sequence data, used universally for nucleotide and amino acid sequences.

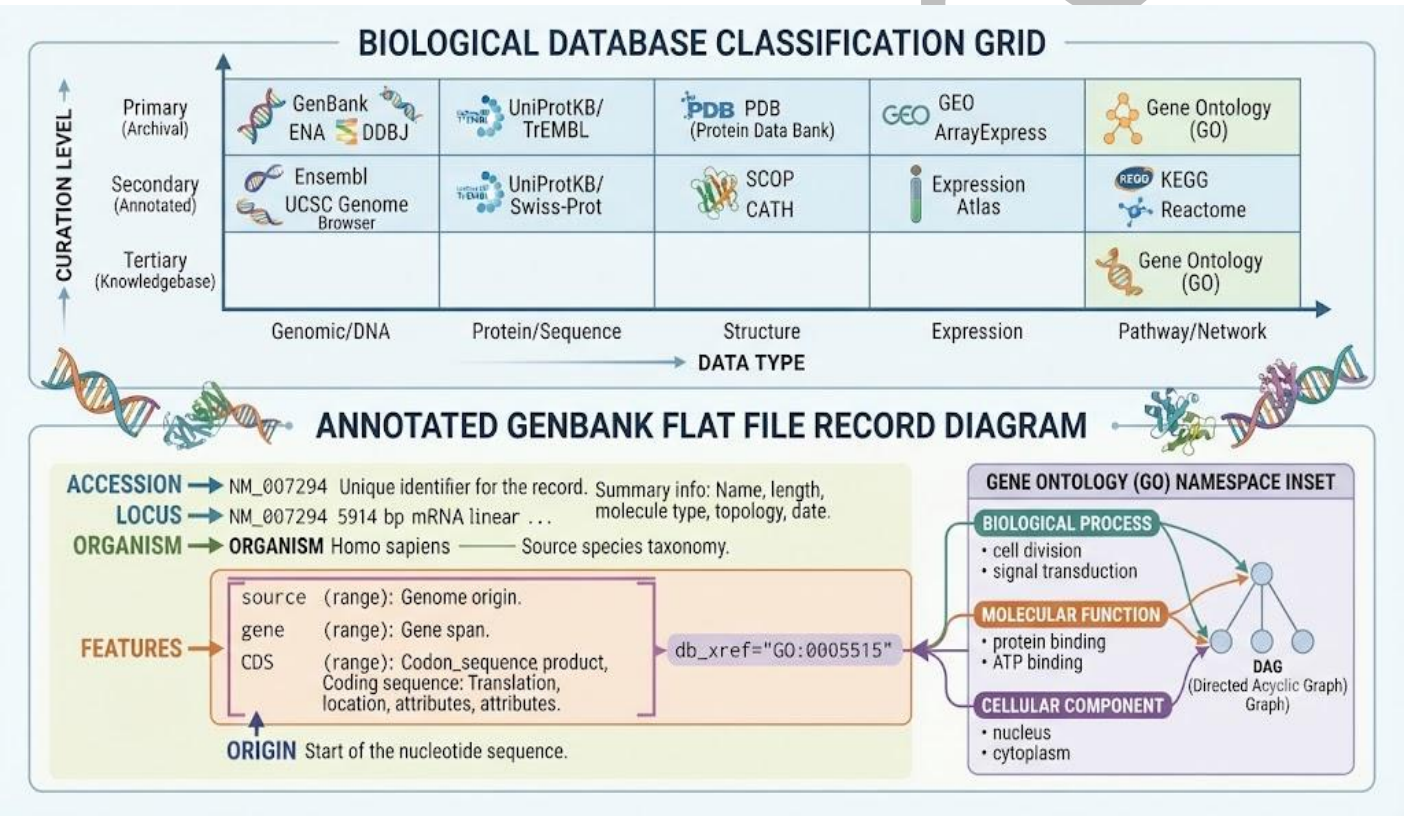

Figure 3.1: Biological database types and design

Pilot questions 3.1

1. Distinguish between primary and secondary biological databases and give one example of each.
2. Distinguish between manual and automatic database curation and explain the trade-off between them.
3. Name three standardised file formats used in biological databases and state what each stores.
4. Explain what the Gene Ontology is and name its three namespaces.

5. Explain how interoperability between biological databases is achieved.

3.2 Nucleotide Sequence Databases

3.2.1 GenBank, EMBL-Bank, and DDBJ

The International Nucleotide Sequence Database Collaboration (INSDC) comprises three partner databases that together form the definitive global repository for nucleotide sequence data: GenBank (maintained by the National Center for Biotechnology Information, NCBI, at the National Institutes of Health in the United States), the European Nucleotide Archive (ENA, formerly EMBL-Bank, maintained by the European Bioinformatics Institute at the Wellcome Genome Campus in Hinxton, United Kingdom), and the DNA Data Bank of Japan (DDBJ, maintained by the National Institute of Genetics in Mishima, Japan); the three databases synchronise data daily, so any sequence submitted to one is automatically available through all three.

GenBank grows exponentially, containing hundreds of billions of nucleotide bases as of the mid-2020s, with the majority of this data consisting of raw sequencing reads submitted through the Sequence Read Archive (SRA, the repository for raw NGS data that accompanies GenBank) rather than assembled and annotated sequences; accessing NCBI databases is done primarily through the Entrez web interface (entrez.ncbi.nlm.nih.gov), which provides unified search across all NCBI databases, or programmatically through the NCBI Entrez Utilities (E-utilities) API; ENA is accessed through the ENA browser (www.ebi.ac.uk/ena) and provides similar programmatic access through its REST API.

Fill in the blank: The International _____ Sequence Database Collaboration (INSDC) comprises GenBank (USA), ENA (UK), and DDBJ (Japan), which synchronise data daily to provide a unified global nucleotide repository.

3.2.2 Genome databases and reference sequences

Reference genome sequences (complete or nearly complete assembled sequences of one or more representative individuals of a species, serving as the standard coordinate system for all subsequent genomic analysis of that species) are maintained in specialised genome databases: the NCBI RefSeq database (Reference Sequence database) provides curated, non-redundant reference sequences for chromosomes, genes, transcripts, and proteins across thousands of organisms, with stable accession numbers and regular updates; the Ensembl genome browser (www.ensembl.org) provides reference genomes and extensive annotation for vertebrates and many other organisms with a focus on comparative genomics.

Organism-specific genome databases are maintained for the most important model organisms and pathogens: SGD (Saccharomyces Genome Database, for baker yeast), FlyBase (for Drosophila melanogaster), WormBase (for Caenorhabditis elegans), TAIR (The Arabidopsis Information Resource, for the model plant Arabidopsis thaliana), and PlasmoDB and VectorBase (for Plasmodium and mosquito vector species respectively, of particular relevance to Nigerian malaria research); these organism-specific databases provide deeper, more specialised annotation than the general-purpose databases, including information on mutant phenotypes, genetic interactions, and community-curated gene function.

Fill in the blank: The NCBI _____ Seq database provides curated, non-redundant reference sequences for chromosomes, genes, transcripts, and proteins across thousands of organisms with stable accession numbers.

3.2.3 Specialised nucleotide databases

Beyond the primary sequence databases, several specialised nucleotide databases serve specific research communities: the Sequence Read Archive (SRA, maintained by NCBI) and its partner European Nucleotide Archive (ENA) short read data repository store raw unassembled sequencing reads from NGS experiments, enabling re-analysis of published datasets; the miRBase database stores microRNA sequences and their targets; the RFAM database stores non-coding RNA sequence families and their secondary structures (using covariance models derived

from multiple sequence alignments); and the Ribosomal Database Project (RDP) and SILVA databases store ribosomal RNA sequences used for taxonomic classification in metagenomics studies.

The dbSNP (database of Single Nucleotide Polymorphisms) and dbVar (database of genomic structural variation) databases at NCBI store known human genetic variants, providing reference sets for variant annotation in clinical genomics; ClinVar stores clinically interpreted variants with their evidence for pathogenicity, essential for interpreting variants identified in patient genomes; COSMIC (Catalogue of Somatic Mutations in Cancer) stores somatic mutations found in cancer samples across many cancer types, used in cancer genomics research; the availability of these specialised databases reflects the depth and breadth of molecular biological knowledge now accessible through bioinformatics infrastructure.

Fill in the blank: _____ stores clinically interpreted human genetic variants with their evidence for pathogenicity, essential for interpreting variants identified in patient genome sequences.

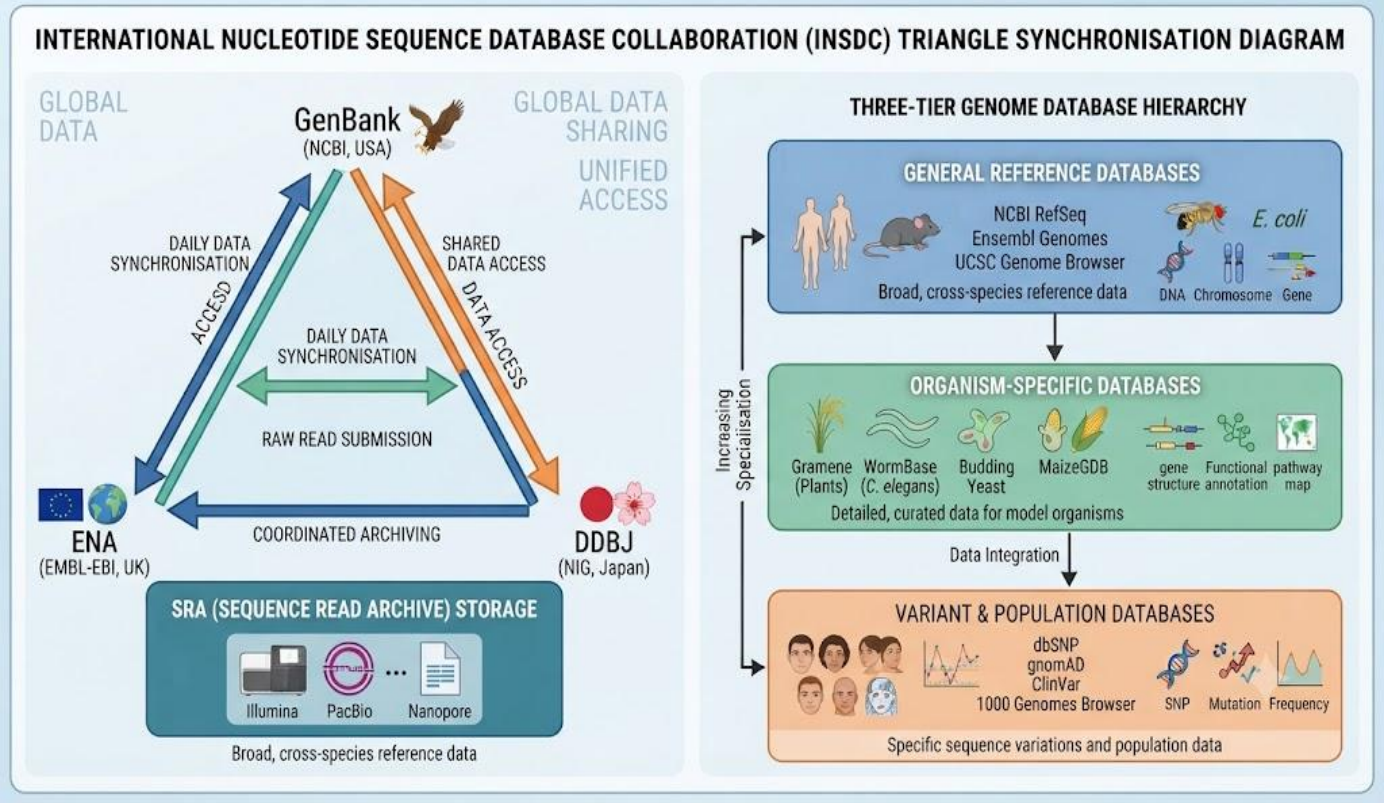

Figure 3.2: Nucleotide sequence databases

Pilot questions 3.2

1. Name the three members of the INSDC and state where each is maintained.
2. Explain what the Sequence Read Archive (SRA) stores and why it is important.
3. Explain what a reference genome sequence is and name one database that provides them.
4. Name two organism-specific genome databases relevant to malaria research.
5. Name two specialised nucleotide variant databases and state what each contains.

3.3 Protein Sequence and Structure Databases

3.3.1 UniProt: Swiss-Prot and TrEMBL

UniProt (Universal Protein Resource, www.uniprot.org) is the primary global repository for protein sequence and functional information, comprising two components with different levels of

curation: Swiss-Prot (the manually reviewed section) and TrEMBL (the unreviewed section); Swiss-Prot entries are reviewed by expert curators who verify the sequence, add functional annotation from the published literature (enzyme activity, subcellular localisation, post-translational modifications, disease associations, active site residues, and binding sites), assign GO terms, and add cross-references to other databases; each Swiss-Prot entry undergoes a thorough review process that ensures high annotation quality.

TrEMBL (originally Translated EMBL, now named UniProtKB/TrEMBL) contains protein sequences translated from nucleotide sequences in EMBL-Bank and other sources that have not yet been manually reviewed; entries are annotated computationally by the UniProt automatic annotation pipeline, which assigns functional information based on sequence similarity to reviewed Swiss-Prot entries using rule-based and machine learning approaches; TrEMBL contains the vast majority of UniProt entries (over 200 million sequences as of the mid-2020s versus approximately 570,000 in Swiss-Prot), reflecting the exponential growth of sequencing data, but with lower annotation reliability than Swiss-Prot; users should be aware of which component of UniProt they are consulting when evaluating the reliability of functional annotations.

Fill in the blank: UniProt comprises two components: _____ -Prot (manually reviewed, high-quality annotation) and TrEMBL (computationally annotated, very large, lower annotation reliability).

3.3.2 The Protein Data Bank

The Protein Data Bank (PDB, www.rcsb.org) is the single worldwide repository for three-dimensional structures of proteins, nucleic acids, and protein-nucleic acid complexes determined by X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, cryo-electron microscopy (cryo-EM), and other experimental methods; established in 1971 at Brookhaven National Laboratory and now maintained by the Research Collaboratory for Structural Bioinformatics (RCSB), the PDB contains over 200,000 structures covering a wide range of organisms and biological contexts.

Each PDB entry contains the three-dimensional atomic coordinates of the structure, experimental details (method, resolution, and quality indicators), and annotation linking the structure to the corresponding protein sequence in UniProt and relevant biological literature; PDB structures are accessed through the RCSB PDB web interface, which provides tools for three-dimensional molecular visualisation (using the Mol* viewer), sequence and structural similarity searching, and download of coordinate files in PDB or mmCIF format; the complementary AlphaFold Protein Structure Database (alphafold.ebi.ac.uk) provides predicted structures for essentially all known proteins, greatly expanding structural coverage beyond experimentally determined structures.

Fill in the blank: The Protein Data Bank (PDB) contains over 200,000 experimentally determined three-dimensional structures of proteins and nucleic acids, determined by X-ray crystallography, _____ spectroscopy, and cryo-electron microscopy.

3.3.3 Protein family and domain databases

Protein family and domain databases organise proteins into groups based on shared sequence patterns, structural folds, or evolutionary relationships, providing a higher-level view of protein function and evolution than individual sequence or structure databases; Pfam (now part of the InterPro database) is the most widely used protein domain database, storing curated families of protein domains defined by multiple sequence alignments and represented as profile hidden Markov models (HMMs) that can be used to detect distantly related members; each Pfam family has a curated description, functional annotation, and a seed alignment of representative members.

InterPro is a meta-database that integrates protein family and domain signatures from multiple member databases (including Pfam, PROSITE, PRINTS, PANTHER, HAMAP, CDD, SUPERFAMILY, and others) into a unified resource, providing a comprehensive view of the domain composition of a protein query; the SCOP (Structural Classification of Proteins) and CATH (Class, Architecture, Topology, Homologous superfamily) databases classify protein structures into hierarchical categories based on structural similarity, revealing evolutionary relationships that may not be detectable at the sequence level; KEGG (Kyoto Encyclopedia of

Genes and Genomes) provides pathway-level annotation linking genes and proteins to metabolic and signalling pathways.

Fill in the blank: _____ integrates protein family and domain signatures from multiple member databases including Pfam, PROSITE, and PANTHER into a unified resource providing comprehensive domain annotation.

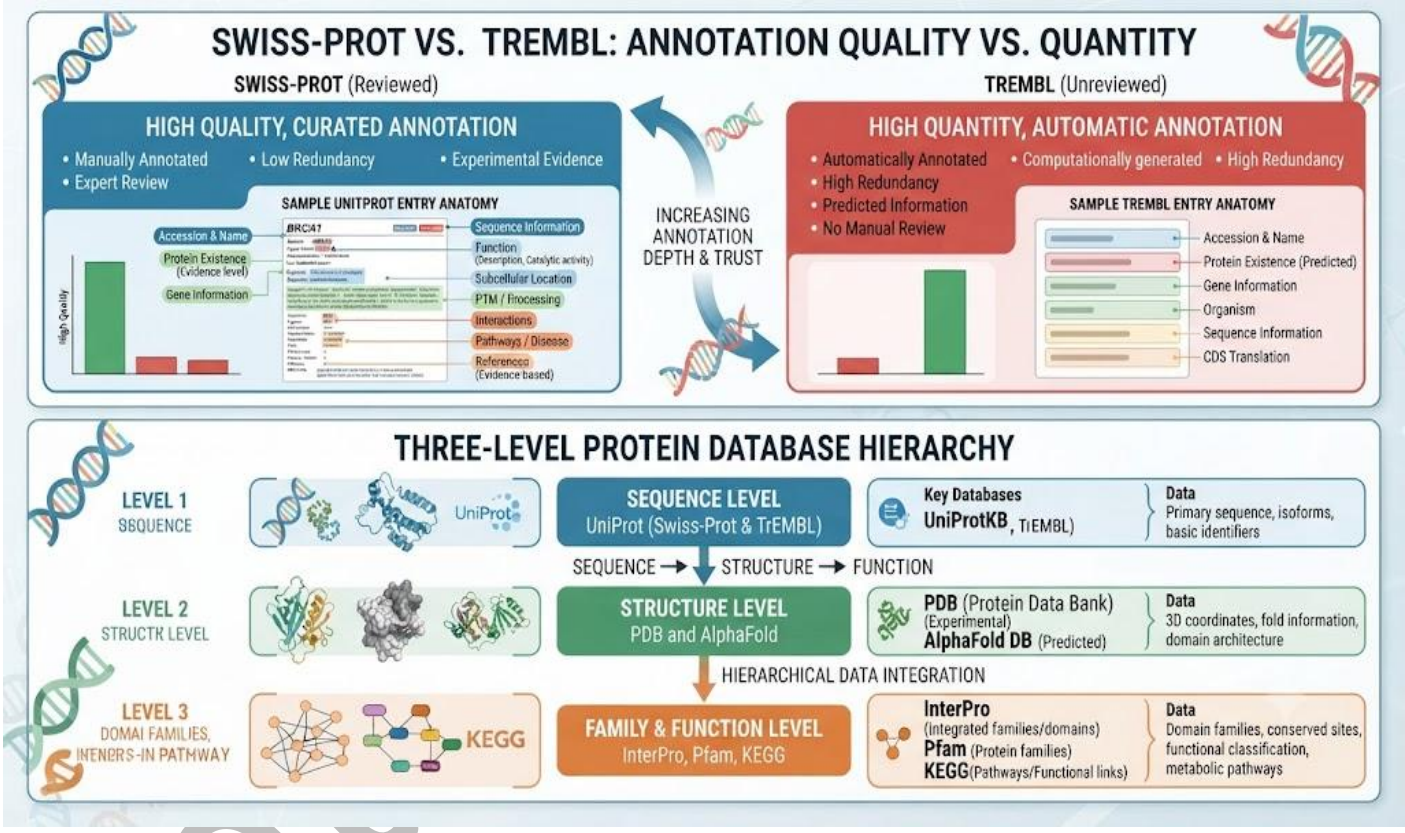

Figure 3.3: Protein sequence and structure databases

Pilot questions 3.3

1. Distinguish between Swiss-Prot and TrEMBL in terms of annotation quality, curation, and number of entries.
2. Name three types of annotation added to a Swiss-Prot entry by expert curators.
3. Explain what the Protein Data Bank contains and how it is accessed.
4. Explain what AlphaFold Protein Structure Database provides.
5. Explain what InterPro is and name three member databases it integrates.

3.4 Database Searching and Access

3.4.1 BLAST and sequence similarity searching

BLAST (Basic Local Alignment Search Tool) is the most widely used tool for searching biological databases by sequence similarity; it identifies sequences in a target database that are similar to a query sequence, using a heuristic algorithm that rapidly identifies short matching words (seeds) between query and database sequences, then extends high-scoring alignments from these seeds; BLAST is available through the NCBI web interface (blast.ncbi.nlm.nih.gov) and as a standalone command-line tool for local database searches; the appropriate BLAST programme depends on the query and database types: BLASTN (nucleotide query vs nucleotide database), BLASTP (protein vs protein), BLASTX (nucleotide vs protein, translating the nucleotide query in all six reading frames), TBLASTN (protein vs nucleotide, translating the database), and TBLASTX (nucleotide vs nucleotide, translating both).

BLAST results are reported as a list of database hits ranked by their Expect value (E-value), a statistical measure representing the number of hits with an equal or better score that would be expected by chance when searching a database of the given size; a lower E-value indicates a more statistically significant match; conventional thresholds are E-value below 10 to the power of minus 5 (0.00001) as a reasonable threshold for a meaningful match, though the appropriate threshold depends on the specific application; each hit is reported with the sequence identity (percentage of aligned positions that are identical), the alignment coverage, and a bitscore (a normalised alignment score that allows comparison across searches of different database sizes).

Fill in the blank: In BLAST results, the _____ value represents the number of hits with an equal or better score expected by chance in a database of the given size; a lower value indicates a more statistically significant match.

3.4.2 Entrez and programmatic database access

The Entrez system is NCBI unified search and retrieval system, providing a single web interface for searching across all NCBI databases simultaneously (including GenBank nucleotide, RefSeq, UniProt protein, PubMed literature, dbSNP, ClinVar, SRA, and many others) using a common

query syntax; Entrez queries can include Boolean operators (AND, OR, NOT) and field tags (for example, "BRCA1[gene] AND Homo sapiens[organism] AND RefSeq[filter]" to retrieve RefSeq transcripts of the human BRCA1 gene), allowing precise retrieval of specific data types from among the billions of entries in NCBI databases.

Programmatic access to NCBI databases is provided through the E-utilities API (a set of web services accessible at eutils.ncbi.nlm.nih.gov), which allows automated retrieval of large numbers of records, submission of sequences, and integration of NCBI data into bioinformatics pipelines; in Python, the Biopython Entrez module provides a high-level interface to the E-utilities, enabling programmatic searching, record retrieval, and parsing in a few lines of code; the EBI also provides REST API access to its databases including UniProt, ENA, and PDB through the EBI search system; programmatic database access is essential for any bioinformatics analysis involving large numbers of records that cannot be retrieved efficiently through a web browser.

Fill in the blank: The NCBI _____ system provides a unified web interface and API for searching across all NCBI databases simultaneously using Boolean operators and field tags.

3.4.3 Interpreting database search results

Interpreting a biological database search result requires evaluating both the statistical significance and the biological significance of the hits returned; the E-value assesses statistical significance (whether the match is likely to have occurred by chance) but a statistically significant match is not necessarily biologically meaningful: a match may be statistically significant because two proteins share a common low-complexity region (such as a repetitive sequence) rather than genuine homology; the percentage identity and alignment coverage provide additional information about how similar the hit is to the query, with higher identity and coverage generally indicating a closer evolutionary relationship.

A retrieved database entry should be evaluated for the quality of its annotation: Swiss-Prot entries with manually curated functional information are more reliable than TrEMBL entries with only computational annotation; an annotation flagged as "by similarity" (assigned based on sequence similarity to a characterised protein rather than direct experimental evidence) should be

treated with more caution than one supported by direct experimental evidence; when using BLAST results to infer protein function, it is important to retrieve the full annotation of the top hits, examine the literature supporting the annotation, and consider whether the functional conservation suggested by sequence similarity is supported by structural or experimental data in the related proteins.

Fill in the blank: A UniProt annotation flagged as "by _____ " has been assigned based on sequence similarity to a characterised protein rather than direct experimental evidence, and should be treated with appropriate caution.

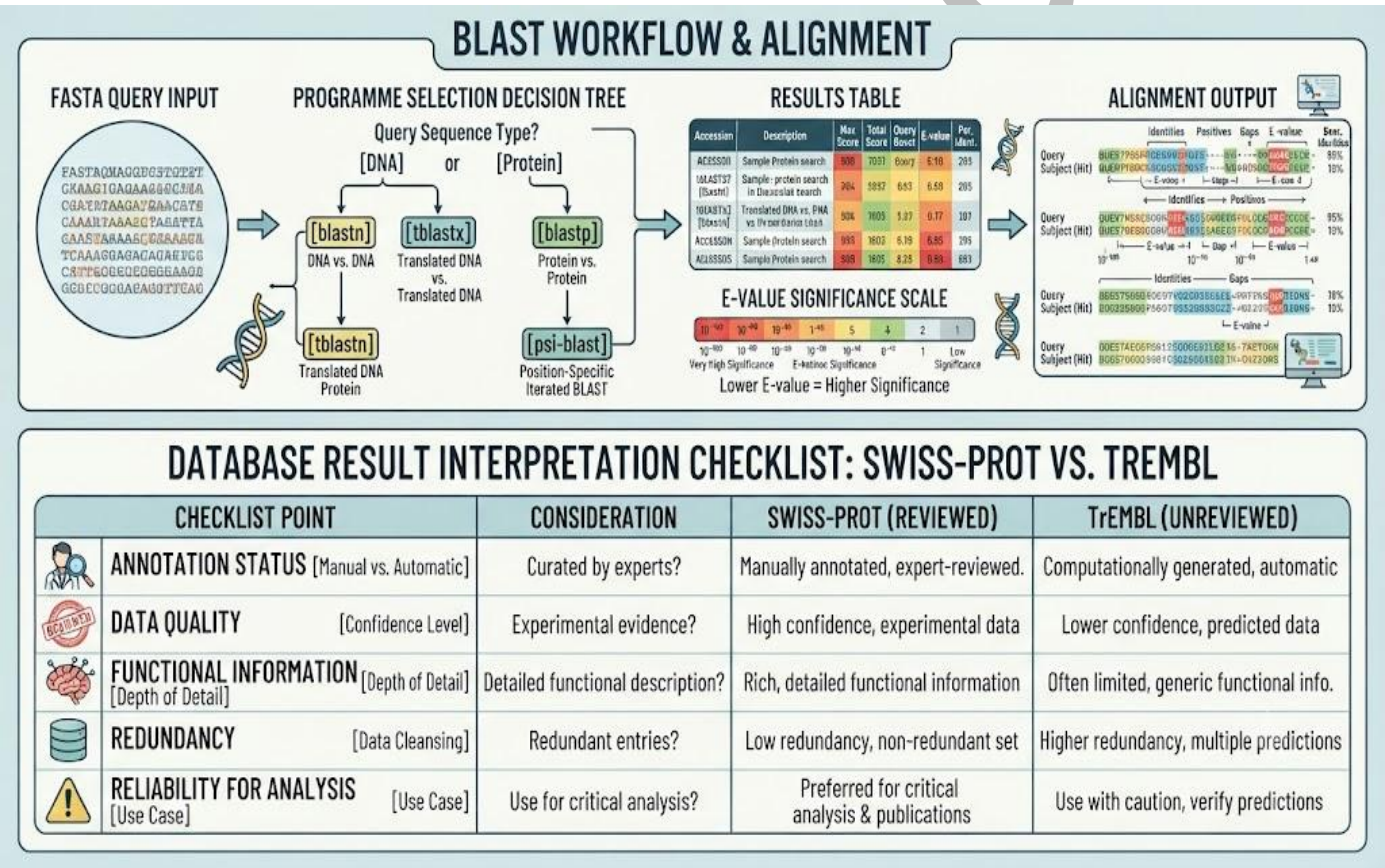

Figure 3.4: Database searching and result interpretation

Pilot questions 3.4

1. Name the five BLAST programme variants and explain when each is used.
2. Explain what the E-value in a BLAST result represents and state a conventional significance threshold.

3. Explain what the Entrez system is and give one example of a field-tagged query.
4. Explain how to access NCBI databases programmatically using Python.
5. Explain what "by similarity" means in a UniProt annotation and why it requires caution.

Series Summary

This Series examined the DNA and protein databases that form the data infrastructure of bioinformatics. Biological databases range from primary repositories of raw experimental data to expert-curated secondary resources, and from broadly annotated general databases to deeply specialised organism-specific collections. The INSDC partners GenBank, ENA, and DDBJ collectively maintain the global nucleotide sequence archive, synchronised daily, while RefSeq and Ensembl provide curated reference genomes and annotation. Gene Ontology terms and standardised file formats underpin the interoperability that allows biological knowledge to flow between databases and analytical tools.

UniProt is the primary protein sequence resource, with Swiss-Prot providing high-quality manually curated annotation for approximately 570,000 proteins and TrEMBL providing computationally annotated coverage for over 200 million. The Protein Data Bank supplies experimental three-dimensional structures, now complemented by AlphaFold predicted structures for essentially all known proteins. InterPro and Pfam organise proteins into evolutionary families and domains. BLAST remains the fundamental database search tool, and understanding E-values, alignment statistics, and evidence codes is essential for correctly interpreting the results of any database search.

Glossary of Terms

- **Accession number:** a stable, unique identifier assigned to a database record at submission, which does not change even if the sequence or annotation is updated.
- **BLAST (Basic Local Alignment Search Tool):** a heuristic algorithm for searching biological databases by sequence similarity, reporting hits ranked by statistical significance (E-value).
- **E-value (Expect value):** the expected number of database hits with an equal or better alignment score than the observed hit that would occur by chance; a lower E-value indicates greater statistical significance.

- **Entrez:** the NCBI unified search and retrieval system providing access to all NCBI databases through a common web interface and E-utilities programmatic API.
- **Gene Ontology (GO):** a hierarchical controlled vocabulary for describing gene and protein functions across three namespaces: molecular function, biological process, and cellular component.
- **INSDC:** the International Nucleotide Sequence Database Collaboration, comprising GenBank, ENA, and DDBJ, which exchange nucleotide sequence data daily.
- **InterPro:** a meta-database integrating protein family and domain signatures from multiple databases including Pfam, PROSITE, and PANTHER.
- **Pfam:** a database of protein domain families defined by multiple sequence alignments and profile hidden Markov models.
- **Swiss-Prot:** the manually reviewed, high-quality protein sequence and annotation section of UniProt, containing approximately 570,000 expert-curated entries.
- **TrEMBL:** the computationally annotated, unreviewed section of UniProt, containing over 200 million protein sequences with automatically assigned annotations.

Concept Check Questions [Series 3]

Objective Questions

1. _____ databases store original experimental data submitted by researchers, while secondary databases store processed or curated data derived from primary sources. (Linked to SLO 3.1)
2. The Gene _____ (GO) organises gene functions into molecular function, biological process, and cellular component namespaces. (Linked to SLO 3.1)
3. The INSDC comprises GenBank (USA), _____ (UK), and DDBJ (Japan), which synchronise data daily. (Linked to SLO 3.2)
4. UniProt comprises _____ -Prot (manually reviewed) and TrEMBL (computationally annotated). (Linked to SLO 3.3)
5. In BLAST results, the _____ value represents the expected number of chance hits of equal or better score in the database. (Linked to SLO 3.4)

Short-Answer Questions

1. Distinguish between primary and secondary biological databases and give one example of each. (Linked to SLO 3.1)
2. Name the three members of the INSDC and state where each is maintained. (Linked to SLO 3.2)
3. Distinguish between Swiss-Prot and TrEMBL in terms of annotation and number of entries. (Linked to SLO 3.3)
4. Name five BLAST programme variants and state when each is used. (Linked to SLO 3.4)
5. Explain what "by similarity" means in a UniProt annotation. (Linked to SLO 3.4)

Long-Answer Questions

1. Describe the principles of biological database design and the major nucleotide sequence databases. (Linked to SLOs 3.1, 3.2)
2. Describe UniProt, the Protein Data Bank, and protein family databases, and explain how to search and interpret biological database results. (Linked to SLOs 3.3, 3.4)

Answers to Concept Check Questions [Series 3]

Objective Questions

1. Primary.
2. Ontology.
3. ENA.
4. Swiss.
5. E.

Short-Answer Questions

1. Primary databases store original experimental data submitted by researchers (e.g. GenBank); secondary databases store curated or derived data processed from primary sources (e.g. Swiss-Prot).
2. GenBank (NCBI, USA), ENA (EBI, UK), and DDBJ (National Institute of Genetics, Japan).
3. Swiss-Prot is manually reviewed by expert curators with high-quality annotation, approximately 570,000 entries; TrEMBL is computationally annotated with lower reliability, over 200 million entries.
4. BLASTN (nucleotide vs nucleotide), BLASTP (protein vs protein), BLASTX (nucleotide vs protein, translating query), TBLASTN (protein vs nucleotide, translating database), TBLASTX (nucleotide vs nucleotide, translating both).
5. "By similarity" means the annotation was assigned based on sequence similarity to a characterised protein rather than direct experimental evidence from the protein itself; it is less reliable because the function may not be conserved despite sequence similarity.

Long-Answer Questions

1. Database design: primary (raw experimental data, e.g. GenBank) vs secondary (curated, e.g. Swiss-Prot); manual vs automatic curation; annotation using controlled vocabularies and Gene Ontology (molecular function, biological process, cellular component); standardised formats (FASTA, GenBank flat file, PDB); interoperability through cross-references, shared ontologies, and APIs. INSDC: GenBank, ENA, DDBJ synchronised daily; RefSeq (curated reference sequences); Ensembl (vertebrate genomes); organism-specific (SGD, FlyBase, PlasmoDB); specialised (dbSNP, ClinVar, COSMIC, SRA).

2. UniProt: Swiss-Prot (manually curated, ~570,000, high quality annotations including function, GO terms, active sites, disease associations) and TrEMBL (computationally annotated, >200 million, lower reliability). PDB: experimental 3D structures (X-ray, NMR, cryo-EM, >200,000); AlphaFold database predicts structures for all known proteins. InterPro (integrates Pfam, PROSITE, PANTHER), Pfam (domain HMMs), KEGG (pathways). BLAST: five programmes for different query/database combinations; results ranked by E-value (lower = more significant, conventional threshold E < 10^-5); evaluate Swiss-Prot vs TrEMBL, evidence codes, experimental vs inferred annotation.

Answers to Pilot Questions [Series 3]

Part 3.1

1. Primary databases store original experimental data directly submitted by researchers (example: GenBank stores nucleotide sequences as submitted); secondary databases store data curated or derived from primary sources (example: Swiss-Prot stores expert-curated protein annotation derived from literature and primary databases).
2. Manual curation: experts review each entry, ensuring high annotation quality; slower, fewer entries. Automatic curation: computational methods annotate entries based on sequence similarity and rules; faster, broader coverage but lower reliability; trade-off is quality versus quantity.
3. FASTA (plain text nucleotide and amino acid sequences), GenBank flat file (annotated nucleotide sequences with structured metadata), PDB format (three-dimensional atomic coordinates of protein/nucleic acid structures).
4. The Gene Ontology is a hierarchical controlled vocabulary for gene and protein function; three namespaces: molecular function (biochemical activity), biological process (broader cellular context), and cellular component (where in the cell).
5. Cross-referencing (identifiers from related databases in each entry), shared ontologies (GO terms meaning the same across databases), and programmatic APIs (standardised interfaces for automated data exchange); INSDC represents the highest level through daily full data synchronisation.

Part 3.2

1. GenBank (NCBI, Bethesda, USA), ENA (European Bioinformatics Institute, Hinxton, UK), DDBJ (National Institute of Genetics, Mishima, Japan).
2. The SRA stores raw unassembled sequencing reads from NGS experiments; it is important because it enables re-analysis of published datasets with improved tools and for secondary analysis of existing data without new experiments.
3. A reference genome is a complete or near-complete assembled genome sequence serving as the standard coordinate system for all subsequent genomic analysis; NCBI RefSeq provides curated reference sequences for thousands of organisms.
4. PlasmoDB (for Plasmodium malaria parasite genomes) and VectorBase (for Anopheles mosquito vector genomes).
5. dbSNP (database of Single Nucleotide Polymorphisms, known human SNPs) and ClinVar (clinically interpreted variants with pathogenicity evidence for use in patient genome interpretation).

Part 3.3

1. Swiss-Prot: manually reviewed, high-quality annotation (function, GO terms, active sites, disease associations, post-translational modifications), approximately 570,000 entries. TrEMBL: computationally annotated, lower annotation reliability, over 200 million entries; users should check which section they are consulting.
2. Enzyme activity and kinetic parameters, subcellular localisation, and disease associations (or active site residues, post-translational modifications, GO terms).
3. The PDB contains over 200,000 experimentally determined three-dimensional structures of proteins and nucleic acids; accessed through www.rcsb.org with the Mol* viewer for 3D visualisation and download of PDB/mmCIF coordinate files.
4. The AlphaFold Protein Structure Database provides computationally predicted three-dimensional structures for essentially all known proteins (over 200 million), greatly extending structural coverage beyond the approximately 200,000 experimentally determined PDB structures.
5. InterPro is a meta-database integrating protein family and domain signatures from multiple databases; three member databases: Pfam (protein domain HMMs), PROSITE (sequence patterns and profiles), and PANTHER (protein families and evolutionary trees).

Part 3.4

1. BLASTN (nucleotide query vs nucleotide database, for searching DNA sequences), BLASTP (protein vs protein, for searching protein databases), BLASTX (nucleotide vs protein, translating query in six frames, for finding protein matches for a new nucleotide sequence), TBLASTN (protein vs nucleotide, translating database, for finding novel genes encoding a known protein), TBLASTX (nucleotide vs nucleotide, translating both, for sensitive cross-species comparison).
2. The E-value represents the expected number of database hits with an equal or better alignment score that would occur by chance in a database of the given size; conventional threshold is E-value below 10 to the power of minus 5 (0.00001) for a meaningful match, though the appropriate threshold depends on the application.
3. Entrez is the NCBI unified search and retrieval system; example query: "BRCA1[gene] AND Homo sapiens[organism] AND RefSeq[filter]" retrieves RefSeq transcripts of the human BRCA1 gene.
4. Using the Biopython Entrez module in Python: Entrez.esearch() performs a search and returns record IDs, Entrez.efetch() retrieves full records for those IDs, and Entrez.read() parses the returned XML or GenBank-formatted records into Python data structures.
5. "By similarity" means the annotation was assigned computationally based on sequence similarity to a Swiss-Prot-reviewed protein rather than experimental evidence from the protein itself; caution is warranted because functional conservation cannot always be assumed from sequence similarity, particularly for distantly related proteins or for specific substrate or cofactor interactions.

References and Attributions

Textual References (Books and External Sources)

- Lesk, A. M. (2019). Introduction to bioinformatics (4th ed.). Oxford University Press.
- Baxevanis, A. D., Bader, G. D., & Wishart, D. S. (Eds.). (2020). Bioinformatics (4th ed.). Wiley-Blackwell.
- UniProt Consortium. (2023). UniProt: The universal protein knowledgebase in 2023. Nucleic Acids Research, 51(D1), D523-D531.
- National Open University of Nigeria (NOUN). (2023). Bioinformatics course material. NOUN Press.

Figures, Plates, Tables, and Boxes

- Figure 3.1: Biological database types and design Attribution: AI generated. Suggested OER search string: "biological database types annotation Gene Ontology controlled vocabulary GenBank Swiss-Prot interoperability OER".
- Figure 3.2: Nucleotide sequence databases Attribution: AI generated. Suggested OER search string: "GenBank EMBL ENA DDBJ INSDC RefSeq Ensembl ClinVar dbSNP nucleotide sequence database OER".
- Figure 3.3: Protein sequence and structure databases Attribution: AI generated. Suggested OER search string: "UniProt Swiss-Prot TrEMBL Protein Data Bank Pfam InterPro KEGG protein database structure domain OER".
- Figure 3.4: Database searching and result interpretation Attribution: AI generated. Suggested OER search string: "BLAST E-value database search result interpretation Swiss-Prot annotation evidence bioinformatics OER".

Course code: BIO 413

Course title: Bioinformatics

Course credit load: 2 Units

Series 4: Data Storage and Retrieval

- **Part 4.1: Biological Data File Formats**
  - Sub Part 4.1.1: Sequence file formats
  - Sub Part 4.1.2: Alignment and variant file formats
  - Sub Part 4.1.3: Structural and expression data formats
- **Part 4.2: Data Storage Systems**
  - Sub Part 4.2.1: Relational databases and SQL
  - Sub Part 4.2.2: NoSQL databases for biological data
  - Sub Part 4.2.3: Cloud storage and data repositories
- **Part 4.3: Sequence Alignment and Multiple Sequence Alignment**
  - Sub Part 4.3.1: Pairwise sequence alignment
  - Sub Part 4.3.2: Multiple sequence alignment
  - Sub Part 4.3.3: Alignment quality and applications
- **Part 4.4: Data Retrieval, Mining, and Integration**
  - Sub Part 4.4.1: Programmatic data retrieval
  - Sub Part 4.4.2: Data mining and knowledge extraction
  - Sub Part 4.4.3: Data integration and federated queries

Introduction

Biological data is generated in an enormous variety of formats, stored in diverse systems, and must be retrieved, processed, and integrated from multiple sources to answer most modern biological questions. Understanding how biological data is stored, what formats it uses, and how it can be retrieved and combined programmatically is as essential to a working bioinformatician as knowledge of the databases themselves, since even perfect database knowledge is of limited value without the practical ability to extract and work with the data they contain.

This Series covers the principal file formats used to store biological data at each stage of a typical bioinformatics analysis, from raw sequences through alignments and variants to structural coordinates and expression matrices; the database management systems used to store biological data at institutional and personal scales; the algorithms and tools for pairwise and multiple sequence alignment; and the methods for programmatic data retrieval, computational data mining, and integration of data from multiple heterogeneous sources into unified analytical datasets.

Series Learning Outcomes

Upon completing this Series, you should be able to:

- **SLO 4.1** describe the major biological data file formats and explain the type of data each stores and its typical use in bioinformatics pipelines,
- **SLO 4.2** describe relational databases, NoSQL databases, and cloud storage as data management systems used in bioinformatics,
- **SLO 4.3** explain pairwise and multiple sequence alignment, describe the algorithms used, and explain their applications, and
- **SLO 4.4** describe programmatic data retrieval, data mining, and data integration approaches used in bioinformatics.

4.1 Biological Data File Formats

4.1.1 Sequence file formats

The FASTA format is the simplest and most universally used biological sequence format: each sequence is represented by a header line beginning with ">" that contains the sequence identifier and optional description, followed by lines of nucleotide or amino acid sequence characters; FASTA files may contain a single sequence or multiple sequences (multi-FASTA); the format is used for reference genome sequences, protein sequences, primer sequences, and any other situation where sequence data needs to be stored or exchanged without quality score information.

The FASTQ format extends FASTA by adding a quality score for each base: each sequence record comprises four lines, the header line (beginning with "@"), the sequence, a separator line ("+"), and the quality string (ASCII-encoded Phred quality scores, where each character represents the probability that the corresponding base call is incorrect); FASTQ is the universal format for raw sequencing reads from NGS platforms and is the input format for quality control tools such as FastQC and for read alignment tools; the GenBank flat file format is the standard annotated sequence format used by NCBI databases, containing both sequence and extensive structured biological annotation in a defined field-based layout.

Fill in the blank: The FASTQ format extends FASTA by adding _____ score information for each base, encoded as ASCII Phred quality scores, making it the standard format for raw NGS sequencing reads.

4.1.2 Alignment and variant file formats

The SAM (Sequence Alignment/Map) format and its compressed binary equivalent BAM are the standard formats for storing the results of aligning short sequencing reads to a reference genome; a SAM file contains a header section (with metadata about the reference genome and alignment run) followed by one line per aligned read, containing the read name, alignment position on the reference, CIGAR string (a compact notation describing the alignment: matches, insertions, deletions, and clipped bases), the read sequence, quality scores, and a set of optional auxiliary fields for additional alignment information; BAM files are compressed binary equivalents of

SAM files, allowing faster access and random retrieval of alignments at specific genomic coordinates using an index file (.bai).

The VCF (Variant Call Format) is the standard format for storing genomic variants (SNPs, indels, structural variants) identified from sequencing data; a VCF file has a header section (with metadata and field definitions) followed by one line per variant, containing the chromosome, position, reference allele, alternative allele, quality score, filter status, and a set of INFO and FORMAT fields carrying additional information about the variant and per-sample genotype calls; BCF is the compressed binary equivalent of VCF; the BED (Browser Extensible Data) format is a simple tab-delimited format for representing genomic intervals (chromosomal coordinates of features such as genes, peaks, or regulatory elements), widely used for annotation and genome browser track display.

Fill in the blank: The _____ (Variant Call Format) is the standard format for storing genomic variants identified from sequencing data, with fields for chromosome, position, reference allele, alternative allele, and genotype calls.

4.1.3 Structural and expression data formats

The PDB format (and its successor mmCIF/PDBx format) stores three-dimensional atomic coordinates of protein and nucleic acid structures: the PDB format is a fixed-column-width text format with ATOM/HETATM records for each atom giving its element, residue, chain, position number, and x, y, z coordinates; mmCIF is a more modern, flexible dictionary-based format preferred for large structures and by the PDB for new depositions; molecular visualisation software such as PyMOL, UCSF Chimera, and the online Mol* viewer can read both formats.

Gene expression data from RNA-seq experiments is typically stored as count matrices (tables where rows are genes and columns are samples, with each cell containing the number of sequencing reads mapped to that gene in that sample) in tab-delimited text or comma-separated values (CSV) format, or in the HDF5 binary format used by tools such as CellRanger for single-cell RNA-seq data; microarray expression data is stored in CEL files (Affymetrix raw data), normalised text files, or in the SOFT and MINIML formats used by the GEO (Gene Expression

Omnibus) database for sharing expression datasets; the GEO database at NCBI is the primary public repository for high-throughput gene expression data.

Fill in the blank: Gene expression data from RNA-seq is typically stored as count _____ (tables of genes by samples with read count values) in text or HDF5 binary formats.

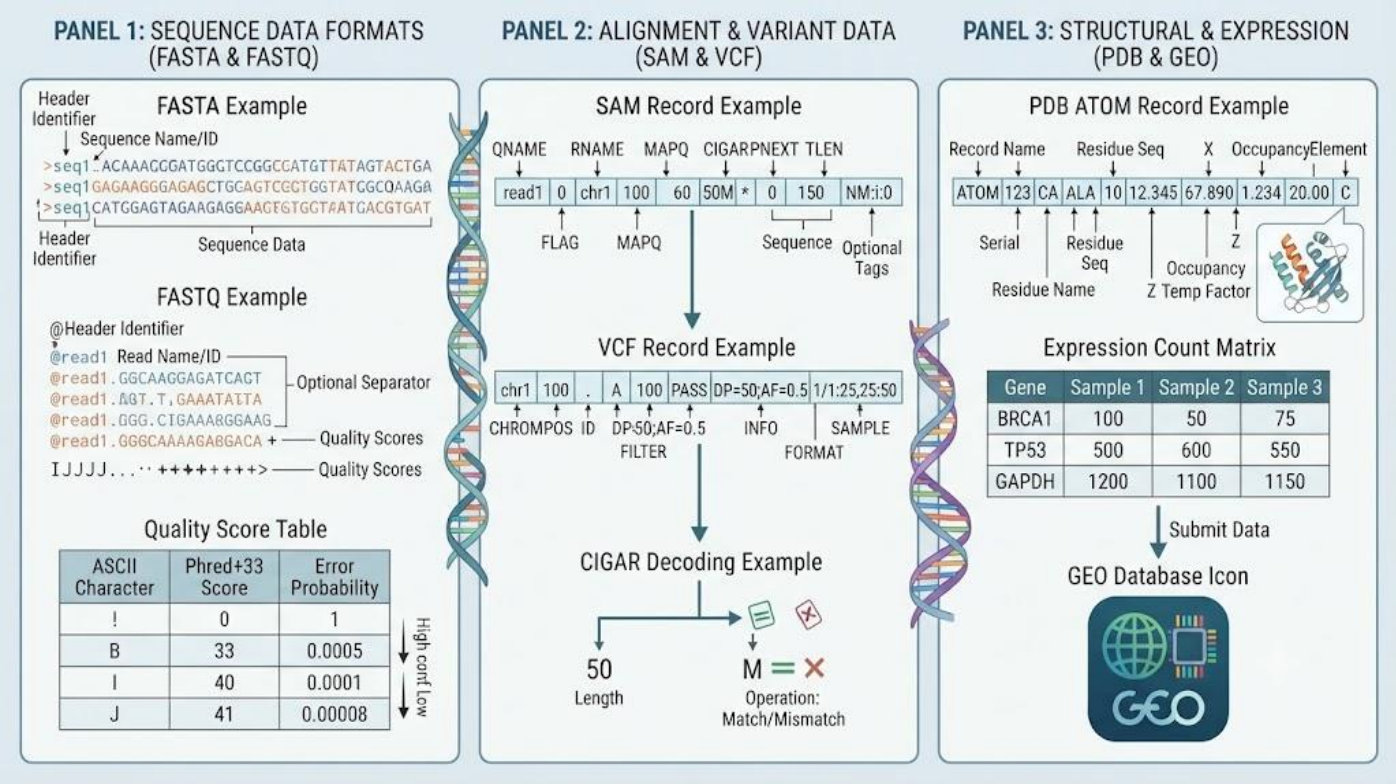

Figure 4.1: Biological data file formats

Pilot questions 4.1

1. Distinguish between FASTA and FASTQ formats and state when each is used.
2. Explain what a CIGAR string in a SAM file represents and give an example.
3. Explain what the VCF format stores and name three fields in a VCF record.
4. Describe the PDB file format and explain what ATOM records contain.
5. Explain what a gene expression count matrix is and what format it is typically stored in.

4.2 Data Storage Systems

4.2.1 Relational databases and SQL

Relational databases organise data into tables (relations) with rows (records) and columns (fields), where relationships between tables are expressed through shared key columns; the structured query language (SQL) is the standard language for creating, querying, and managing relational databases, and is the most widely used database management approach for structured biological data; SQL queries allow data to be selected (SELECT), filtered (WHERE), joined across tables (JOIN), grouped and aggregated (GROUP BY, COUNT, SUM), and modified (INSERT, UPDATE, DELETE); relational databases are particularly well-suited to storing highly structured biological data with well-defined relationships between entities, such as the metadata associated with sequence entries, experiment records, or clinical sample information.

The major public biological databases are implemented on relational database management systems (RDBMS): GenBank uses an Oracle RDBMS, Ensembl uses MySQL, and many local bioinformatics databases are implemented in PostgreSQL or SQLite; for personal bioinformatics database work, SQLite (a lightweight file-based relational database requiring no server) is often the most practical choice for storing and querying moderate-sized datasets; proficiency in basic SQL queries is a useful skill for bioinformaticians who need to query local or institutional biological data repositories, extract specific subsets of data, or join data from multiple related tables.

Fill in the blank: _____ is the standard language for creating and querying relational databases, using commands such as SELECT, WHERE, JOIN, and GROUP BY to retrieve and manipulate structured data.

4.2.2 NoSQL databases for biological data

NoSQL (Not Only SQL) databases are non-relational database systems that store data in formats other than tabular rows and columns, and are particularly suited to biological data types that do not fit neatly into a relational schema; the major NoSQL database types and their biological applications include: document databases (such as MongoDB, which store data as flexible

JSON-like documents, suitable for storing heterogeneous biological records with variable fields, such as genomic variant annotations or literature-derived biological facts); key-value stores (such as Redis, providing very fast retrieval of values by a unique key, useful for caching frequently accessed database lookups in bioinformatics pipelines).

Graph databases (such as Neo4j) store data as nodes and relationships, making them well-suited to biological network data such as protein-protein interaction networks, gene regulatory networks, and metabolic pathway networks, where the relationships between entities are as important as the entities themselves; column-family stores (such as Apache HBase and Google BigTable) are optimised for very large datasets where reads and writes involve many rows and few columns, useful for storing large-scale genomic data; the SPARQL query language and RDF (Resource Description Framework) data model are used by semantic web databases in biology to store and query biological knowledge in a structured, interoperable format.

Fill in the blank: _____ databases store data as nodes and relationships, making them well-suited to biological network data such as protein-protein interaction and gene regulatory networks.

4.2.3 Cloud storage and data repositories

Cloud storage systems provide scalable, durable, and accessible storage for large bioinformatics datasets without requiring local infrastructure; the major cloud object storage services used in bioinformatics are Amazon S3 (Simple Storage Service), Google Cloud Storage, and Microsoft Azure Blob Storage, all of which store arbitrary data files (objects) in named containers (buckets) and provide programmatic access through command-line tools (aws s3, gsutil) and software development kit (SDK) libraries for Python and other languages; many major bioinformatics databases now provide direct cloud-accessible copies of their data through AWS Open Data, Google Public Datasets, and similar programmes, enabling analysis without local download.

Specialised bioinformatics data repositories and analysis platforms built on cloud infrastructure include the NCBI SRA (Sequence Read Archive), which provides cloud-accessible storage and computational tools for raw sequencing data through the SRA Toolkit and the AWS and GCP

hosted SRA data; the European Genome-phenome Archive (EGA) for controlled-access human genomic and phenotypic data; and analysis platforms such as Terra (terra.bio, for collaborative genomics analysis using the WDL workflow language) and Galaxy (usegalaxy.org, an open-source web-based platform for bioinformatics analysis requiring no programming knowledge, particularly valuable for training in resource-limited settings such as many Nigerian institutions).

Fill in the blank: Galaxy (usegalaxy.org) is an open-source web-based bioinformatics analysis platform requiring no _____ knowledge, making it particularly accessible for training in resource-limited settings.

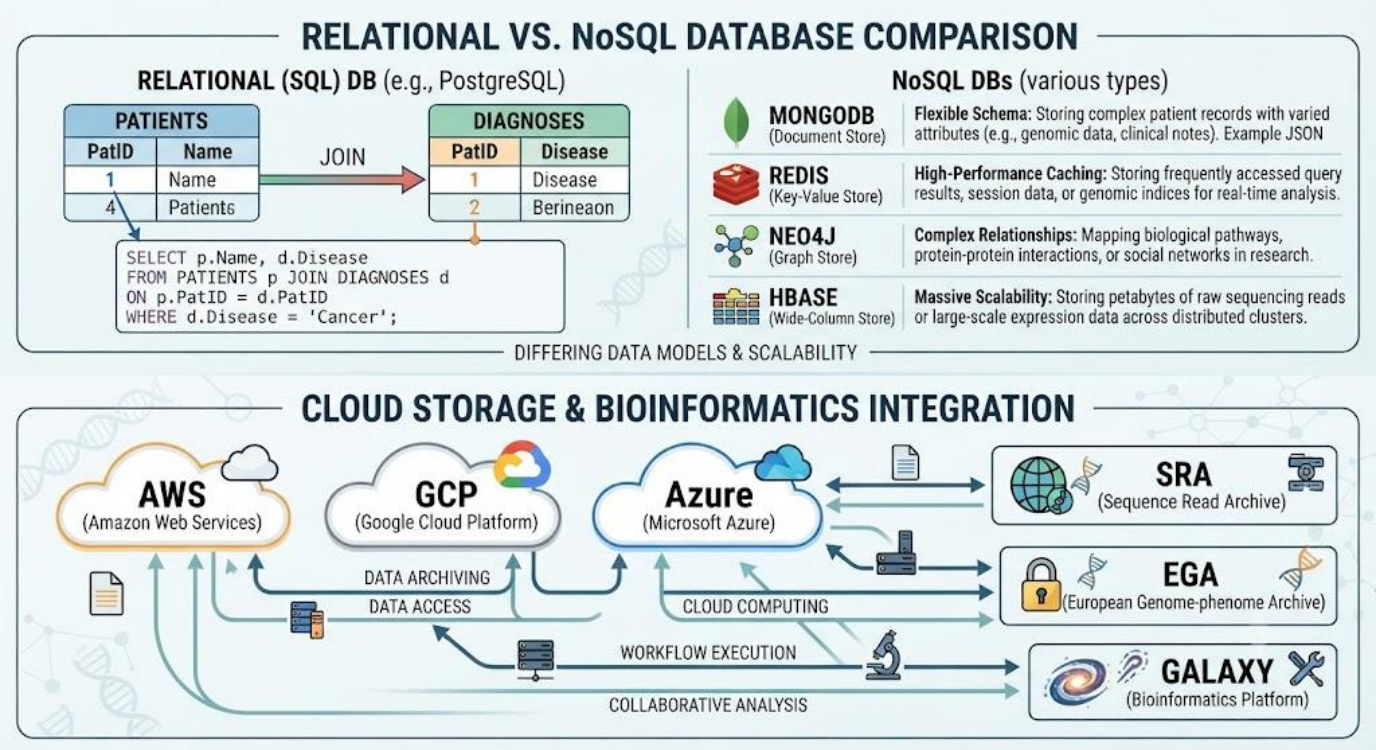

Figure 4.2: Data storage systems in bioinformatics

Pilot questions 4.2

1. Explain what a relational database is and state the language used to query it.
2. Write a simple SQL query to retrieve the gene name and length from a table called "genes" for all human genes longer than 500 base pairs.
3. Name four types of NoSQL database and give one biological application for each.
4. Explain what cloud object storage is and name three services used in bioinformatics.

5. Explain what Galaxy is and why it is valuable for bioinformatics training in Nigeria.

4.3 Sequence Alignment and Multiple Sequence Alignment

4.3.1 Pairwise sequence alignment

Pairwise sequence alignment is the process of arranging two biological sequences (nucleotide or amino acid) to identify regions of similarity that may reflect structural, functional, or evolutionary relationships; the two fundamental types of pairwise alignment are global alignment (aligning the full length of both sequences from end to end, appropriate when both sequences are of similar length and are expected to be homologous throughout, implemented by the Needleman-Wunsch dynamic programming algorithm) and local alignment (finding the highest-scoring region of similarity between two sequences, appropriate when sequences share only a local conserved region, implemented by the Smith-Waterman algorithm).

Alignment quality is assessed using a substitution matrix (a scoring system for comparing pairs of characters at each aligned position: for amino acids, the BLOSUM and PAM matrix families assign positive scores to conservative substitutions likely to preserve protein function, and negative scores to radical substitutions; BLOSUM62 is the most widely used matrix for protein alignment) and gap penalties (costs assigned for introducing insertions or deletions, typically a gap-open penalty for starting a gap and a lower gap-extension penalty for each additional position, rewarding compact gaps over multiple small gaps); the alignment score is the sum of substitution matrix scores and gap penalties over all aligned positions.

Fill in the blank: The _____ 62 substitution matrix is the most widely used scoring matrix for protein sequence alignment, assigning positive scores to conservative amino acid substitutions and negative scores to radical ones.

4.3.2 Multiple sequence alignment

Multiple sequence alignment (MSA) is the simultaneous alignment of three or more related sequences, producing a matrix in which each column represents a position believed to be homologous across all sequences in the alignment; MSA is fundamental to phylogenetic analysis (where the aligned positions define the characters used to infer evolutionary trees), identification of conserved functional residues (positions that are invariant across diverse homologs are likely to be structurally or functionally important), protein structure prediction (MSA conservation patterns inform secondary structure prediction and homology modelling), and the construction of sequence profiles (HMMs) for sensitive database searching.

The most widely used MSA programmes are ClustalW and its successor Clustal Omega (progressive alignment: first builds a guide tree by pairwise comparison of all sequences, then aligns sequences progressively from the most similar pairs outward to the most divergent), MUSCLE (iterative refinement: starts from a fast approximate alignment then iteratively improves it by realigning subsets of sequences), MAFFT (fast and accurate, particularly for large alignments, using Fourier transform-based methods for initial alignment), and T-Coffee (consistency-based: uses information from pairwise alignments to improve the final MSA); the choice of programme depends on the number of sequences, their divergence, and the available computing time.

Fill in the blank: Multiple sequence alignment is used to identify _____ residues across related proteins, which are often structurally or functionally important, and to construct sequence profiles for sensitive database searching.

4.3.3 Alignment quality and applications

MSA quality can be assessed using several approaches: visual inspection (checking that conserved residues are aligned in columns, that gap patterns are biologically plausible, and that highly divergent sequences have not been over-aligned); alignment scores and consistency measures (algorithms such as T-Coffee calculate a consistency score based on agreement between the MSA and pairwise alignments); and comparison with structural alignments (where

available, comparing the sequence alignment with a structure-based alignment of the same proteins to assess whether structurally equivalent positions are correctly aligned).

Applications of pairwise and multiple sequence alignment span the breadth of bioinformatics: BLAST database searching is based on seed-and-extend local alignment; phylogenetic tree inference requires an MSA as input; protein structure prediction by homology modelling requires accurate pairwise alignment between the query sequence and the template structure; gene prediction in newly sequenced genomes uses cross-species alignment to identify conserved coding regions; and the construction of protein family HMMs in Pfam uses MSA of known family members; the ability to produce high-quality alignments and to assess their quality is therefore a central bioinformatics skill.

Fill in the blank: Phylogenetic tree inference requires a _____ sequence alignment as input, since the aligned columns define the characters from which the evolutionary tree is reconstructed.

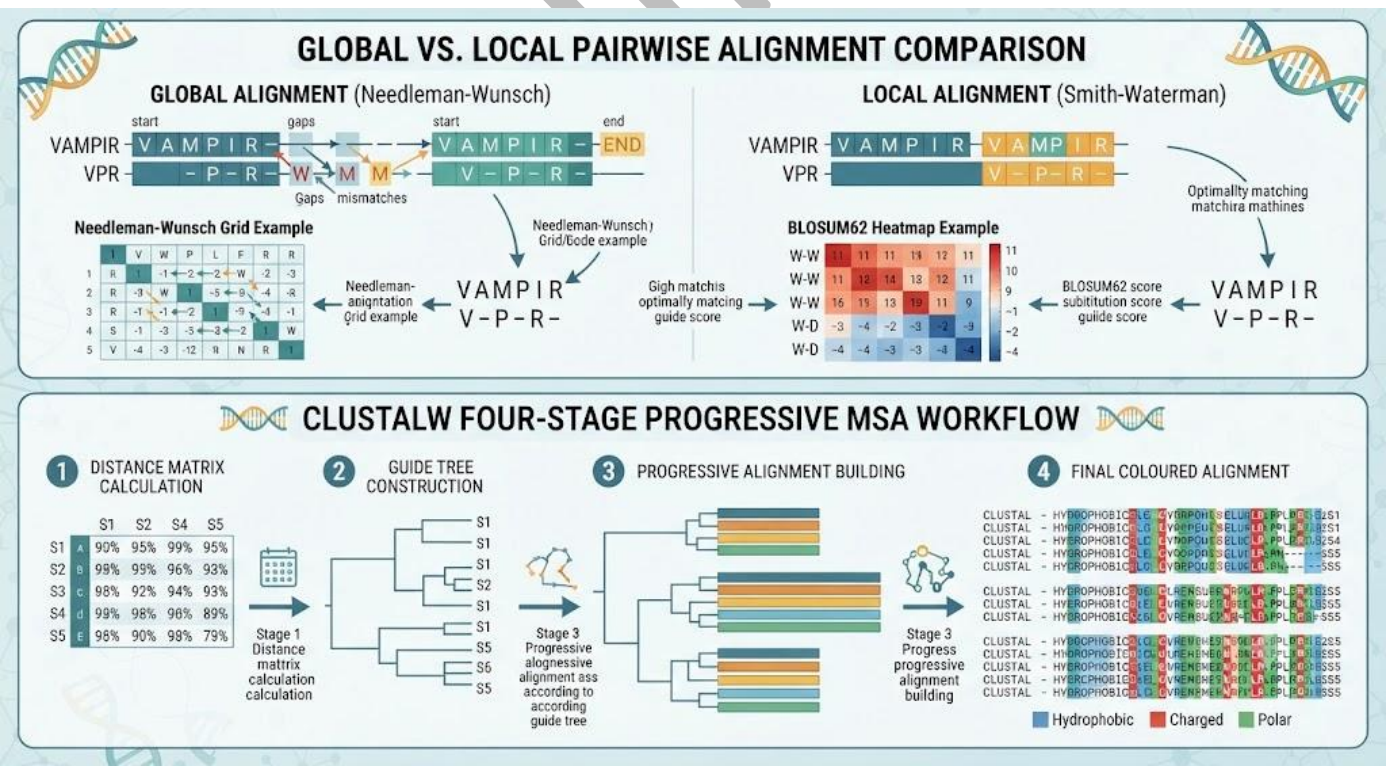

Figure 4.3: Sequence alignment methods

Pilot questions 4.3

1. Distinguish between global and local pairwise sequence alignment and name the algorithm for each.
2. Explain what a substitution matrix is and name two types used in protein alignment.
3. Explain what multiple sequence alignment is and state three applications.
4. Describe the progressive alignment strategy used by ClustalW.
5. Name three MSA programmes other than ClustalW and briefly state a strength of each.

4.4 Data Retrieval, Mining, and Integration

4.4.1 Programmatic data retrieval

Programmatic data retrieval is the automated extraction of data from databases and file systems using code rather than manual point-and-click operations; it is essential in bioinformatics because most analyses require data from multiple sources (sequence data from GenBank, annotation from UniProt, expression data from GEO, variant data from dbSNP) and because the volume of data required often makes manual retrieval impractical; the primary tools for programmatic retrieval from NCBI are the E-utilities API (accessible directly through HTTP requests or through the Biopython Entrez module in Python); for EBI resources, the EBI REST APIs provide programmatic access to UniProt, ENA, PDB, and InterPro data.

A typical programmatic retrieval workflow in Python using Biopython involves: importing the Entrez module and setting the user email address (required by NCBI to identify the request source); using Entrez.esearch() to search a database and retrieve a list of matching record IDs; using Entrez.efetch() with those IDs to retrieve the full records in the desired format (GenBank, FASTA, XML); parsing the returned records using Entrez.read() or SeqIO.parse() into Python data structures; and processing, filtering, or writing the extracted data to local files or a local database for downstream analysis; this workflow can retrieve hundreds or thousands of records automatically in the time it would take to retrieve one manually.

Fill in the blank: In the Biopython Entrez module, _____ .esearch() searches a database and returns a list of record IDs, which are then passed to Entrez.efetch() to retrieve the full records.

4.4.2 Data mining and knowledge extraction

Biological data mining is the application of computational methods to extract patterns, regularities, and biological insights from large biological datasets; common data mining tasks in bioinformatics include: clustering (grouping genes, proteins, or samples with similar expression or sequence properties into clusters using algorithms such as k-means, hierarchical clustering, or DBSCAN, to identify co-expressed gene modules or related protein families); classification (building predictive models that assign new sequences, variants, or samples to defined categories, such as predicting whether a sequence encodes a membrane protein or whether a variant is pathogenic); and pattern discovery (finding recurring sequence motifs, structural patterns, or functional modules in biological data).

Text mining of the biological literature (extracting structured biological facts, such as protein interactions, drug targets, and gene-disease associations, from the free text of published papers) is an increasingly important data mining approach, given that much biological knowledge is buried in the text of millions of published papers rather than in structured databases; tools such as PubTator, STRING, and DisGeNET use text mining of PubMed abstracts to build databases of biological relationships; machine learning approaches including deep learning are increasingly applied to biological data mining tasks including variant pathogenicity prediction, protein function annotation, and drug-target interaction prediction.

Fill in the blank: _____ mining of the biological literature extracts structured biological facts such as protein interactions and gene-disease associations from the free text of published papers.

4.4.3 Data integration and federated queries

Data integration is the process of combining data from multiple heterogeneous sources (databases, file formats, organisms, and experimental platforms) into a unified dataset for analysis; in bioinformatics, integrating data across databases is complicated by differences in identifiers (a gene may be referred to by different names or accession numbers in different

databases), data formats, annotation depth and quality, and the timeliness of database updates; identifier mapping (converting between database identifiers, for example from Ensembl gene IDs to UniProt accession numbers) is a frequent practical challenge, addressed by tools such as BioMart (biomart.org), UniProt ID mapping, and the g:Profiler suite.

Federated queries allow a single query to retrieve and integrate data from multiple databases simultaneously without physically copying all data to a single location; the BioMart federated query system (available through Ensembl and other databases) allows users to build complex queries that join data across genome annotation, expression, variation, and homology databases in a single request; the SPARQL query language for RDF (semantic web) biological databases enables federated queries across linked open biological data sources including WikiData, UniProt, Rhea, and the Open PHACTS discovery platform; as biological databases increasingly adopt linked data principles, federated querying is becoming a practical approach to large-scale biological data integration.

Fill in the blank: _____ (biomart.org) is a federated query system that allows users to join and retrieve data across genome annotation, expression, variation, and homology databases in a single request.

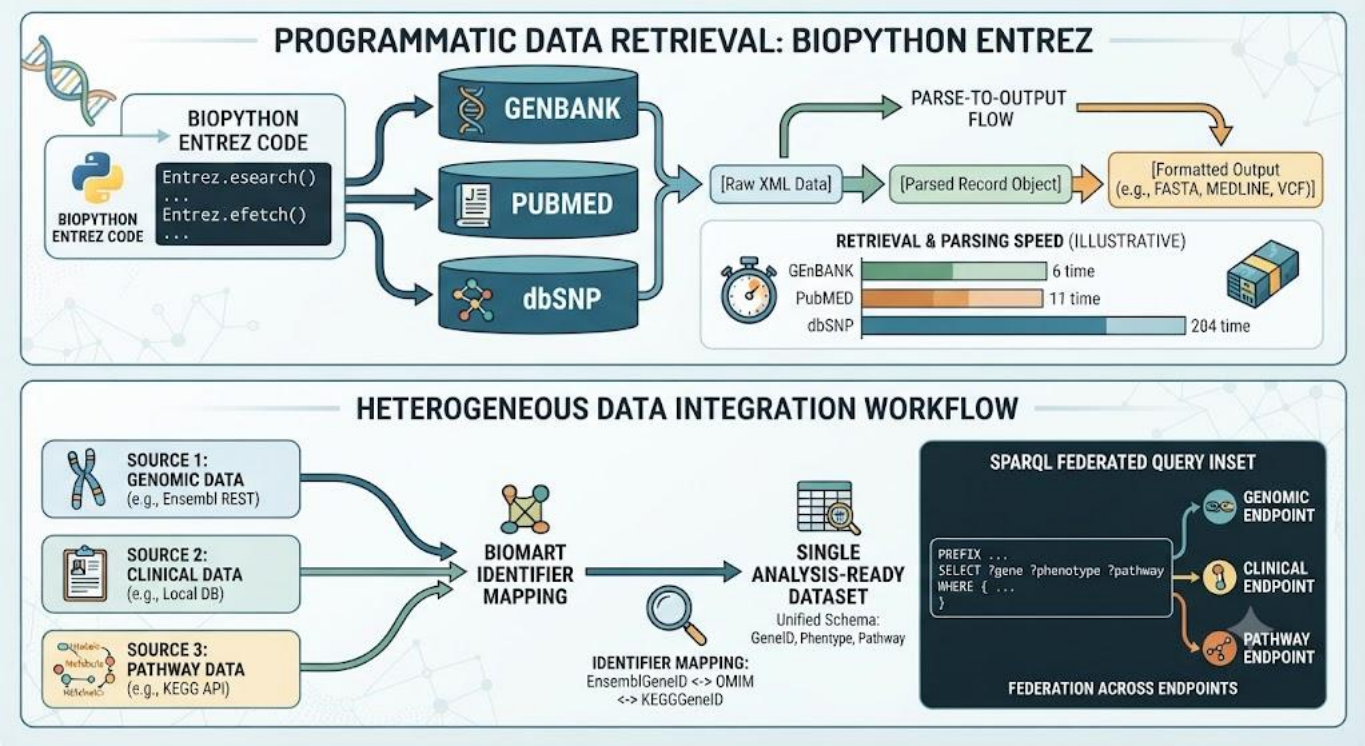

Figure 4.4: Data retrieval, mining, and integration

Pilot questions 4.4

1. Explain why programmatic data retrieval is essential in bioinformatics.
2. Describe the steps of a typical programmatic retrieval workflow using Biopython Entrez.
3. Name three data mining tasks used in bioinformatics and give one example of each.
4. Explain what text mining of the biological literature does and name one tool that uses it.
5. Explain what data integration is in bioinformatics and name one tool used for identifier mapping.

Series Summary

This Series covered the full chain from data storage format to retrieval and integration. Biological data exists in a rich ecology of specialised formats: FASTA and FASTQ for sequences, SAM/BAM for alignments, VCF for variants, PDB/mmCIF for structures, and count matrices for expression data, each designed for efficient storage and processing of a specific data type. At the storage level, relational databases and SQL handle structured biological metadata, while NoSQL databases accommodate network, document, and large-scale genomic data, and cloud platforms provide scalable storage and analysis infrastructure accessible globally.

Sequence alignment, both pairwise (using Needleman-Wunsch for global and Smith-Waterman for local alignment) and multiple (using ClustalW, MUSCLE, MAFFT, and T-Coffee), is the analytical foundation for database searching, phylogenetics, structure prediction, and functional annotation. Programmatic retrieval through APIs such as NCBI E-utilities and Biopython automates the extraction of large datasets efficiently. Data mining extracts patterns through clustering, classification, and text mining, while data integration through BioMart and federated SPARQL queries unifies heterogeneous data sources into analysis-ready datasets that can address complex biological questions.

Glossary of Terms

- **BAM format:** the compressed binary equivalent of the SAM (Sequence Alignment/Map) format, storing aligned short reads with their reference positions, quality scores, and CIGAR alignment descriptions.
- **BioMart:** a federated biological database query system that retrieves and joins data across genome annotation, expression, variation, and homology databases in a single query.
- **BLOSUM matrix:** a family of substitution scoring matrices derived from alignments of conserved protein blocks; BLOSUM62 is the most widely used for protein sequence alignment.
- **CIGAR string:** a compact notation in SAM format describing how a sequencing read aligns to a reference genome, encoding matches, mismatches, insertions, deletions, and clipped bases.

- **Count matrix:** a gene-by-sample table of read counts produced by RNA-seq analysis, used as input for differential expression analysis.
- **Global alignment:** a pairwise alignment that extends the full length of both sequences from end to end, implemented by the Needleman-Wunsch dynamic programming algorithm.
- **Local alignment:** a pairwise alignment that identifies the highest-scoring region of similarity between two sequences, implemented by the Smith-Waterman algorithm.
- **Multiple sequence alignment (MSA):** the simultaneous alignment of three or more sequences into a matrix where each column represents a homologous position across all sequences.
- **NoSQL database:** a non-relational database management system storing data in document, key-value, graph, or column-family formats, suited to heterogeneous or network biological data.
- **VCF (Variant Call Format):** the standard format for storing genomic variants identified from sequencing data, with fields for chromosome, position, reference and alternative alleles, quality, and genotype calls.

Concept Check Questions [Series 4]

Objective Questions

1. The FASTQ format extends FASTA by adding per-base _____ score information encoded as ASCII Phred values. (Linked to SLO 4.1)
2. The VCF (Variant _____ Format) is the standard format for storing genomic variants identified from sequencing data. (Linked to SLO 4.1)
3. _____ is the standard language for querying relational databases, using SELECT, WHERE, and JOIN commands. (Linked to SLO 4.2)
4. The _____ 62 substitution matrix is the most widely used scoring matrix for protein pairwise sequence alignment. (Linked to SLO 4.3)
5. _____ (biomart.org) is a federated query system for joining data across genome annotation, expression, and variation databases. (Linked to SLO 4.4)

Short-Answer Questions

1. Distinguish between FASTA and FASTQ formats. (Linked to SLO 4.1)
2. Explain what a CIGAR string in a SAM file represents. (Linked to SLO 4.1)
3. Distinguish between global and local pairwise sequence alignment. (Linked to SLO 4.3)
4. Describe the progressive alignment strategy used by ClustalW. (Linked to SLO 4.3)
5. Explain what data integration is in bioinformatics and name one identifier mapping tool. (Linked to SLO 4.4)

Long-Answer Questions

1. Describe the major biological data file formats and the data storage systems used in bioinformatics. (Linked to SLOs 4.1, 4.2)
2. Describe pairwise and multiple sequence alignment and explain programmatic data retrieval, mining, and integration. (Linked to SLOs 4.3, 4.4)

Answers to Concept Check Questions [Series 4]

Objective Questions

1. Quality.
2. Call.
3. SQL.
4. BLOSUM.
5. BioMart.

Short-Answer Questions

1. FASTA stores nucleotide or amino acid sequences with a header line starting with ">" followed by sequence; FASTQ extends this with a per-base quality score string (ASCII Phred encoded), making it the standard for raw NGS reads.
2. A CIGAR string is a compact notation in SAM format describing the alignment of a read to the reference: M (match), I (insertion relative to reference), D (deletion from reference), S (soft clip); e.g. "5M2I3M" means 5 matches, 2 insertions, 3 matches.
3. Global alignment (Needleman-Wunsch) aligns the full length of both sequences end-to-end; local alignment (Smith-Waterman) identifies the highest-scoring region of similarity between two sequences without requiring alignment of their full lengths.
4. ClustalW first calculates all pairwise distances between sequences, then builds a guide tree by neighbour-joining, then aligns sequences progressively starting with the most similar pairs and adding progressively more divergent sequences following the guide tree order.
5. Data integration combines data from multiple heterogeneous sources (databases, formats, identifiers) into a unified dataset; BioMart is a federated query system and identifier mapping tool for converting between Ensembl, UniProt, and other database identifiers.

Long-Answer Questions

1. File formats: FASTA (sequence), FASTQ (sequence + quality, raw NGS reads), GenBank flat file (annotated sequence), SAM/BAM (aligned reads, CIGAR string, read position), VCF/BCF (variants, chromosome/position/REF/ALT/genotype), BED (genomic intervals), PDB/mmCIF (3D atomic coordinates), count matrix (gene expression RNA-seq, tab-delimited or HDF5). Storage: relational databases (SQL, tables, JOINs; GenBank Oracle, Ensembl MySQL, local

SQLite); NoSQL (document/MongoDB for heterogeneous records, graph/Neo4j for networks, key-value/Redis for fast lookup, column-family/HBase for large-scale genomic); cloud (AWS S3, GCP, Azure; SRA, EGA, Galaxy for training).

2. Pairwise alignment: global (Needleman-Wunsch, full length, both sequences similar) vs local (Smith-Waterman, highest-scoring local region); substitution matrices (BLOSUM62 for protein, positive for conservative substitutions) and gap penalties (open and extension). MSA (ClustalW: guide tree + progressive; MUSCLE: iterative; MAFFT: fast large datasets; T-Coffee: consistency); applications: phylogenetics, conserved residue identification, HMM construction. Programmatic retrieval: Biopython Entrez (esearch+efetch+parse) for NCBI; EBI REST APIs. Data mining: clustering, classification, text mining (PubTator, STRING). Integration: BioMart, identifier mapping, SPARQL federated queries.

Answers to Pilot Questions [Series 4]

Part 4.1

1. FASTA stores sequences with a ">" header and sequence lines, no quality information; FASTQ adds a per-base quality score string (ASCII Phred) as the fourth line of each four-line record; FASTA is used for reference sequences and protein databases, FASTQ for raw NGS reads.
2. A CIGAR string describes how a sequencing read aligns to a reference: M = match/mismatch, I = insertion (relative to reference), D = deletion (from reference), S = soft-clipped (sequence present in read but not in alignment); example "3M2I5M" means 3 matched positions, 2 inserted bases, 5 more matched positions.
3. VCF stores genomic variants (SNPs, indels, structural variants) with one variant per line; three fields: CHROM (chromosome), POS (position), REF (reference allele), ALT (alternative allele), QUAL (quality score).
4. The PDB format stores three-dimensional atomic coordinates in fixed-column-width text; ATOM records contain element type, atom name, residue name, chain identifier, residue number, and x, y, z Cartesian coordinates in Angstroms for each atom in the structure.
5. A gene expression count matrix is a table with genes as rows and samples as columns, where each cell contains the number of sequencing reads mapped to that gene in that sample; typically stored as tab-delimited text (TSV) or HDF5 binary format.

Part 4.2

1. A relational database organises data into tables with rows and columns, with relationships expressed through shared key columns; queried using SQL (Structured Query Language).
2. SELECT gene_name, length FROM genes WHERE organism = "Homo sapiens" AND length > 500;
3. Document (MongoDB): storing variable-format variant annotation records; Key-value (Redis): fast lookup of protein sequences by accession; Graph (Neo4j): protein-protein interaction network storage; Column-family (HBase): large-scale genomic data with many rows.
4. Cloud object storage stores arbitrary data files (objects) in named containers (buckets) accessible through the internet; three services: Amazon S3, Google Cloud Storage, and Microsoft Azure Blob Storage.
5. Galaxy is an open-source web-based bioinformatics analysis platform requiring no programming knowledge; it is valuable for Nigerian training because it allows students to run standard analyses on institutional or public servers without local computing resources or programming expertise.

Part 4.3

1. Global alignment (Needleman-Wunsch) aligns the complete length of both sequences end-to-end, appropriate when sequences are expected to be homologous throughout; local alignment (Smith-Waterman) finds the highest-scoring local region of similarity between two sequences, appropriate when only part of each sequence is conserved.
2. A substitution matrix assigns scores to all pairs of aligned characters based on the likelihood or frequency of that substitution in biologically related sequences; protein types: BLOSUM matrices (derived from blocks of conserved sequence) and PAM matrices (derived from phylogenetically estimated substitution frequencies).
3. MSA simultaneously aligns three or more sequences into a matrix where each column represents a homologous position; three applications: phylogenetic tree inference (aligned columns are the input characters), identification of conserved functional residues (invariant positions across diverse homologs), and construction of HMMs for sensitive database searching.
4. ClustalW progressive alignment: (1) calculate all pairwise distances; (2) build a guide tree by neighbour-joining; (3) align the most similar sequence pair first; (4) progressively add more divergent sequences following the guide tree order; (5) final MSA output.
5. MUSCLE (iterative refinement: fast initial alignment then iterative improvement, good for moderate number of sequences), MAFFT (very fast and accurate, FFT-based, suited for large

alignments), T-Coffee (consistency-based using pairwise alignments, higher accuracy for divergent sequences).

Part 4.4

1. Programmatic retrieval is essential because most analyses require data from multiple sources in large volumes that make manual retrieval impractical; automated retrieval can fetch hundreds or thousands of records in the time manual retrieval would take for one.

(1) Import Entrez module and set user email; (2) use Entrez.esearch() to search database and retrieve matching record IDs; (3) use Entrez.efetch() to retrieve full records in desired format; (4) parse with Entrez.read() or SeqIO.parse(); (5) process and write outputs to local files or database.

2. Clustering (grouping co-expressed genes by k-means or hierarchical clustering to find gene modules), classification (predicting variant pathogenicity using machine learning on variant features), and text mining (extracting protein interactions from PubMed abstracts using PubTator).
3. Text mining extracts structured biological facts from the free text of published papers (protein interactions, gene-disease associations, drug targets); tool example: PubTator (NCBI) or STRING (which uses text mining to build protein interaction networks from PubMed).
4. Data integration combines data from multiple heterogeneous sources with different formats and identifiers into a unified dataset; BioMart (biomart.org) is a widely used tool for converting between Ensembl gene IDs, UniProt accessions, gene symbols, and other database identifiers.

References and Attributions

Textual References (Books and External Sources)

- Lesk, A. M. (2019). Introduction to bioinformatics (4th ed.). Oxford University Press.
- Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press.
- Mount, D. W. (2004). Bioinformatics: Sequence and genome analysis (2nd ed.). Cold Spring Harbor Laboratory Press.
- National Open University of Nigeria (NOUN). (2023). Bioinformatics course material. NOUN Press.

Figures, Plates, Tables, and Boxes

- Figure 4.1: Biological data file formats Attribution: AI generated. Suggested OER search string: "FASTA FASTQ SAM BAM VCF BED PDB gene expression count matrix bioinformatics file formats OER".
- Figure 4.2: Data storage systems in bioinformatics Attribution: AI generated. Suggested OER search string: "SQL relational database NoSQL MongoDB Neo4j cloud storage AWS Galaxy bioinformatics data management OER".
- Figure 4.3: Sequence alignment methods Attribution: AI generated. Suggested OER search string: "sequence alignment global local Needleman Wunsch Smith Waterman BLOSUM multiple sequence alignment ClustalW MSA OER".
- Figure 4.4: Data retrieval, mining, and integration Attribution: AI generated. Suggested OER search string: "programmatic data retrieval Biopython Entrez data integration BioMart SPARQL data mining bioinformatics OER".

Course code: BIO 413

Course title: Bioinformatics

Course credit load: 2 Units

Series 5: Genomics and Proteomics

- **Part 5.1: Principles of Genomics**
    - Sub Part 5.1.1: The genome and genomic organisation
    - Sub Part 5.1.2: Genome sequencing and assembly
    - Sub Part 5.1.3: Genome annotation
- **Part 5.2: Functional Genomics and Transcriptomics**
    - Sub Part 5.2.1: RNA-seq and transcriptomics
    - Sub Part 5.2.2: Differential expression analysis
    - Sub Part 5.2.3: Epigenomics and single-cell genomics
- **Part 5.3: Comparative Genomics and Phylogenomics**
    - Sub Part 5.3.1: Comparative genomics
    - Sub Part 5.3.2: Phylogenomics
    - Sub Part 5.3.3: Population genomics
- **Part 5.4: Proteomics and Systems Biology**
    - Sub Part 5.4.1: Proteomics: approaches and data
    - Sub Part 5.4.2: Protein structure prediction and function
    - Sub Part 5.4.3: Systems biology and network analysis

Introduction

Genomics and proteomics represent the highest-level integration of bioinformatics, biology, and technology, applying the databases, algorithms, and computing infrastructure described in earlier Series to address questions about the complete genetic content of organisms and the total complement of proteins they produce. These disciplines have transformed biology from a gene-by-gene science to one capable of analysing entire biological systems, revealing the molecular basis of evolution, disease, and cellular function at an unprecedented scale.

This final Series covers the principles of genomics from genome organisation through sequencing and assembly to annotation; functional genomics through RNA-seq transcriptomics, differential expression analysis, epigenomics, and single-cell genomics; comparative genomics and phylogenomics for understanding evolutionary relationships at the genome scale and population genomics for studying genetic variation; and proteomics and systems biology, completing the integration of bioinformatics across all levels of biological organisation from gene to network.

Series Learning Outcomes

Upon completing this Series, you should be able to:

- **SLO 5.1** describe genomic organisation, genome sequencing and assembly, and genome annotation,
- **SLO 5.2** describe RNA-seq transcriptomics, differential expression analysis, epigenomics, and single-cell genomics,
- **SLO 5.3** explain comparative genomics, phylogenomics, and population genomics, and
- **SLO 5.4** describe proteomics approaches and explain systems biology and biological network analysis.

5.1 Principles of Genomics

5.1.1 The genome and genomic organisation

The genome of an organism is the complete set of its genetic material, comprising all its chromosomes and the sequences they contain, including both coding sequences (genes encoding proteins or functional RNAs) and non-coding sequences (introns, regulatory elements, repetitive sequences, and large amounts of sequence with unknown or no function); the size and organisation of genomes vary enormously across life: the human genome is approximately 3.2 billion base pairs (3.2 Gb) and contains approximately 20,000 protein-coding genes (which constitute only about 1.5 percent of the total sequence), while bacterial genomes are typically 1 to 10 Mb and are much more compact with a higher proportion of coding sequence.

Eukaryotic genomes are organised on multiple linear chromosomes within the nucleus and are packaged into chromatin (DNA wrapped around histone octamers to form nucleosomes, further folded into higher-order structures); repetitive sequences constitute a large proportion of eukaryotic genomes: transposable elements (mobile genetic elements that can copy or move themselves to new genomic locations) make up approximately 45 percent of the human genome; short tandem repeats (STRs) and microsatellites (tandem arrays of short repeat units of 2 to 6 bp) are important markers in forensic genetics and population studies; the functional organisation of the genome (which sequences are expressed in which tissues and developmental stages, and how gene expression is regulated by the chromatin environment) is the subject of functional genomics.

Fill in the blank: Transposable elements constitute approximately _____ percent of the human genome and are one major class of repetitive sequence that makes up the non-coding majority of eukaryotic genomes.

5.1.2 Genome sequencing and assembly

Genome sequencing is the process of determining the complete nucleotide sequence of an organism genome, which for large eukaryotic genomes requires assembling millions to billions of short sequencing reads into a contiguous sequence; two major assembly strategies are used:

reference-guided assembly (aligning reads to an existing reference genome sequence, appropriate when a closely related reference is available) and de novo assembly (assembling reads into contigs and scaffolds from scratch, required for new organisms without a close reference and for detecting large structural variants not represented in the reference); de novo assembly is computationally intensive and algorithmically challenging, particularly in regions of repetitive sequence where individual reads cannot be uniquely placed.

De novo genome assembly proceeds through several stages: read overlap detection (identifying reads that share sequence, indicating they originate from overlapping regions of the genome), graph construction (building a graph representation of the overlapping reads, such as an overlap-layout-consensus graph or a de Bruijn graph), contig assembly (resolving the graph into linear contiguous sequences by following paths through it), scaffolding (ordering and orienting contigs into scaffolds using long-range information from mate-pair reads, long reads, or optical mapping), and gap filling (sequencing across gaps between contigs within scaffolds); assembly quality is assessed by statistics including N50 (the length such that half of all assembled sequence is in contigs of this length or longer, with higher N50 indicating a more contiguous assembly).

Fill in the blank: The N50 statistic is the length such that half of all assembled sequence is contained in contigs of this length or _____ , with a higher value indicating a more contiguous and complete genome assembly.

5.1.3 Genome annotation

Genome annotation is the process of identifying and describing the functional elements within a genome sequence, including protein-coding genes, non-coding RNA genes, regulatory elements, and repetitive sequences; structural annotation (identifying the location and boundaries of genes: intron and exon positions, start and stop codons, splice sites) uses a combination of ab initio gene prediction (computational models of gene structure trained on known genes, such as AUGUSTUS and SNAP) and evidence-based approaches (using RNA-seq data showing expressed transcripts, protein sequence evidence from related organisms, and homology to known genes in other organisms to validate and refine gene predictions).

Functional annotation (assigning biological functions to the predicted genes) uses BLAST similarity searching to identify homologs in characterised databases, InterPro domain scanning to identify protein family and domain membership, GO term assignment to describe molecular function, biological process, and cellular component, and KEGG pathway mapping to place genes in metabolic and signalling contexts; annotation pipelines for prokaryotic genomes (such as Prokka and RAST) are substantially simpler than for eukaryotes, reflecting the simpler gene structures (no introns) and more compact genomic organisation of bacteria; the quality of genome annotation is critical because downstream analyses including transcriptomics and comparative genomics depend entirely on the accuracy of the gene models.

Fill in the blank: _____ annotation identifies the location and boundaries of genes (introns, exons, splice sites), while functional annotation assigns biological functions to the predicted gene products.

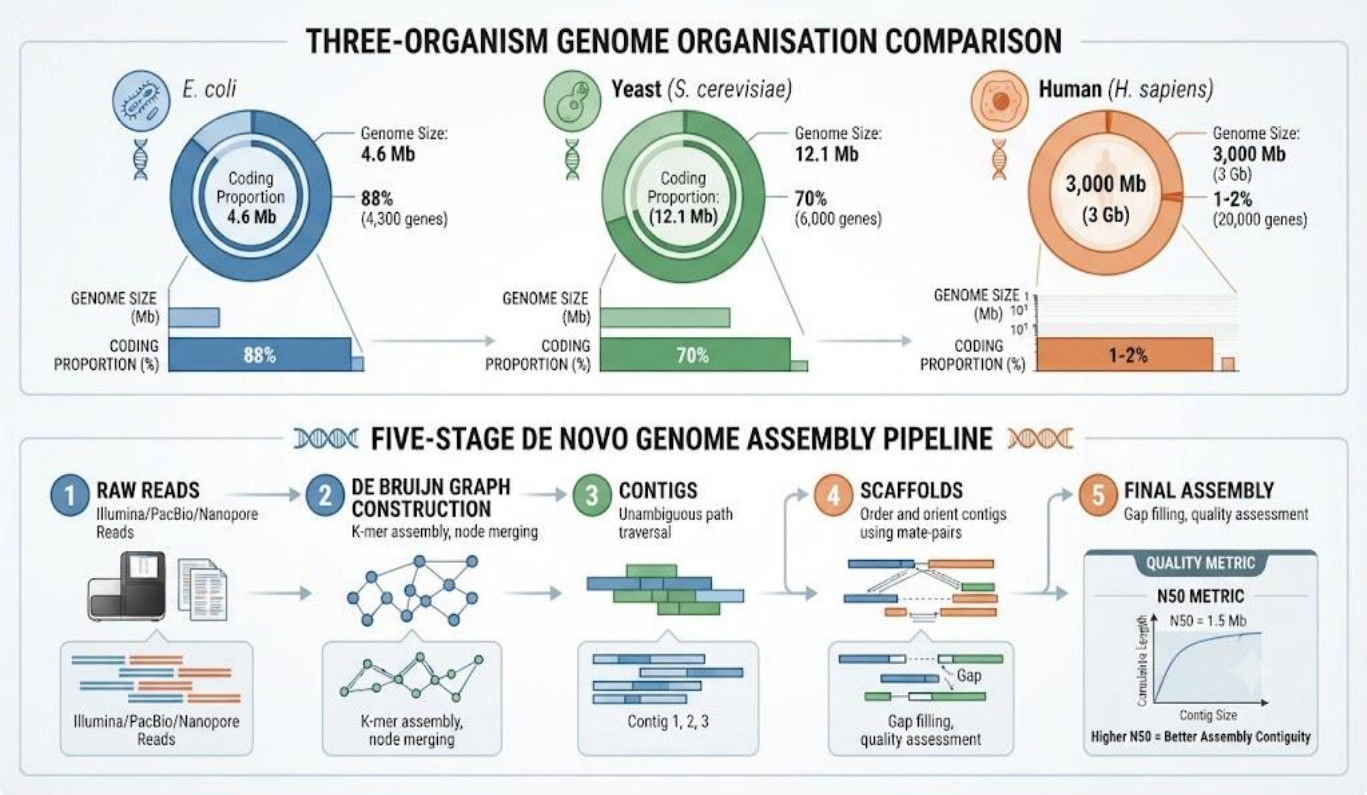

Figure 5.1: Genomics principles

Pilot questions 5.1

1. State the size of the human genome in base pairs and the approximate number of protein-coding genes it contains.
2. Explain what transposable elements are and what proportion of the human genome they constitute.
3. Distinguish between reference-guided and de novo genome assembly.
4. Explain what the N50 statistic measures in genome assembly.
5. Distinguish between structural and functional genome annotation.

5.2 Functional Genomics and Transcriptomics

5.2.1 RNA-seq and transcriptomics

RNA sequencing (RNA-seq) is the application of next-generation sequencing to the transcriptome (the complete set of RNA transcripts produced by a cell, tissue, or organism under defined conditions), enabling quantitative measurement of the expression level of all expressed genes simultaneously; an RNA-seq experiment involves extracting total RNA or poly(A)-selected mRNA from a biological sample, converting it to complementary DNA (cDNA), preparing a sequencing library, and sequencing on an NGS platform (typically Illumina) to generate millions of reads; the reads are then aligned to a reference genome or transcriptome using a splice-aware aligner (such as STAR or HISAT2) and the number of reads mapping to each gene is counted to produce a count matrix.

RNA-seq has largely replaced microarrays as the standard platform for gene expression profiling because it provides higher dynamic range (can detect both very low and very high expression levels), does not require prior knowledge of gene sequences, can detect novel transcripts, splicing variants, and fusion genes not represented in the array design, and produces digital count data that can be compared directly across experiments; long-read RNA-seq (using Oxford Nanopore or PacBio platforms) directly sequences full-length mRNA molecules, resolving alternative splicing patterns without the ambiguity inherent in short-read data; the raw count matrix output of RNA-seq serves as the input for differential expression analysis.

Fill in the blank: RNA-seq reads are aligned to a reference genome using a _____ -aware aligner such as STAR or HISAT2 that correctly handles reads spanning exon-exon junctions.

5.2.2 Differential expression analysis

Differential expression (DE) analysis identifies genes that are significantly more or less expressed in one condition (for example, a diseased tissue) compared to another (healthy tissue), after accounting for the natural variability between biological replicates; the standard pipeline for count-based DE analysis in R uses the DESeq2 or edgeR package: raw counts are normalised to account for differences in sequencing depth and RNA composition between samples; the normalised counts are modelled using a negative binomial distribution (chosen because RNA-seq count data typically shows overdispersion, with variance greater than the mean, inconsistent with the simpler Poisson distribution); statistical tests are applied to identify genes with significantly different expression.

DE results are reported as a table of genes with their fold change (the ratio of expression in one condition versus the other, typically expressed on a log2 scale so that a log2 fold change of 1 represents a doubling of expression) and their adjusted p-value (the probability of observing the detected fold change by chance, corrected for multiple testing using the Benjamini-Hochberg false discovery rate correction, since tens of thousands of genes are tested simultaneously); results are typically visualised using a volcano plot (fold change on the x-axis, negative log10 adjusted p-value on the y-axis, with significantly differentially expressed genes highlighted) and a heatmap (showing the expression pattern of the top DE genes across all samples).

Fill in the blank: DE analysis results report log2 fold change and _____ -adjusted p-values (using Benjamini-Hochberg false discovery rate correction) to identify significantly differentially expressed genes.

5.2.3 Epigenomics and single-cell genomics

Epigenomics is the genome-wide study of epigenetic modifications: chemical modifications of DNA (primarily cytosine methylation at CpG dinucleotides, which generally represses gene expression) and histone proteins (acetylation, methylation, phosphorylation, and other

modifications that regulate chromatin accessibility and gene expression), and the three-dimensional organisation of chromosomes within the nucleus (studied by Hi-C chromosome conformation capture sequencing); ENCODE (Encyclopedia of DNA Elements) and the Roadmap Epigenomics project have systematically mapped these epigenetic features across hundreds of human cell types, providing a comprehensive view of genome regulation.

Single-cell RNA sequencing (scRNA-seq) measures gene expression in individual cells rather than in bulk cell populations, revealing the cellular heterogeneity hidden in bulk RNA-seq averages; scRNA-seq workflows involve dissociating tissue into single cells, capturing individual cells into droplets or wells with a unique barcode sequence (using platforms such as 10x Genomics Chromium), reverse-transcribing each cell mRNA with the cell barcode and a unique molecular identifier (UMI) for deduplication, sequencing the barcoded library, and computationally demultiplexing reads to reconstruct the expression profile of each individual cell; downstream analysis (using packages such as Seurat or Scanpy) clusters cells by expression similarity and identifies cell-type-specific marker genes.

Fill in the blank: Single-cell RNA sequencing uses _____ sequences added to each cell mRNA during library preparation to assign reads to their cell of origin during computational demultiplexing.

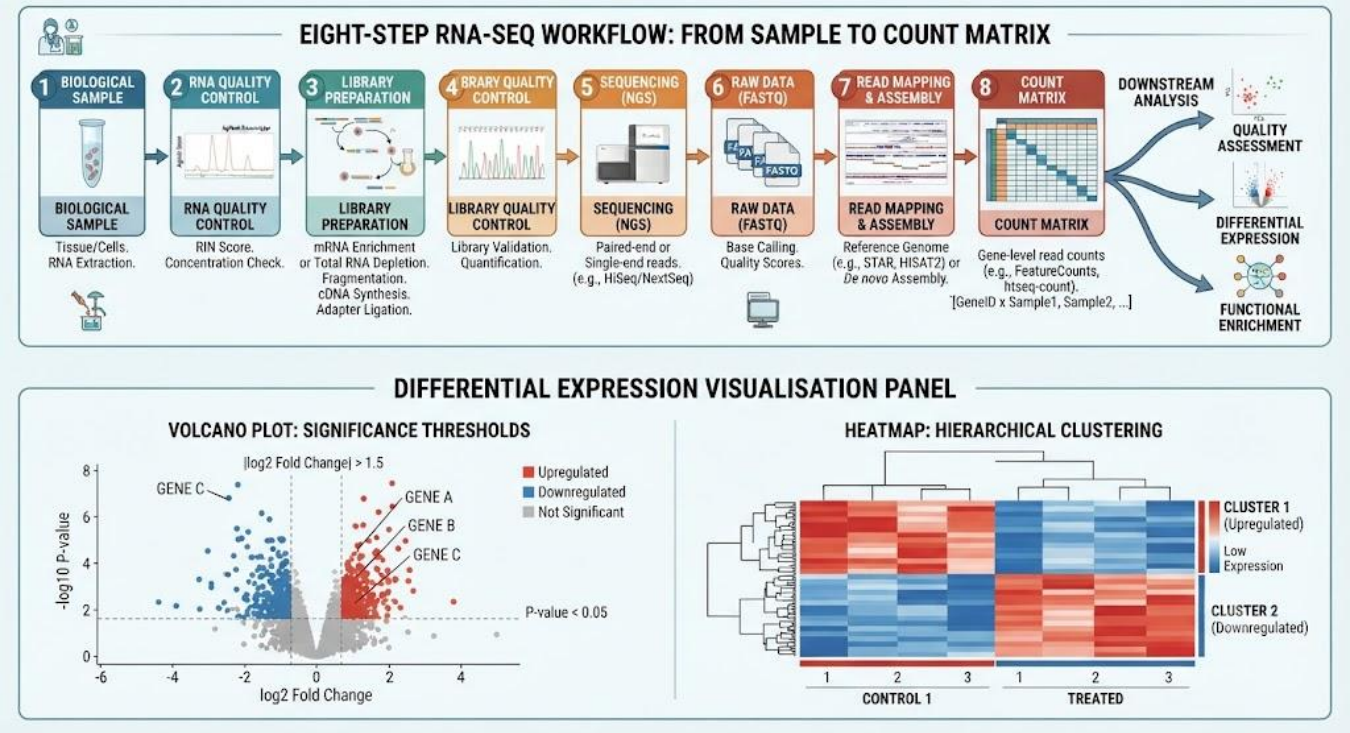

Figure 5.2: Functional genomics and transcriptomics

Pilot questions 5.2

1. Describe the steps of an RNA-seq experiment from sample to count matrix.
2. State two advantages of RNA-seq over microarrays for gene expression profiling.
3. Explain what log2 fold change means in differential expression analysis.
4. Explain what a volcano plot shows and how to interpret it.
5. Explain what single-cell RNA-seq reveals that bulk RNA-seq cannot.

5.3 Comparative Genomics and Phylogenomics

5.3.1 Comparative genomics

Comparative genomics is the study of similarities and differences in genome structure, content, and organisation between different species, with the goal of understanding genome evolution and identifying functionally important elements; whole-genome alignment (aligning the complete sequences of two or more genomes to identify syntenic regions, conserved sequences, and

structural rearrangements) is performed using tools such as MUMmer (for closely related genomes, using maximal unique match seeds) and LASTZ (for more divergent genomes, using gapped alignment); the identification of conserved non-coding sequences (sequences that are similar between species despite not encoding proteins) is a powerful approach for discovering functional regulatory elements.

Orthologs (genes in different species that diverged by speciation and typically retain the same function, as opposed to paralogs which diverged by gene duplication and may have different functions) are identified by comparative genomics tools such as OrthoFinder and the OrthoMCL algorithm, and the resulting ortholog groups are used to transfer functional annotation from well-characterised model organisms to newly sequenced genomes; gene family expansion and contraction (changes in the number of genes belonging to a given family across species) can be detected and analysed using comparative genomics, revealing adaptive evolution of gene repertoires in response to ecological and physiological pressures.

Fill in the blank: _____ are genes in different species that diverged by speciation and typically retain the same biological function, distinguished from paralogs which diverged by gene duplication.

5.3.2 Phylogenomics

Phylogenomics is the inference of evolutionary relationships among organisms using genome-scale sequence data, providing substantially more statistical power and resolution than phylogenetics based on one or a few genes, because many independent loci provide independent evidence for the same evolutionary relationship; phylogenomic datasets are constructed by identifying orthologous genes shared across all taxa in the analysis, aligning each gene separately, and then combining the alignments either by concatenation (treating the combined alignment as a single super-matrix and inferring a single tree) or by coalescent-based methods (inferring individual gene trees and then combining them to estimate the species tree, accounting for the discordance between gene trees caused by incomplete lineage sorting).

Phylogenomic analysis has resolved many previously uncertain evolutionary relationships, including the placement of basal animal lineages, the relationships among placental mammal

orders, and the deep phylogeny of bacteria and archaea; the tools most widely used for phylogenomic tree inference include IQ-TREE (a maximum likelihood tree inference programme that implements model testing and ultrafast bootstrap support estimation), RAxML (another widely used maximum likelihood tool), and BEAST (Bayesian evolutionary analysis for time-calibrated trees estimating divergence times); phylogenomic trees are used to date evolutionary events, identify horizontal gene transfer, and reconstruct ancestral genomes.

Fill in the blank: In phylogenomics, the _____ approach to combining gene data concatenates multiple ortholog alignments into a super-matrix and infers a single tree, while coalescent methods account for discordance among individual gene trees.

5.3.3 Population genomics

Population genomics is the application of genome-wide sequence data to study genetic variation within and between populations, addressing questions including the genetic basis of adaptation, the history of population size changes (demographic history), patterns of gene flow and admixture between populations, and the identification of loci under natural selection; whole-genome resequencing (sequencing the genomes of many individuals from one or more populations and comparing the variants they carry) enables population genomic analysis at the highest resolution, while SNP genotyping arrays provide a cost-effective alternative for large sample sizes.

Key population genomic statistics and methods include: allele frequency spectrum analysis (examining the distribution of variant frequencies across the genome to infer demographic history and selection); Fst (a measure of genetic differentiation between populations, ranging from 0 for identical populations to 1 for completely differentiated populations); linkage disequilibrium (LD) analysis (examining the correlation of allele frequencies between nearby SNPs, which reflects recombination history and identifies haplotype blocks); and genome-wide association studies (GWAS, which test each of hundreds of thousands of SNPs for statistical association with a disease or trait in a large population sample, identifying genomic regions containing causal variants); tools widely used in population genomics include PLINK, ADMIXTURE, TreeMix, and the GATK variant calling pipeline.

Fill in the blank: Fst is a measure of genetic _____ between populations, ranging from 0 (identical allele frequencies) to 1 (completely different allele frequencies).

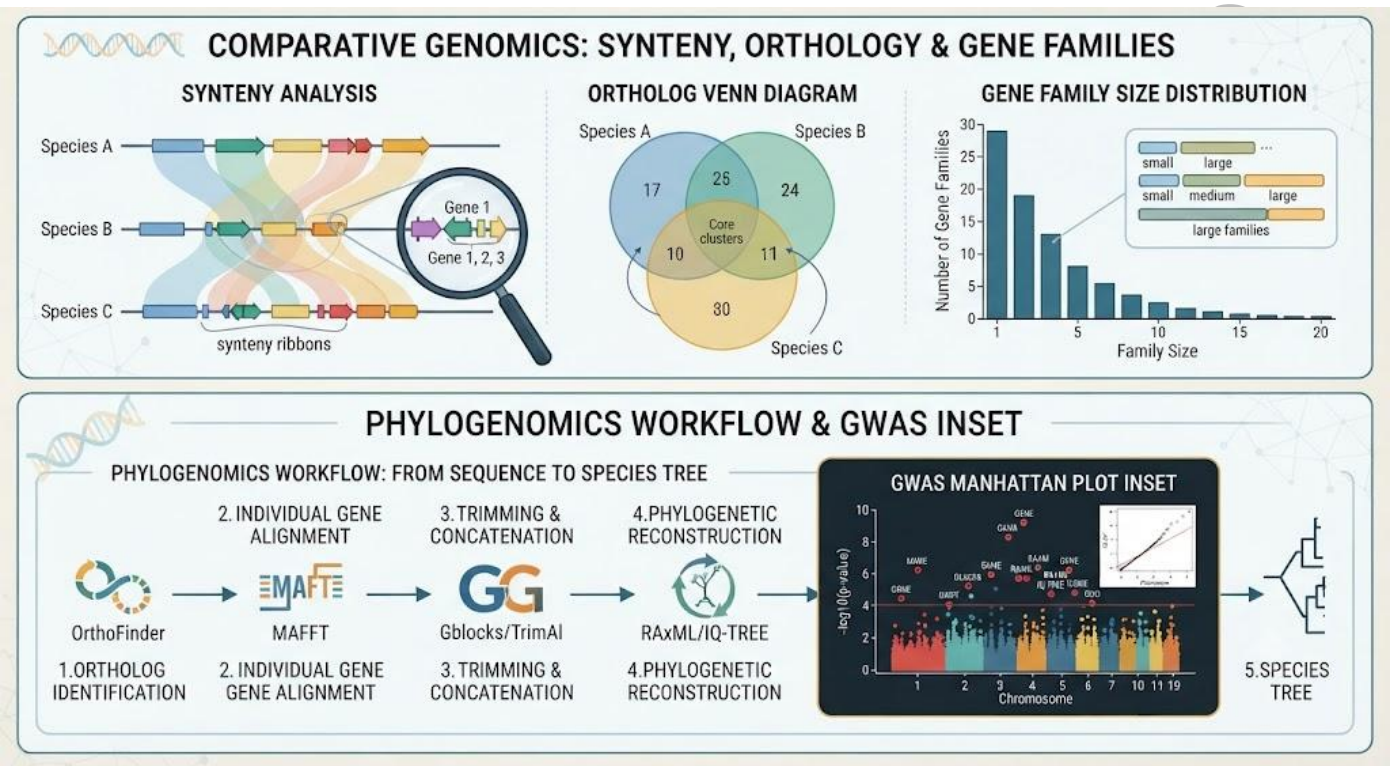

Figure 5.3: Comparative genomics and phylogenomics

Pilot questions 5.3

1. Explain what comparative genomics is and name one tool used for whole-genome alignment.
2. Distinguish between orthologs and paralogs.
3. Explain what phylogenomics is and how it differs from single-gene phylogenetics.
4. Name two phylogenomic tree inference tools and state the method each uses.
5. Explain what Fst measures in population genomics.

5.4 Proteomics and Systems Biology

5.4.1 Proteomics: approaches and data

Proteomics is the large-scale study of the complete protein complement (proteome) of a cell, tissue, or organism under defined conditions, including the identity, abundance, modification state, and interaction partners of all proteins; the proteome is far more complex than the genome because each gene can produce multiple transcript isoforms (through alternative splicing) and protein variants (through post-translational modifications such as phosphorylation, glycosylation, ubiquitination, and proteolytic cleavage), and because protein abundance does not simply reflect mRNA abundance due to differences in translation efficiency and protein stability.

The major proteomics approaches are: bottom-up (shotgun) proteomics (proteins digested to peptides, separated by LC, identified by MS/MS, and quantified using label-free methods or chemical labelling approaches such as iTRAQ or TMT); top-down proteomics (intact proteins analysed directly without prior digestion, preserving information about post-translational modifications and protein isoforms but technically more demanding); targeted proteomics (selected reaction monitoring, SRM, or parallel reaction monitoring, PRM, measuring the abundance of a specific set of proteins with high sensitivity and reproducibility); and protein-protein interaction proteomics (affinity purification coupled with mass spectrometry, AP-MS, or proximity ligation approaches such as BioID, identifying the interaction partners of a protein of interest).

Fill in the blank: _____ proteomics analyses intact proteins without prior digestion, preserving information about post-translational modifications and protein isoforms at the cost of greater technical complexity.

5.4.2 Protein structure prediction and function

Protein structure prediction is the computational determination of the three-dimensional structure of a protein from its amino acid sequence; the three levels of protein structure prediction are secondary structure prediction (predicting which amino acids adopt alpha-helix, beta-sheet, or coil conformations, achieved with approximately 80 percent accuracy by methods such as

Psipred using evolutionary information from multiple sequence alignments), tertiary structure prediction by homology modelling (building a three-dimensional model of a protein based on its alignment to a protein of known structure, the template, using tools such as MODELLER and Swiss-Model), and ab initio structure prediction (predicting structure from sequence without a template, historically limited to small proteins but revolutionised by AlphaFold2 in 2021).

AlphaFold2 (developed by DeepMind, published in Nature in 2021) uses a deep learning architecture (Evoformer) that processes multiple sequence alignments and pairwise structural features to predict protein structures with experimental accuracy for the majority of proteins; the AlphaFold Protein Structure Database (AFDB) provides predicted structures for essentially all proteins in UniProt; protein function prediction from structure uses tools such as ProFunc, which identifies functional residues and binding sites by comparison with structurally characterised proteins; molecular docking (computational prediction of how a small molecule drug or a protein ligand binds to a protein target) uses structural information to support drug discovery and the understanding of protein-protein interactions.

Fill in the blank: AlphaFold2, developed by _____ and published in 2021, uses a deep learning architecture to predict protein three-dimensional structures with experimental accuracy for the majority of proteins.

5.4.3 Systems biology and network analysis

Systems biology takes a holistic, quantitative approach to understanding biological systems, treating them as integrated networks of interacting components (genes, transcripts, proteins, and metabolites) rather than studying individual molecules in isolation; the core data types of systems biology are multi-omics data (combining genomics, transcriptomics, proteomics, and metabolomics measurements from the same biological system), biological interaction networks (protein-protein interaction networks, gene regulatory networks, and metabolic networks), and dynamic mathematical models (ordinary differential equations or agent-based models describing how the state of the biological system changes over time).

Biological network analysis uses graph theory to study the properties and topology of biological networks: hub proteins (highly connected nodes in a protein interaction network, often essential

for cell viability and frequently implicated in disease), network modules (clusters of highly interconnected nodes representing protein complexes or co-regulated gene sets), and shortest paths (the minimum number of interactions connecting two nodes, relevant to signal transduction cascades and drug target identification); tools widely used for network analysis include Cytoscape (a desktop application for network visualisation and analysis), the STRING database (providing protein interaction networks with confidence scores for over 5,000 organisms), and the WGCNA R package (weighted gene co-expression network analysis, for identifying co-expressed gene modules from RNA-seq data).

Fill in the blank: _____ proteins are highly connected nodes in a biological network that are often essential for cell viability and frequently implicated in disease.

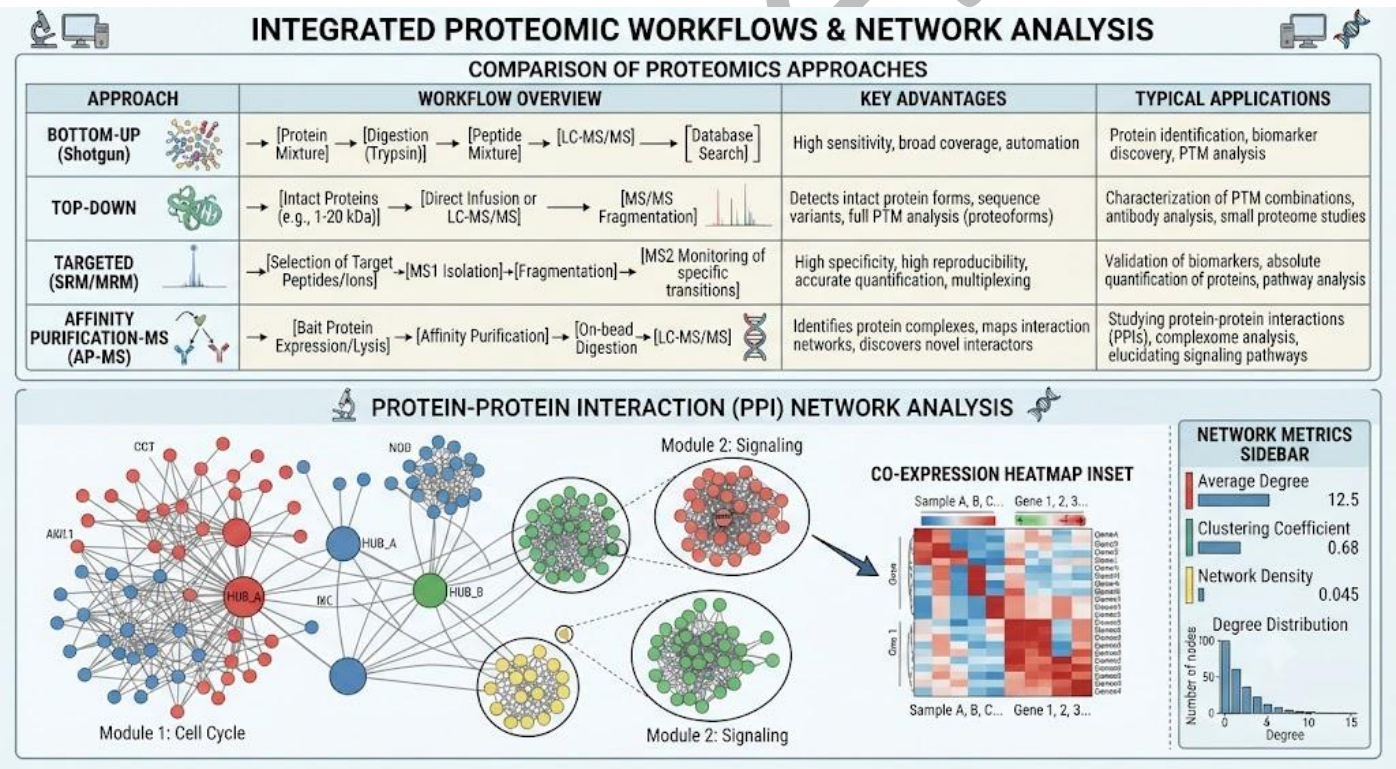

Figure 5.4: Proteomics and systems biology

Pilot questions 5.4

1. Explain why the proteome is more complex than the genome.
2. Distinguish between bottom-up and top-down proteomics.
3. Explain what protein homology modelling is and name one tool used for it.

4. Explain why AlphaFold2 was considered revolutionary in structural bioinformatics.
5. Explain what a hub protein is in a biological network and why it is important.

Series Summary

This final Series brought together the full scope of genomics and proteomics as integrated applications of bioinformatics. Genomics begins with the organisation of genomes, progresses through sequencing and de novo assembly to produce contigs and scaffolds assessed by N50 statistics, and culminates in structural and functional genome annotation using ab initio prediction and experimental evidence. Functional genomics through RNA-seq quantifies the transcriptome, with differential expression analysis identifying biologically significant expression changes, while epigenomics and single-cell genomics reveal the regulatory landscape and cellular diversity hidden in bulk data.

Comparative genomics identifies orthologs, syntenic regions, and gene family evolution across species, while phylogenomics leverages genome-scale data to resolve evolutionary relationships with far greater statistical power than single-gene approaches. Population genomics applies these tools to variation within and among populations, addressing adaptation, demographic history, and the genetic basis of traits through GWAS. Proteomics complements genomics by characterising the actual protein complement and its modification state, while systems biology integrates multi-omics data into network models that reveal the emergent properties of biological systems, completing the bioinformatics perspective from sequence to biological understanding.

Glossary of Terms

- **AlphaFold2:** a deep learning system developed by DeepMind that predicts protein three-dimensional structure from amino acid sequence with experimental accuracy, published in 2021.
- **Differential expression analysis:** the statistical identification of genes with significantly different expression levels between conditions using RNA-seq count data; tools include DESeq2 and edgeR.
- **Epigenomics:** the genome-wide study of epigenetic modifications including DNA methylation, histone modifications, and chromatin three-dimensional organisation.

- **Genome annotation:** the process of identifying and describing functional elements in a genome sequence, including protein-coding genes, non-coding RNAs, regulatory elements, and repeats.
- **Hub protein:** a highly connected node in a biological protein interaction network, often essential for cell viability and frequently implicated in disease.
- **N50:** the genome assembly statistic defined as the length such that half of all assembled sequence is contained in contigs of this length or longer.
- **Ortholog:** a gene in a different species that diverged from a common ancestor by speciation, typically retaining the same biological function.
- **Phylogenomics:** the inference of evolutionary relationships using genome-scale sequence data, providing greater statistical power than single-gene phylogenetics.
- **Population genomics:** the application of genome-wide sequence data to study genetic variation within and between populations, including demographic history, adaptation, and GWAS.
- **Shotgun proteomics:** a mass spectrometry approach for global protein identification in which proteins are digested to peptides, separated by LC, and identified by MS/MS database searching.

Concept Check Questions [Series 5]

Objective Questions

1. Transposable elements constitute approximately _____ percent of the human genome. (Linked to SLO 5.1)
2. The N50 statistic is the length such that half of all assembled sequence is contained in contigs of this length or _____ . (Linked to SLO 5.1)
3. RNA-seq reads are aligned using a splice-_____ aligner such as STAR that handles reads spanning exon-exon junctions. (Linked to SLO 5.2)
4. _____ are genes in different species that diverged by speciation and typically retain the same biological function. (Linked to SLO 5.3)
5. _____ proteins are highly connected nodes in a biological interaction network and are often essential for cell viability. (Linked to SLO 5.4)

Short-Answer Questions

1. Distinguish between structural and functional genome annotation. (Linked to SLO 5.1)
2. State two advantages of RNA-seq over microarrays. (Linked to SLO 5.2)
3. Explain what phylogenomics is and how it differs from single-gene phylogenetics. (Linked to SLO 5.3)
4. Explain why the proteome is more complex than the genome. (Linked to SLO 5.4)
5. Explain what AlphaFold2 is and why it was revolutionary. (Linked to SLO 5.4)

Long-Answer Questions

1. Describe genomic organisation, genome assembly, and genome annotation. (Linked to SLO 5.1)
2. Describe RNA-seq transcriptomics and differential expression analysis, and explain comparative genomics, phylogenomics, proteomics, and systems biology. (Linked to SLOs 5.2, 5.3, 5.4)

Answers to Concept Check Questions [Series 5]

Objective Questions

1. 45.
2. Longer.
3. Aware.
4. Orthologs.
5. Hub.

Short-Answer Questions

1. Structural annotation identifies the location and boundaries of genes (exons, introns, start/stop codons, splice sites); functional annotation assigns biological functions to the predicted gene products (using BLAST, InterPro, GO terms, KEGG pathway mapping).
2. Higher dynamic range (detects very low and very high expression) and does not require prior knowledge of gene sequences (can detect novel transcripts and isoforms).
3. Phylogenomics infers evolutionary trees using genome-scale data (thousands of orthologous genes) rather than one or a few genes; it provides far greater statistical power and resolves relationships that are ambiguous with single-gene approaches.
4. Each gene can produce multiple protein isoforms through alternative splicing; proteins undergo post-translational modifications (phosphorylation, glycosylation, ubiquitination) that create further diversity; protein abundance does not simply reflect mRNA abundance due to translation efficiency and protein stability differences.
5. AlphaFold2 is a deep learning system from DeepMind that predicts protein three-dimensional structures from amino acid sequences with experimental accuracy; it was revolutionary because protein structure prediction had been an unsolved problem for 50 years, and AlphaFold2 solved it at scale for all known proteins.

Long-Answer Questions

1. Genome organisation: human genome 3.2 Gb, ~20,000 protein-coding genes (~1.5% of sequence), ~45% transposable elements, packaged in chromatin on 46 chromosomes; bacterial genomes smaller, more compact. Assembly: reference-guided (existing reference) vs de novo (new organisms); de novo stages: overlap detection, De Bruijn graph, contig

assembly, scaffolding, gap filling; N50 measures contiguity. Annotation: structural (ab initio gene prediction with AUGUSTUS + RNA-seq/protein evidence) + functional (BLAST homology, InterPro domains, GO terms, KEGG pathways).

2. RNA-seq: extract RNA, cDNA library, Illumina sequencing, splice-aware alignment (STAR), count matrix; advantages over microarrays: higher dynamic range, novel transcript detection. DE analysis: DESeq2/edgeR, negative binomial model, log2 fold change, Benjamini-Hochberg FDR-adjusted p-values; volcano plot and heatmap visualisation. Comparative genomics: whole-genome alignment, ortholog identification (OrthoFinder), gene family evolution. Phylogenomics: multi-gene super-matrix or coalescent methods, IQ-TREE/BEAST. Population genomics: GWAS, Fst, LD. Proteomics: shotgun LC-MS/MS, top-down, targeted SRM, AP-MS. AlphaFold2: structure prediction. Systems biology: network analysis, hub proteins, Cytoscape, STRING, WGCNA.

Answers to Pilot Questions [Series 5]

Part 5.1

1. The human genome is approximately 3.2 billion base pairs (3.2 Gb) and contains approximately 20,000 protein-coding genes.
2. Transposable elements are mobile genetic elements capable of copying or moving themselves to new positions in the genome; they constitute approximately 45 percent of the human genome and are a major component of non-coding repetitive sequence.
3. Reference-guided assembly aligns reads to an existing reference genome (appropriate when a close reference is available); de novo assembly assembles reads from scratch into contigs and scaffolds (required for new organisms or for detecting structural variants not in the reference).
4. N50 is the contig length such that half of all assembled sequence is contained in contigs of this length or longer; higher N50 indicates a more contiguous assembly.
5. Structural annotation identifies the location and boundaries of genes (exons, introns, splice sites, start/stop codons); functional annotation assigns biological functions to predicted gene products using homology, domain analysis, and pathway mapping.

Part 5.2

(1) Extract total RNA or poly-A selected mRNA; (2) convert to cDNA and prepare NGS library; (3) sequence on Illumina; (4) align reads to reference genome using STAR or HISAT2; (5) count reads mapping to each gene to produce a gene x sample count matrix.

1. Higher dynamic range (detects very low and very high expression levels in the same experiment) and does not require prior knowledge of gene sequences (can detect novel transcripts, isoforms, and fusion genes not represented in any array design).
2. Log2 fold change is the base-2 logarithm of the ratio of expression in one condition versus another; a log2FC of 1 means expression has doubled, log2FC of -1 means it has halved, and log2FC of 0 means no change.
3. A volcano plot has log2 fold change on the x-axis and -log10(adjusted p-value) on the y-axis; genes in the upper right are significantly upregulated (large positive fold change, small adjusted p-value); genes in the upper left are significantly downregulated; genes in the lower portion are not significantly differentially expressed.
4. scRNA-seq reveals cellular heterogeneity hidden in bulk RNA-seq averages: different cell types within a tissue, rare cell populations, cell states, and developmental trajectories that are obscured when all cells are averaged together.

Part 5.3

(1) Comparative genomics studies similarities and differences in genome structure, content, and organisation between species to understand genome evolution and identify functional elements; MUMmer is a tool for whole-genome alignment of closely related genomes.

(2) Orthologs are genes in different species that diverged by speciation and typically retain the same function; paralogs are genes within or across species that diverged by gene duplication and may have different functions.

(3) Phylogenomics infers evolutionary relationships using genome-scale sequence data (thousands of orthologous genes); it differs from single-gene phylogenetics by providing far greater statistical power and phylogenetic signal, resolving relationships that are ambiguous or conflicting when based on a single gene.

(4) IQ-TREE (maximum likelihood tree inference with model testing and ultrafast bootstrap support) and BEAST (Bayesian inference producing time-calibrated trees with divergence time estimates).

(5) Fst (fixation index) measures genetic differentiation between populations: 0 indicates identical allele frequencies (no differentiation) and 1 indicates completely different allele frequencies (complete differentiation); it identifies genomic regions that differ between populations, potentially reflecting natural selection.

Part 5.4

(1) Each gene can produce multiple protein isoforms through alternative splicing; proteins undergo post-translational modifications creating additional diversity; protein abundance is not proportional to mRNA abundance because translation efficiency and protein stability vary.
(2) Bottom-up proteomics digests proteins to peptides with trypsin, separates by LC, and identifies by MS/MS; it is comprehensive but loses intact protein information. Top-down proteomics analyses intact proteins directly without digestion, preserving modification and isoform information but is technically more demanding and has lower proteome coverage.
(3) Homology modelling builds a three-dimensional protein model based on alignment to a protein of known structure (template); tools: MODELLER and Swiss-Model.
(4) AlphaFold2 predicts protein three-dimensional structures from amino acid sequences with experimental accuracy using a deep learning Evoformer architecture; it was revolutionary because protein structure prediction had been an unsolved grand challenge in biology for over 50 years, and AlphaFold2 solved it for essentially all known proteins.
(5) A hub protein is a highly connected node (high degree) in a protein interaction network; it is important because hub proteins are often essential for cell viability (their removal disrupts many cellular functions), are frequently implicated in disease, and are attractive drug targets.

References and Attributions

Textual References (Books and External Sources)

- Lesk, A. M. (2019). Introduction to bioinformatics (4th ed.). Oxford University Press.
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.
- National Open University of Nigeria (NOUN). (2023). Bioinformatics course material. NOUN Press.

Figures, Plates, Tables, and Boxes

- Figure 5.1: Genomics principles Attribution: AI generated. Suggested OER search string: "genomics genome organisation assembly annotation de novo contig scaffold N50 gene prediction OER".
- Figure 5.2: Functional genomics and transcriptomics Attribution: AI generated. Suggested OER search string: "RNA-seq transcriptomics differential expression DESeq2 volcano plot heatmap single cell scRNA-seq epigenomics OER".
- Figure 5.3: Comparative genomics and phylogenomics Attribution: AI generated. Suggested OER search string: "comparative genomics phylogenomics ortholog synteny population genomics GWAS phylogenetic tree IQ-TREE OER".
- Figure 5.4: Proteomics and systems biology Attribution: AI generated. Suggested OER search string: "proteomics shotgun top-down systems biology protein interaction network hub module Cytoscape STRING bioinformatics OER".