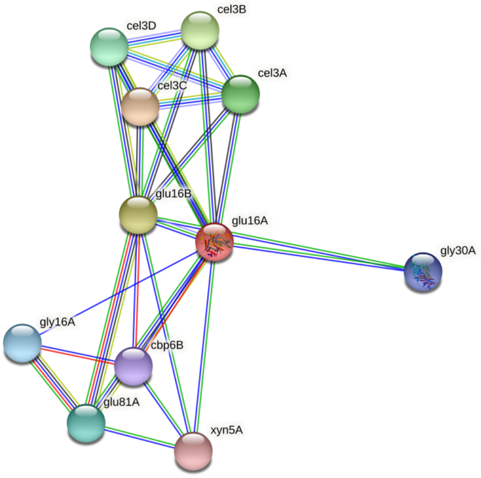
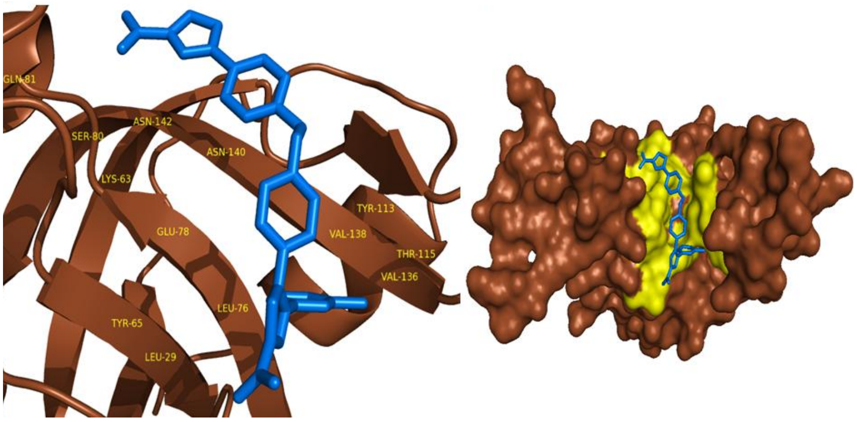
الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research
جامعة غرداية
University of Ghardaia
Faculty of Natural Sciences, Life and Earth Sciences
Département of Biology

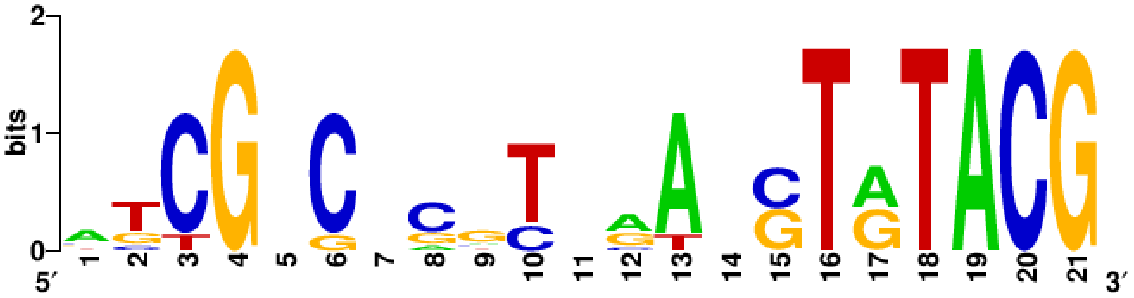
Course Handout of

Bioinformatics

Master 1 Applied Microbiology







Mahfoud BAKLI, PhD

2025-2026

SYNOPSIS (Version Française)

Le syllabus du cours de la matière, Bioinformatique tel que présenté dans le canevas de Master est présenté ci-dessous:

Intitulé du Master : Microbiologie appliquée Semestre : 2

Unité d'Enseignement	VHS	V.H hebdomadaire				Coeff	Crédits	Mode d'évaluation	
	14-16 sem	C	TD	TP	Autres			Continu	Examen
UE UE méthodologie									
Matière 1: Epidémiologie, bio-sécurité et hygiène hospitalière	60	1h30	3h	1h	55h00	03	05	40%	40%
Matière 2: Bioinformatique	**45h**	1h30	**1h30**	**-**	65h00	**02**	**04**	**40%**	**40%**

Objectifs de l'enseignement

La spécialité Bio-Informatique vise à donner une compétence tant en informatique qu'en bio-informatique à des scientifiques non informaticiens ayant une première compétence en biologie. Au premier semestre de l'année de Master1, les étudiants issus des licences de Microbiologie approfondiront leurs compétences en Biologie par le biais d'unités d'enseignement de la spécialité Microbiologie comprenant notamment une unité de bio-informatique. Au second semestre, ces étudiants suivront une formation de premier niveau en Bioinformatique permettant de comprendre de manière approfondie les liens multiples existant entre la séquence, la structure et la fonction des structures biologiques. Elle traite également le contrôle de la fonction des structures biologiques par mécanisme ingénieux mis en place au cours de l'évolution.

Connaissances Préalables Recommandées

Connaissance de l'usage de l'outil informatique, connaissances avancées en Biologie moléculaire, en Biologie générale, en physiologie cellulaire, en physiologie moléculaire et en Biochimie structurale.

Contenu de la matière

1. Introduction à la bioinformatique (la définition des concepts propres à l'information biologique et à son analyse par ordinateur)
2. Description des banques de données utiles au biologiste ainsi que leurs consultations via l'internet; banques de données biologiques
3. Utilisation des logiciels appropriés
4. Traitements bioinformatiques des séquences biologiques
5. Principes de bases de l'alignement de séquences et comparaison des méthodes d'alignement de séquences, dendrogrammes et taxonomie
6. Analyse, visualisation et modélisation de structures
   a. visualisation des structures tridimensionnelles
   b. Prédiction des structures secondaires, tertiaire et quaternaires des protéiques
   c. Choix des méthodes de prédiction est justifié selon le contexte de la recherche alors que les démarches expérimentales pour valider les prédictions sont également discutées
7. Etablir la relation entre la structure et la fonction

SYNOPSIS (English Version)

The syllabus for the Bioinformatics course, as presented in the Master's program framework, is detailed below:

Master's Program Title: Applied Microbiology   Semester : 2

Teaching unit	**Total Hours**	**Weekly Hours**				**Coeff**	**Credits**	**Assessment Mode**	
	14-16 weeks	**L**	**DW**	**PW**	**Others**			**Continuous**	**Exam**
UE methodology									
Course 1: Epidemiology, Biosafety, and Hospital Hygiene	60	1h30	3h	1h	55h00	03	05	40%	40%
Course 2: Bioinformatics	**45h**	1h30	**1h30**	**-**	65h00	**02**	**04**	**40%**	**40%**

Teaching Objectives

The Bioinformatics specialization aims to provide competence in both computer science and bioinformatics to non-computer scientist professionals with a primary background in biology. In the first semester of the Master's year 1, students from Microbiology undergraduate degrees will deepen their skills in Biology through teaching units from the Microbiology specialization, including notably a bioinformatics unit. In the second semester, these students will follow a first-level training in Bioinformatics, enabling a thorough understanding of the multiple links existing between the sequence, structure, and function of biological structures. It also covers the control of biological function through ingenious mechanisms established during evolution.

Recommended Prerequisite Knowledge

Knowledge of computer usage, advanced knowledge in Molecular Biology, General Biology, Cell Physiology, Molecular Physiology, and Structural Biochemistry.

Course Content

1. Introduction to Bioinformatics (definition of concepts specific to biological information and its computer analysis)
2. Biological Databases
3. Use of Appropriate Software
4. Bioinformatic Analysis of Biological Sequences
5. Basic Principles of Sequence Alignment and Comparison of Sequence Alignment Methods, Dendrograms and Taxonomy
6. Analysis, Visualization, and Modeling of Structures
   a. Visualization of three-dimensional structures
   b. Prediction of secondary, tertiary, and quaternary structures of proteins
   c. Choice of prediction methods is justified according to the research context, and the experimental approaches to validate the predictions are also discussed
7. Establishing the Relationship between Structure and Function

CONTENT

LIST OF FIGURES

LIST OF TABLES

Chapter 1. Introduction to bioinformatics

1.1. Concepts and Definitions

Bioinformatics is an interdisciplinary field emerging at the confluence of the ongoing revolutions in molecular biology and computer science. It is formally defined as the application of computational databases and algorithms for the analysis of biological data, including protein sequences, genes, and the complete DNA content that constitutes an organism (the genome) (Bayat 2002). A principal challenge in modern biology is the interpretation of the vast quantities of sequence and structural data produced by high-throughput initiatives such as genome-sequencing projects and proteomics. The computational tools of bioinformatics are essential for elucidating fundamental mechanisms in biology, encompassing macromolecular structure and function, biochemical pathways, disease pathogenesis, and evolutionary processes (Zhang *et al.* 2025).

Institutional definitions from the National Institutes of Health (NIH) and the National Human Genome Research Institute (NHGRI) characterize bioinformatics as a discipline concerned with the research, development, and application of computational tools to expand the use of biological data, including its acquisition, storage, organization, analysis, and visualization. The closely related field of computational biology is distinctively focused on the development of data-analytical theories, mathematical modeling, and computational simulation techniques for the study of biological systems (Pevsner 2015).

The term "bioinformatics" was originally defined by Ben Hesper and Paulien Hogeweg as "the study of informatic processes in biotic systems" (Hesper and Hogeweg 1970). This conceptualization represented bioinformatics in its broadest sense, interpreting biological systems as integrated informational networks.

The modern usage of the term was established in the mid-1990s, notably in foundational works by Harper (1994) and Boguski (1994), whose studies applied the term specifically to the computational analysis of biological data (Harper 1994; Boguski 1994).

Subsequently, the term has evolved to denote the application of computational methodologies to analyze and interpret biological data. This evolution toward a more applied discipline is further elaborated by Russ B Altman (1998) and Altman and Dugan (2003), who conceptualize bioinformatics through two paradigms: one concerning the management of information flow according to the central dogma of molecular biology (**Fig. 1**), and another addressing the computational scientific method itself, which includes software design and validation, data sharing, reproducible research workflows, and experimental interpretation (Altman and Dugan 2003; Altman 1998).

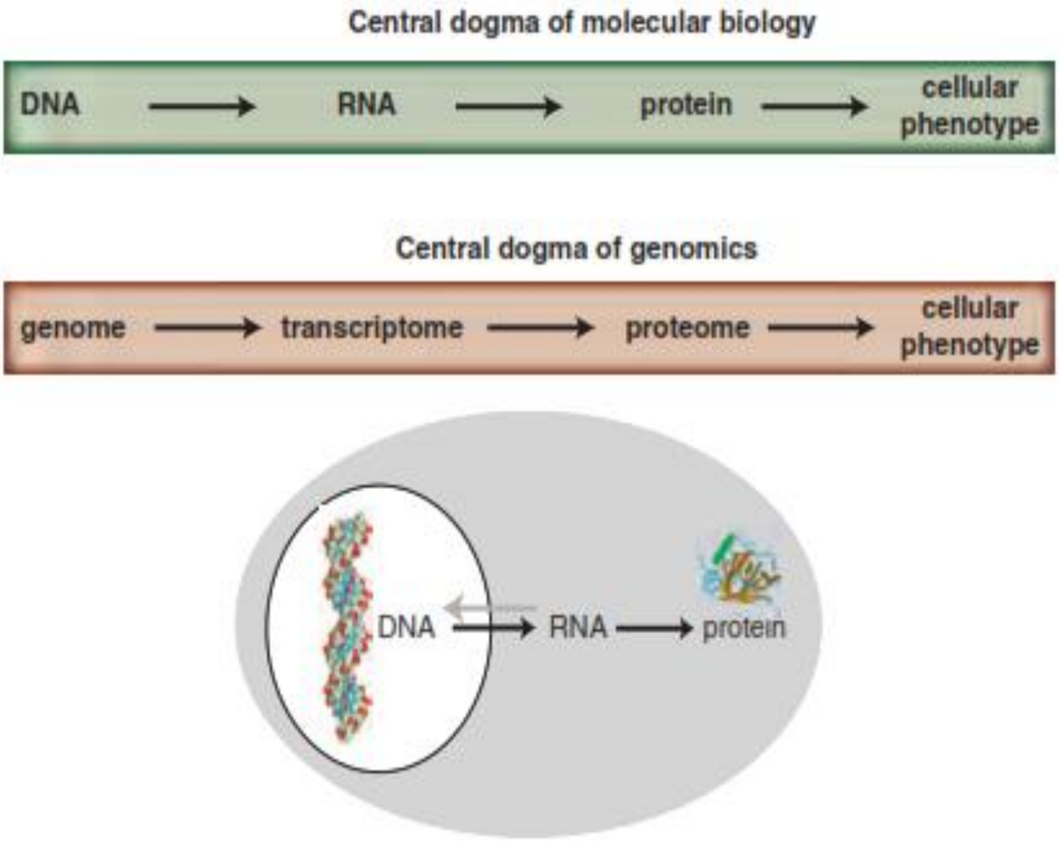

Central dogma of molecular biology
DNA → RNA → protein → cellular phenotype
Central dogma of genomics
genome → transcriptome → proteome → cellular phenotype
DNA
RNA
protein

Figure 1. Bioinformatics and the cell: Managing molecular data. Bioinformatics has emerged to manage the vast growth of molecular sequence data. Major databases (EMBL, GenBank, SRA, DDBJ) store quadrillions of nucleotides, with complementary resources for DNA, RNA and proteins. The field focuses on analyzing these sequences to address biological questions at the cellular and molecular levels (Pevsner 2015).

While bioinformatics is primarily focused on the analysis of molecular sequences, it provides the foundational methodology for the closely linked disciplines of genomics (dedicated to determining and analyzing entire genome sequences) and functional genomics (which utilizes genome-wide assays to characterize gene and protein function). These fields now enable the comprehensive characterization of an individual's genome, transcriptome, proteome, metabolome, epigenetic modifications, and microbiome (Clark and Lillard Jr 2024).

Collectively, bioinformatics—or computational biology—constitutes a multidisciplinary science that employs computational and statistical approaches to acquire, store, analyze, and disseminate biological data and knowledge (**Fig. 2**). Driven by continuous technological and methodological advances in biomedical research, the field has become indispensable for managing and interpreting the exponential growth of data generated across diverse omics disciplines, including genomics, proteomics, metabolomics, and pharmacogenomics. This foundational role provides a critical bridge to the historical development of the field, as discussed in the following section (Demirbaga *et al.* 2024).

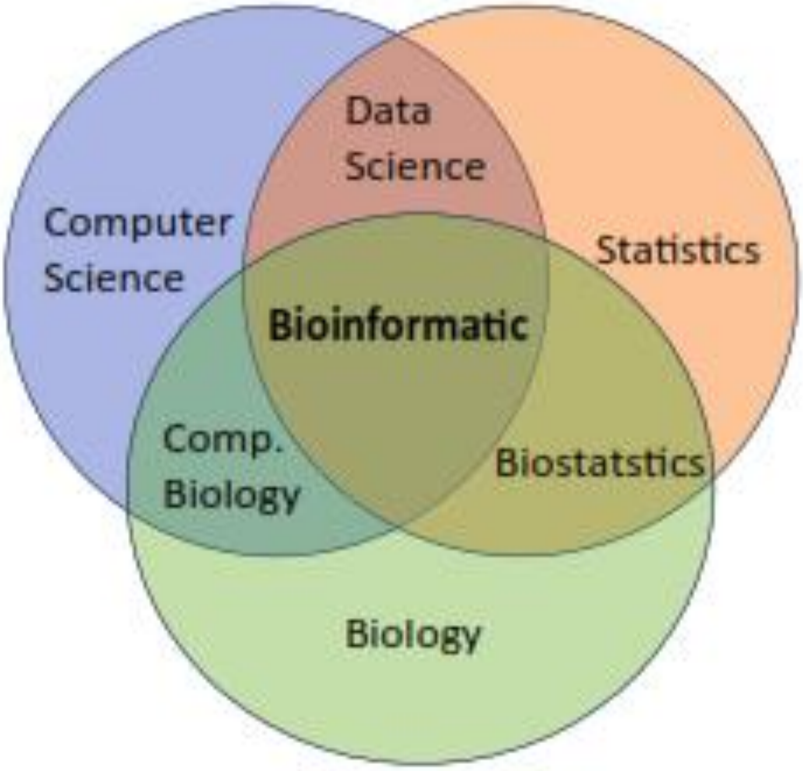


Figure 2. Bioinformatics as a multidisciplinary field (Demirbaga *et al.* 2024)**.**

1.2. Bioinformatics History

As seen above, the term "bioinformatics" was coined by Paulien Hogeweg in 1979; however, the systematic application of its principles began earlier in the 1960s through the work of Margaret Dayhoff, who is considered the field's pioneer (**Fig. 3**) (Badar 2023).

Figure 3. Margaret Oakley Dayhoff (1925-1983). A pioneering bioinformatician. Photo: Ruth E. Day-hoff, M.D.; US National Library of Medicine (Wünschiers 2025).

The origins of bioinformatics date to the 1950s, marked by Frederick Sanger's determination of insulin's protein structure (1958 Nobel Prize in Chemistry). Fortran was the programming language used by Margaret Dayhoff, who along with Robert S. Ledley, developed COMPROTEIN, an early program for protein structure determination that served as the first de novo sequence assembler and utilized a three-letter amino acid code; Dayhoff later introduced the standard single-letter code (Wünschiers 2025).

The 1970s saw significant growth with the development of the Needleman-Wunsch algorithm for sequence alignment, the establishment of the first nucleotide database (EMBL), and the creation of the PAM substitution model by Dayhoff's group. The following decades introduced key methods such as maximum likelihood phylogenetics in the 1980s and the BLAST algorithm for local alignment in the 1990s, with major milestones summarized in **Fig. 4** (Badar 2023).

The field has been driven by three major breakthroughs: the development of efficient DNA sequencing methods, the rise of supercomputing and the Internet, and the launch of genome projects. The completion of the Human Genome Project in 2003, alongside the sequencing of numerous other organisms, catalyzed the development of cost-effective, high-throughput next-generation sequencing (NGS) platforms.

Recent milestones include the development of the AI system AlphaFold in 2018, which represented a major advance in protein structure prediction, and the critical role of bioinformatics in the rapid genomic analysis of SARS-CoV-2 (Badar 2023).

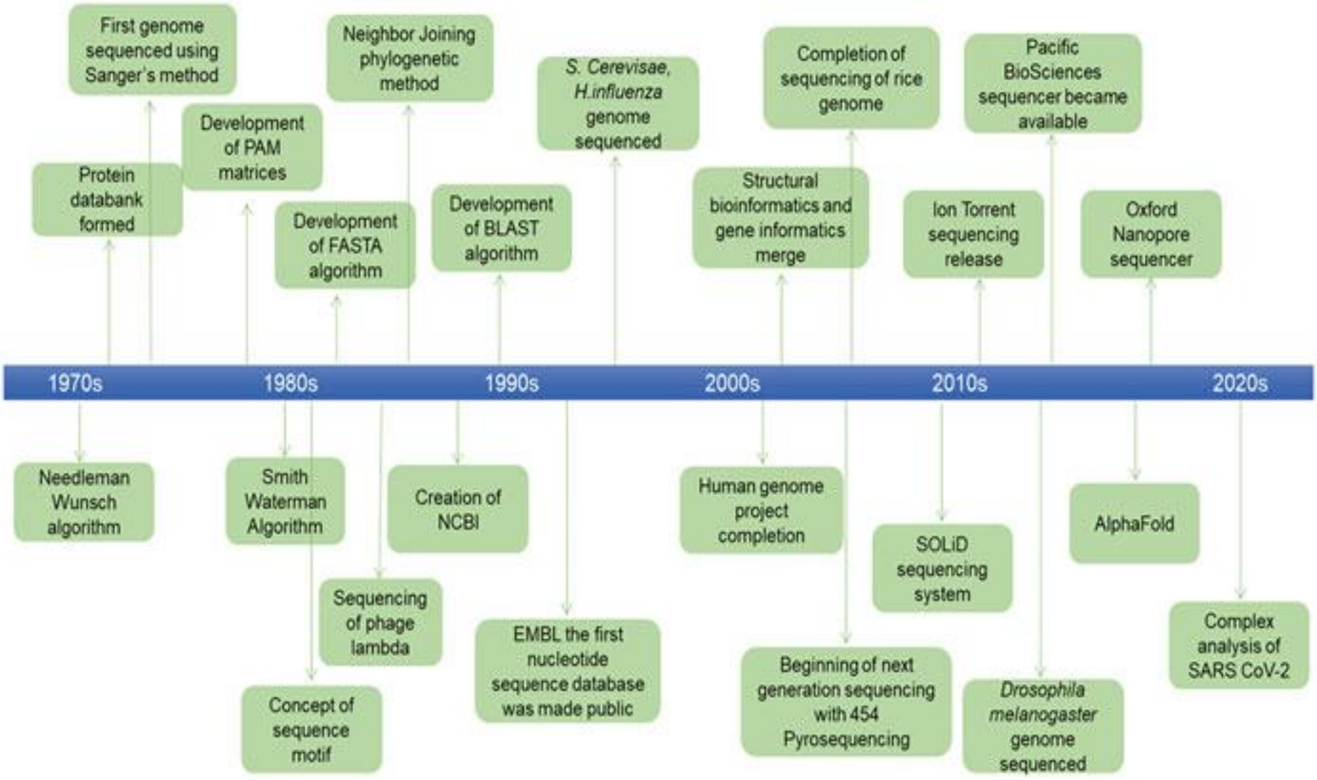

First genome sequenced using Sanger's method
Neighbor Joining phylogenetic method
S. Cerevisae, H.influenza genome sequenced
Completion of sequencing of rice genome
Pacific BioSciences sequencer became available
Development of PAM matrices
Protein databank formed
Development of FASTA algorithm
Development of BLAST algorithm
Structural bioinformatics and gene informatics merge
Ion Torrent sequencing release
Oxford Nanopore sequencer
1970s
1980s
1990s
2000s
2010s
2020s
Needleman Wunsch algorithm
Smith Waterman Algorithm
Creation of NCBI
Human genome project completion
SOLiD sequencing system
AlphaFold
Sequencing of phage lambda
EMBL the first nucleotide sequence database was made public
Beginning of next generation sequencing with 454 Pyrosequencing
Drosophila melanogaster genome sequenced
Complex analysis of SARS CoV-2
Concept of sequence motif

Figure 4. Key historical milestones in bioinformatics from the 1970s to the present. The field's foundations were established with the 1950s publication of insulin's protein sequence. Margaret Dayhoff, considered the first bioinformatician, pioneered systematic work in the 1960s by developing COMPROTEIN with Robert S. Ledley. This first *de novo* sequence assembler used three-letter amino acid codes; Dayhoff later introduced the standard single-letter code. The 1970s saw substantial progress including the Needleman-Wunsch algorithm for global sequence alignment, establishment of the EMBL nucleotide database, and creation of the PAM substitution model by Dayhoff's team. Although Paulien Hogeweg coined the term "bioinformatics" in 1979, methodological advances continued with 1980s maximum likelihood methods for phylogenetics. The 1990s sequence data explosion drove BLAST's development for rapid local alignment. Three transformative breakthroughs propelled the field: efficient DNA sequencing technologies, expanded supercomputing and Internet capabilities, and large-scale genome projects. The Human Genome Project's 2003 completion, building on shotgun sequencing of *Haemophilus influenzae* (1995) and *Drosophila melanogaster* (2010), accelerated next-generation sequencing (NGS) platforms. 2010s landmarks included novel sequencing technologies (SOLiD, Ion Torrent, Nanopore) and AlphaFold's 2018 protein folding breakthrough recognized by CASP. Most recently, bioinformatics proved indispensable for COVID-19 response through rapid SARS-CoV-2 genomic analysis and surveillance (Badar 2023).

1.3. Human Genome Project

Building on technological advances in molecular biology, the Human Genome Project (HGP) was launched to advance biomedical science. This 13-year international effort, led by the U.S. Department of Energy and National Institutes of Health and completed in 2003, had two primary goals: to identify all human genes (estimated at 20,000-25,000) and to determine the complete sequence of the 3 billion DNA base pairs that make up human DNA. The project's completion coincided with the 50th anniversary of Watson and Crick's discovery of DNA's structure. Although coordinated by U.S. agencies, the HGP was a global collaboration with significant contributions from the United Kingdom's Wellcome Trust, Japan, France, Germany, China, and others. The project was enabled by advances in database technology, information retrieval, and international networking capabilities (Singh 2025b).

The completion of the HGP marked the beginning of the post-genome era, shifting focus from sequencing to understanding biological function. A surprising finding was that humans possess only about 25,000 protein-coding genes—far fewer than the previously estimated 100,000 and only slightly more than the simple roundworm C. elegans. This revelation demonstrated that biological complexity arises not from gene number but from how genes are regulated and networked. This understanding has established that sequence information alone cannot fully explain biological function. Modern biology now requires integrated approaches combining computational modeling with experimental science, forming the essential foundation for effective bioinformatics research (Singh 2025b).

1.4. Genome Data Statistics

The National Center for Biotechnology Information (NCBI), part of the National Library of Medicine at the U.S. National Institutes of Health, was established in 1988 as a central repository for molecular biology information. It provides public databases and develops computational tools for the analysis of genomic data. Among these resources, the NCBI hosts GenBank, the NIH genetic sequence database, which serves as a comprehensive, annotated collection of all publicly available DNA sequences (Sayers *et al.* 2024).

Monthly submissions to GenBank exceed three million new sequences. As of GenBank Release 253.0 (June 2023), the database contained approximately 244 million sequences comprising over 1.4 trillion nucleotide bases, with continuous growth driven by ongoing worldwide submissions. The database includes sequences from more than 165,000 organisms, ranging from individual genes to complete genomes, and spans a wide taxonomic spectrum including humans, plants, invertebrates, and microorganisms (Sayers *et al.* 2023).

As shown in **Fig. 5**, both the size and number of sequences in GenBank have exhibited exponential growth, a trend sustained beyond the completion of the Human Genome Project. This expansion reflects continued scientific efforts to sequence diverse organisms and growing interest in exploring global biological diversity (Singh 2025b).

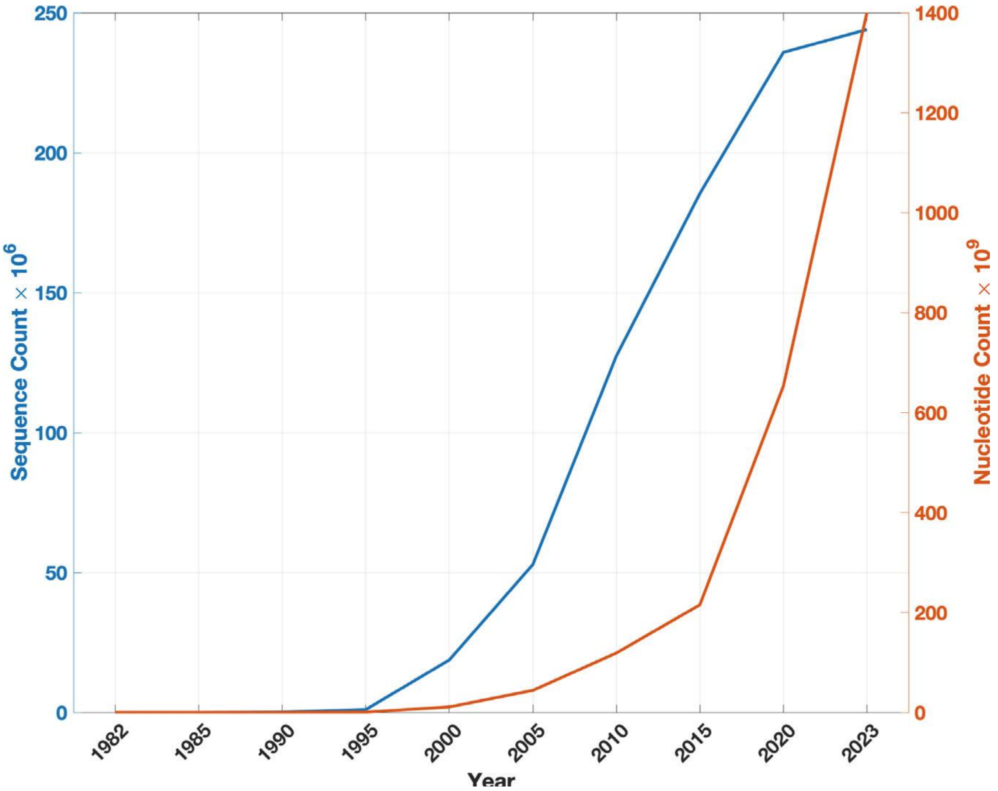


Figure 5. Exponential growth of the GENBANK nucleotide database since its inception through June 2023, showing a consistent increase in both the total number of sequence entries and the volume of nucleotide base pairs stored, reflecting sustained worldwide submission activity and expanding genomic sequencing efforts across diverse organisms (Singh 2025b).

1.5. Emergence of Artificial Intelligence in Computational Biology and Bioinformatics

Artificial intelligence is fundamentally reshaping the disciplines of computational biology and bioinformatics by introducing novel methodologies for large-scale biological data analysis. Recent advancements in AI have not only enhanced existing computational approaches but have also enabled entirely new forms of scientific inquiry. The technology's capacity to identify complex patterns, generate predictive models, and extract meaningful insights from massive datasets has created unprecedented opportunities for addressing longstanding challenges in biological research. This transformation is particularly valuable given the rapidly expanding volume and heterogeneity of biological data (Nalina *et al.* 2025).

1.5.1. Revolutionizing Key Research Areas

AI applications are revolutionizing several key areas of the life sciences:

1.5.1.1. Drug Discovery

AI systems are revolutionizing drug discovery by analyzing the genetic and molecular foundations of diseases to identify potential therapeutic compounds. This approach significantly accelerates the drug development pipeline, from target identification to lead optimization, facilitating the creation of more effective and targeted treatments (Ferreira and Carneiro 2025).

1.5.1.2. Cancer Genomics

In cancer genomics, machine learning algorithms are deployed to detect somatic mutations, chromosomal aberrations, and molecular biomarkers from sequencing data. Deep learning algorithms are particularly valuable for creating predictive models, as they can analyze complex historical patient data to identify subtle patterns and prognostic trends that would be difficult for clinicians to detect manually (**Fig. 6**).

These computational insights are critical for informing targeted therapy selection and enabling more personalized and effective treatment strategies for cancer patients (Calvino *et al.* 2025).

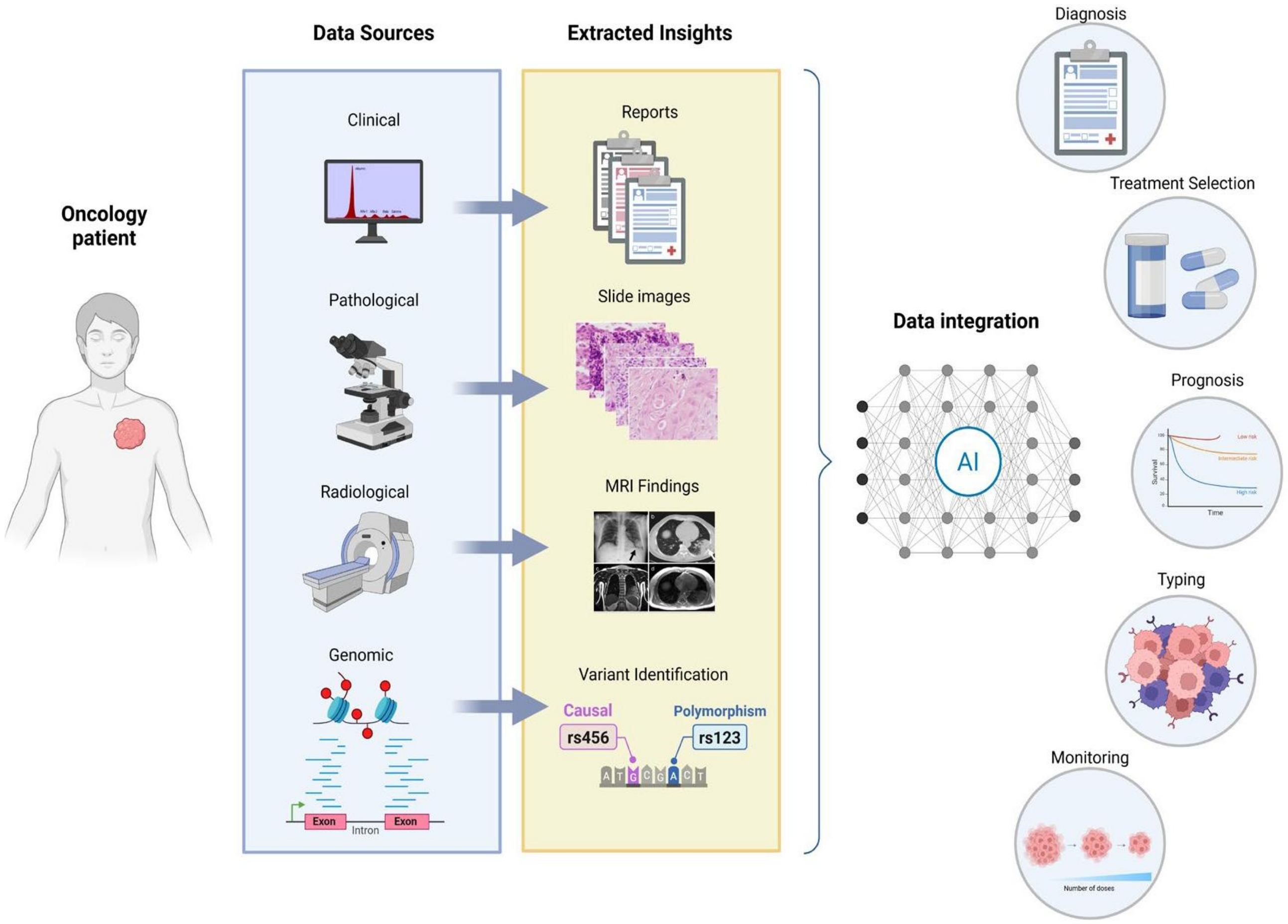


Figure 6. Multimodal Data Integration in Cancer Genomics. AI leverages integrated clinical, pathological, radiological, and genomic data to assist in diagnosis, prognosis, treatment selection, and monitoring (Calvino *et al.* 2025).

1.5.1.3. Genome Editing

AI models enhance the precision and efficacy of genome-editing technologies, such as CRISPR-Cas9. By improving guide RNA design, predicting off-target effects, and optimizing editing conditions, AI increases the safety and reliability of genetic interventions for research and therapeutic purposes (Vamathevan *et al.* 2019).

1.5.1.4. Predictive Medicine

The field of predictive medicine leverages increasingly sophisticated AI algorithms to analyze individual genetic profiles, lifestyle data, and clinical histories. This enables a more accurate assessment of disease susceptibility and progression, paving the way for early interventions and personalized health management plans (Karim *et al.* 2021).

1.5.1.5. Population Genetics

AI techniques are applied to population genetics to reconstruct human evolutionary history, migration patterns, and demographic events. By analyzing genetic variation across diverse populations, these models provide profound insights into our species' origins, dispersal, and adaptation (Le *et al.* 2020).

1.5.2. Methodological Frameworks and Implementation

1.5.2.1. Protein Structure Prediction

The field of protein structure prediction has been revolutionized by deep learning systems like AlphaFold. These systems employ comprehensive pipelines involving extensive data curation, feature extraction, model training, and structural validation (Sibli *et al.* 2025). They leverage advanced neural network architectures and substantial computational resources to achieve unprecedented accuracy in predicting three-dimensional protein structures from amino acid sequences (Wang *et al.* 2025).

1.5.2.2. Metagenomic Analysis

In metagenomics, researchers process complex sequencing data from platforms such as Illumina. This involves sophisticated preprocessing steps to remove low-quality reads and adapter sequences (Yan and Wang 2022). Subsequent analysis often uses deep learning frameworks like TensorFlow and PyTorch for taxonomic classification. Optimization involves architecture design, hyperparameter tuning, and validation against curated reference databases (Mughal 2021).

1.5.2.3. Biological Network Analysis

For the analysis of biological networks, scientists first curate protein-protein interaction data from public repositories, followed by meticulous data normalization and quality control. Graph neural networks (GNNs) have emerged as particularly valuable tools for these analyses due to their innate ability to model complex relational data structures (Esposito *et al.* 2020).

Model development involves specialized optimization techniques and validation against known biological complexes and pathways (Ferreira and Carneiro 2025).

1.5.3. Ethical and Privacy Considerations

The integration of AI into biology raises critical ethical and privacy concerns that must be addressed through thoughtful governance and technical safeguards. As AI applications increasingly rely on sensitive genetic and health information, ensuring proper data anonymization, secure storage, ethical use guidelines, and equitable access becomes essential for maintaining scientific integrity and public trust (Weltz *et al.* 2022).

1.6. From Data to Theory: The Evolution of Bioinformatics

Bioinformatics serves as the data science of life, employing multidisciplinary methods to analyze biological data across all scales, from molecules to populations. It is a rapidly evolving, interdisciplinary field that has expanded in response to advances in other disciplines. Its historical progression, as outlined in **Fig. 7**, is marked by distinct stages, each initiated by a landmark event: the sequence-oriented stage (beginning in 1952 with Chargaff's rules), the omics-driven stage (launched in 1990 with the Human Genome Project), and the current AI-powered stage (initiated in 2018 with AlphaFold). The field is now poised to enter a theory-guided stage (beyond 2024), where biological theory will direct AI modeling and experimentation (Zhang 2024).

The primary challenges ahead are inherently data-centric. High-quality data is crucial for formulating robust theory, which in turn is the key to developing more successful AI applications.

These future models will be characterized by greater efficiency (fewer parameters, reduced training cost), enhanced predictive power, and improved explainability grounded in biological reasoning. This creates a critical feedback loop—data informs theory, and theory guides AI—a cycle that is accelerating a fundamental paradigm shift in biological research (Zhang 2024).

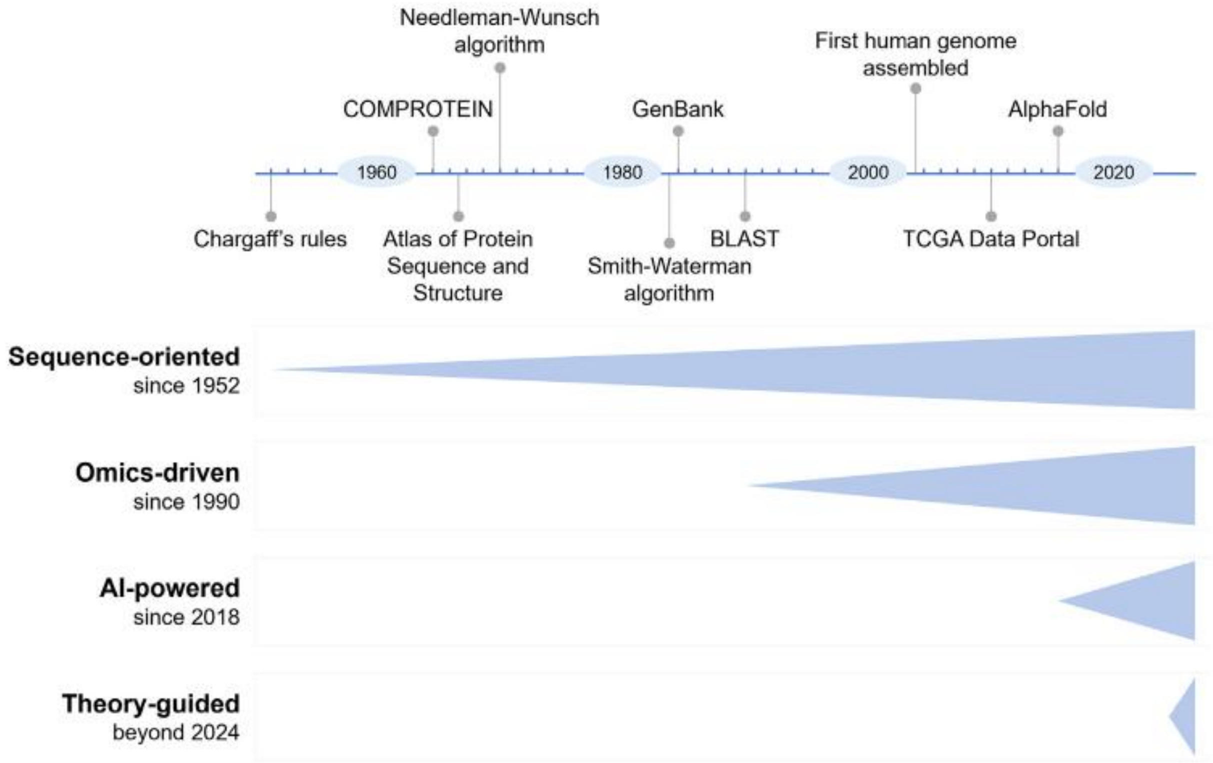


Figure 7. Schematic representation of four stages with major landmarks in bioinformatics. Bioinformatics has evolved through four phases: a sequence-oriented era (beginning circa 1952), an omics-driven era (from ~1990), a contemporary AI-powered era (since ~2018), and an anticipated theory-guided era (post-2024). Key developments include Chargaff's rules (1952), early computational tools (COMPROTEIN, 1962), the Atlas of Protein Sequence and Structure (1965), the Needleman-Wunsch (1970) and Smith-Waterman (1981) algorithms, the GenBank database (1982), the BLAST tool (1990), the TCGA Data Portal (2010), and AlphaFold (2018) (Zhang 2024).

1.7. Applications of Bioinformatics

Bioinformatics is broadly applied across medicine, agriculture, and biotechnology. It facilitates drug discovery, enables early disease detection, and supports gene therapy development. In agriculture, it aids crop improvement through genetic modifications. Bioinformatics tools analyze protein structures, propose novel compounds, and integrate multi-omics data, as illustrated in **fig. 8** (Iqbal and Kumar 2023).

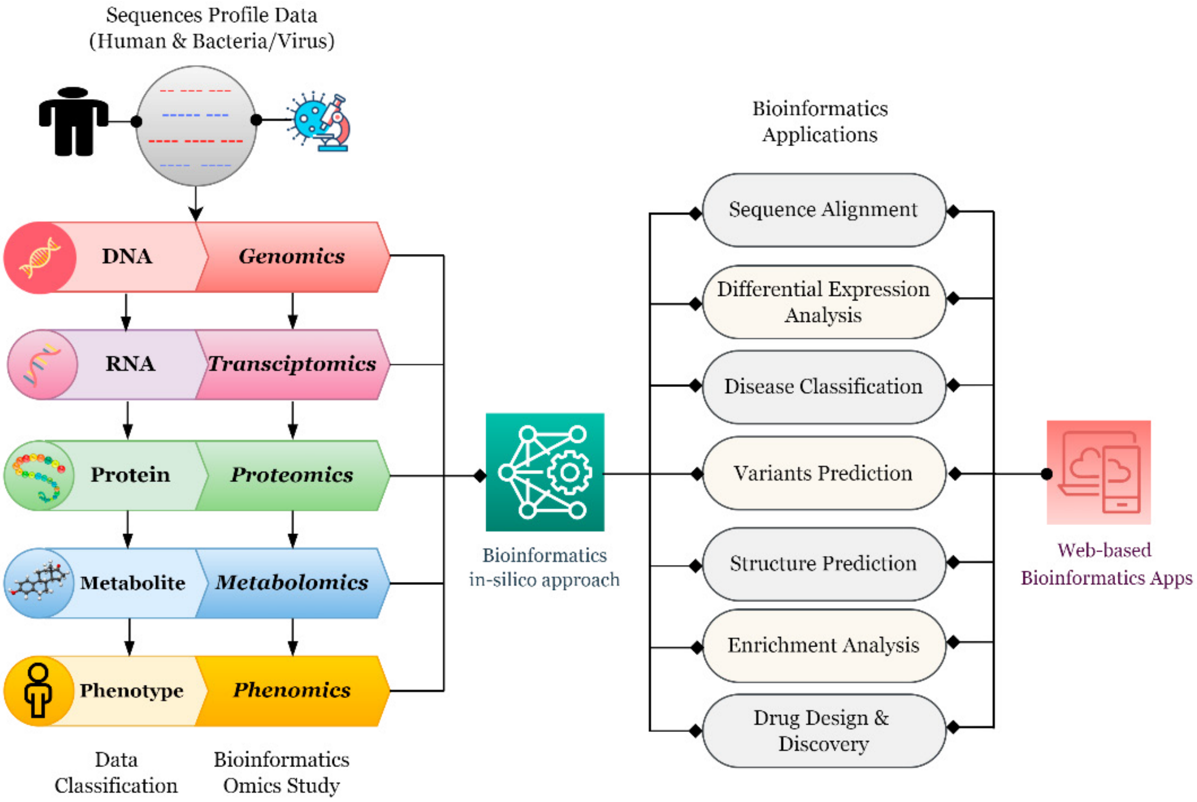


Figure 8. Omics profile data integration with various bioinformatics applications (Iqbal and Kumar 2023).

Building upon the applications in medicine, agriculture, and multi-omics integration highlighted above, bioinformatics is also fundamental to numerous other advanced fields (Singh 2025b);

1.7.1. Protein Structure Prediction

A primary challenge is the computational prediction of a protein's three-dimensional structure from its amino acid sequence, which remains a unsolved problem critical for advancing *in-silico* drug design.

1.7.2. Evolutionary and Comparative Genomics

This field analyzes sequence homology to determine evolutionary relationships and infer gene function. It computationally distinguishes orthologs, diverged through speciation, from paralogs, originated from gene duplication. This accurate classification is essential for reconstructing reliable phylogenetic trees and understanding functional gene evolution across diverse species.

1.7.3. Gene Expression Analysis

Bioinformatics tools are essential for analyzing genome-wide differential gene expression data from technologies like microarrays, crucial for understanding cellular responses in diseases like cancer.

1.7.4. Systems Biology Modeling

This includes constructing models of Gene Regulatory Networks (GRNs) to understand transcriptional control and simulating metabolic pathways to link genotype to molecular physiology.

1.7.5. Single-Cell Genomics

Techniques like single-cell RNA sequencing (scRNA-seq) require advanced computational methods to analyze gene expression at the individual cell level, revealing cellular heterogeneity within tissues.

1.7.6. Synthetic Biology

This multidisciplinary domain employs bioinformatics to design synthetic routes, create artificial genes, and optimize genomes for producing biofuels, medicines, and other important substances using genetically modified microorganisms.

1.7.7. Epigenomics

Computational tools analyze epigenetic modifications (e.g., DNA methylation, histone marks) to understand their role in gene regulation, development, and disease.

1.7.8. Machine Learning and AI Integration

Advanced algorithms are deployed to predict disease outcomes from complex datasets, identify biomarkers, categorize cancer subtypes, and assist in diagnostic imaging analysis.

1.7.9. Forensic Science

Bioinformatics enables the analysis of genetic evidence for criminal investigations, enhancing the precision of DNA profiling and the interpretation of degraded samples.

1.8. Goals of Bioinformatics

Bioinformatics development is structured around three primary objectives:

(i) The creation and maintenance of specialized databases to organize and store biological datasets derived from research. These resources enable efficient access to existing information and support the submission of new data. Key examples include nucleotide sequence repositories such as GenBank, EMBL, and DDBJ.

(ii) The design of computational tools and software for analyzing complex biological data, facilitating investigations that are challenging to address experimentally. Tool development requires integration of computational techniques and biological theory. For instance, BLAST and its variants are widely used for detecting local sequence alignments and identifying homologous relationships.

(iii) The utilization of stored data—through detailed analysis, interpretation, and application—to advance scientific research. Bioinformatics aims to generate novel biological insights and establish a unified understanding of biological principles, enabling the detection of overarching patterns in biological systems. The core domains of bioinformatics are summarized in **Fig. 9** (Badar 2023).

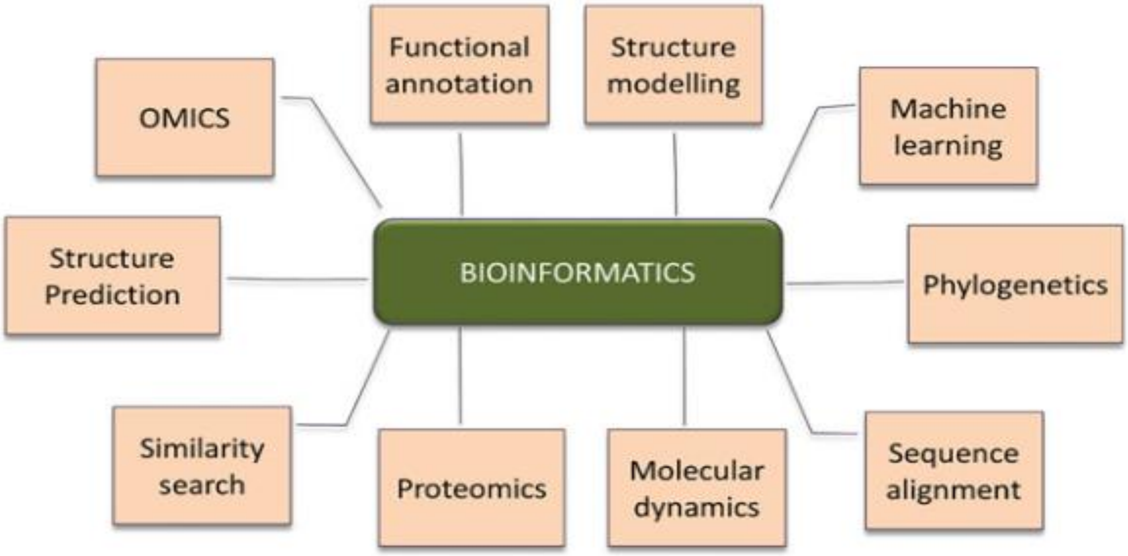
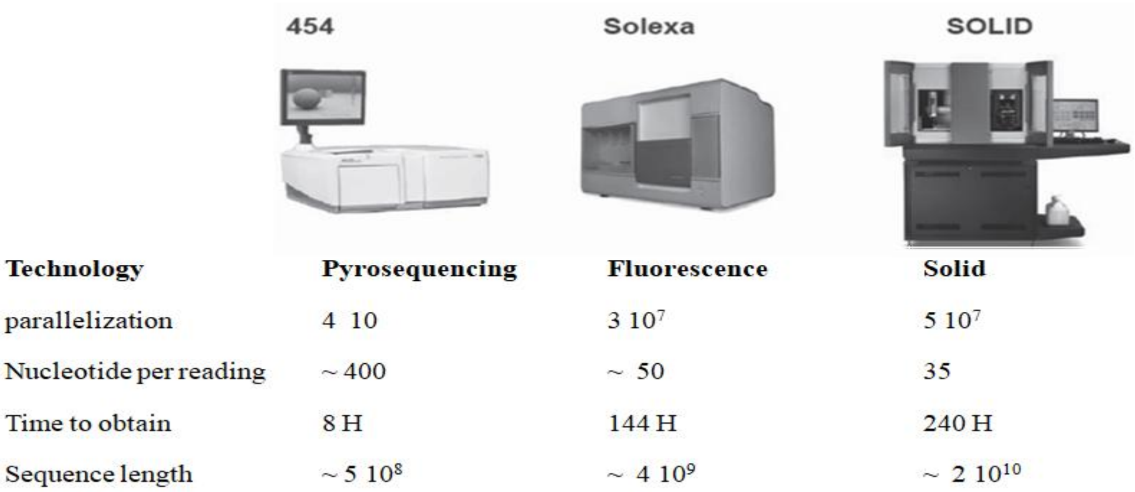


Figure 9. Core methodologies in bioinformatics. Sequence alignment, functional annotation of biological macromolecules, protein structure prediction, and evolutionary relationship analysis (Badar 2023).

Chapter 2. Biological Databases

2.1. Fundamentals of Biological Databases

The rapid advancement of high-throughput, cost-effective next-generation sequencing (NGS) technologies has led to an exponential growth in biological data. The capabilities of these techniques are so powerful that sequencing costs have plummeted by a factor of 500,000 over two decades; the cost of sequencing a single human genome fell from $100 million in 2001 to just $700 in 2019 (source: https://www.genome.gov/). This massive data output from whole-genome sequencing programs has created an urgent need for sophisticated data management solutions (Deléage *et al.* 2021). To manage this vast amount of information, numerous biological databases have been developed. These databases serve as centralized, digital repositories that store data in a structured format, enabling efficient organization, annotation, cross-referencing, and retrieval through specialized search tools (Danielewski *et al.* 2025). The process typically begins with NGS technologies, which generate enormous quantities of short sequence fragments (reads) that vary in length depending on the platform used. As illustrated in **Figure 10**, these methods differ in their characteristics, such as read length, throughput, and speed. Bioinformatic tools are then essential to assemble these fragments into a complete sequence for a small genome or to align them to a reference genome for larger projects (Deléage *et al.* 2021).



	454	Solexa	SOLID
Technology	Pyrosequencing	Fluorescence	Solid
parallelization	4 10	$3\ 10^{7}$	$5\ 10^{7}$
Nucleotide per reading	~ 400	~ 50	35
Time to obtain	8 H	144 H	240 H
Sequence length	$\sim 5\ 10^{8}$	$\sim\ 4\ 10^{9}$	$\sim\ 2\ 10^{10}$

Figure 10. Next Generation Sequencing (NGS) (Deléage *et al.* 2021).

The standard workflow for this genome analysis, as shown in **Fig. 11**, consists of three main stages: data acquisition and quality control, read alignment and mapping, and variant calling **(Singh and Kumar 2024)**.

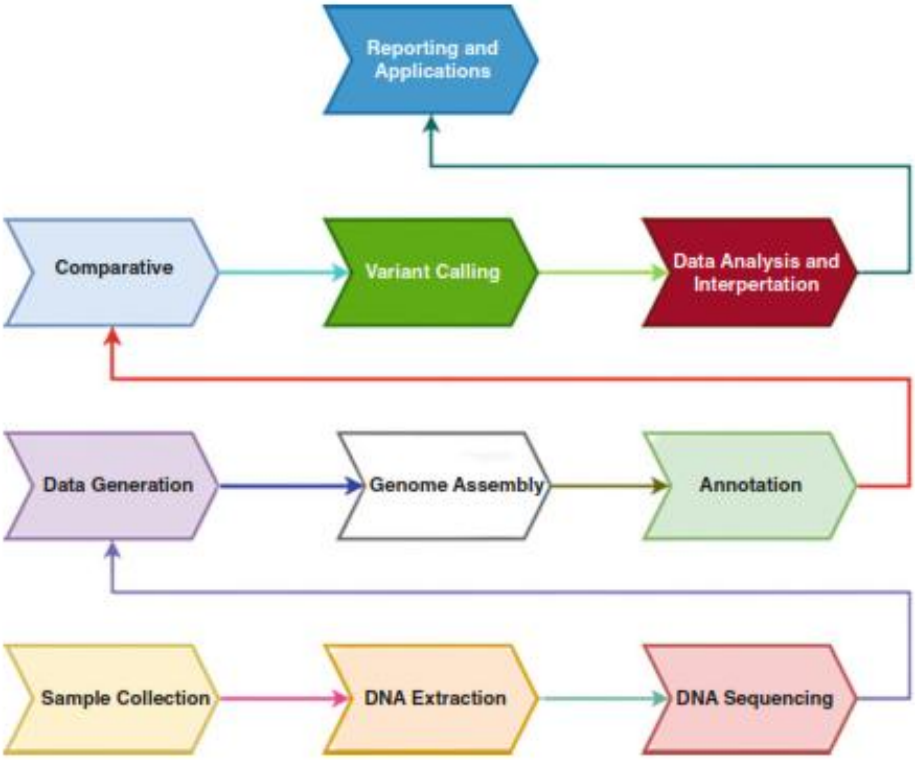


Figure 11. Basic work flow of genome sequencing processing (Singh and Kumar 2024).

Biological databases are broadly classified along two main criteria: data coverage and level of curation. Based on data coverage, they fall into comprehensive databases (e.g., GenBank), which aggregate data from a wide range of species, and specialized databases (e.g., WormBase), which focus on a single species or a specific type of data. Based on data curation, they are categorized as primary or secondary databases. Primary databases (e.g., the raw sequence data in GenBank) archive experimentally-derived data submitted directly by researchers. In contrast, secondary databases (e.g., Ensembl, UCSC Genome Browser) contain highly curated, often computationally processed information derived from primary sources. Some resources, like UniProt, function as hybrid databases, containing both primary sequences and expertly curated annotations (Zou *et al.* 2015).

Recognizing that information is fragmented across these specialized resources, integrated retrieval systems have been developed. The Entrez system, maintained by the National Center for Biotechnology Information (NCBI), is a prime example. The NCBI (https://www.ncbi.nlm.nih.gov/) itself classifies biological databases into two main types: Comprehensive databases, which store data from many organisms and multiple sequence types (nucleotide, protein, genomic), and Specialized databases, which focus on specific organisms (e.g., human or mouse) or data from particular sequencing technologies (Villalba and Matte 2021).

The Entrez system provides a unified interface to search across dozens of distinct molecular databases, which are grouped into six major categories: Literature, Genomes, Genes, Proteins, Health, and Chemicals **(Table 1)** (Tiwary 2022).

Using Boolean operators and allows data to be downloaded in multiple formats. Other systems, like the EMBL-EBI's BioStudies database, aim to archive comprehensive metadata and data from entire biological studies, representing the next step in integrated data management (Tiwary 2022).

Table 1. Entrez databases (Tiwary 2022).

Database	Area	Description
PubMed	Literature	Biomedical abstracts and citations
PubMed central	Literature	Full text articles from journals

Nucleotide	Genomes	DNA and RNA sequences
Genome	Genomes	Genome sequencing projects of different species
Gene	Genes	Detailed information on gene loci
GEO profiles	Genes	Gene expression profiles
HomoloGene	Genes	Homologous genes from different species
Protein	Proteins	Protein sequences
Structure	Proteins	Biomolecular structures
PubChem compound	Chemicals	Chemical information of compounds with structures
Online Mendelian inheritance in man (OMIM)	Genes	A catalogue of human genes and genetic disorders including phenotypes and linkage data
BioProject	Diverse data	Comprehensive collection of research studies including diverse data types
BioSample	Diverse data	A resource of annotated biological samples from diverse studies
LitCovid	Literature	A COVID-19-specific curated literature database

This integrated data architecture directly accelerates scientific discovery. The efficiency of resources like GenBank, Ensembl, and Entrez has fueled a measurable explosion in bioinformatics output. Consequently, PubMed publications in this field have grown dramatically over the past 20 years (**Figure 12**), a trend powered by advances in computational capacity and software capable of leveraging these rich data sources **(Singh and Kumar 2024)**.

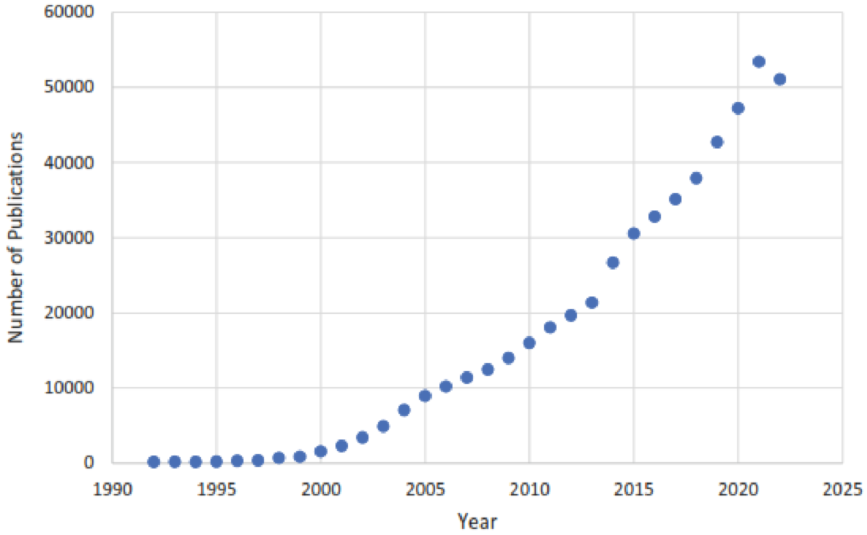


Figure 12. The amount of publications related to bioinformatics in PubMed in the recent 20 years (Singh and Kumar 2024)**.**

2.2. History of Biological Databases

A core goal of biology is to understand the instructions that make life work, which are encoded in DNA and protein sequences. As scientists discovered more and more of these sequences, they faced a new challenge: how to collect, manage, and share this vast amount of information effectively. The solution was the creation of biological databases. These organized collections of data have become essential tools. They allow researchers to use computers to analyze information, creating a powerful link between traditional lab experiments and modern, data-driven discovery (Baxevanis *et al.* 2020).

The first biological database, a protein sequence resource, was established by Margaret Dayhoff in 1965. Dayhoff also pioneered the development of the Point Accepted Mutation (PAM) substitution matrix and introduced the standardized one-letter code for amino acids. In the early 1980s, the EMBL Data Library (now the European Nucleotide Archive, https://www.ebi.ac.uk/ena) initiated a systematic catalog of published biological data. A timeline summarizing key historical milestones in biological databases is presented in **Figure 13** (Villalba and Matte 2021).

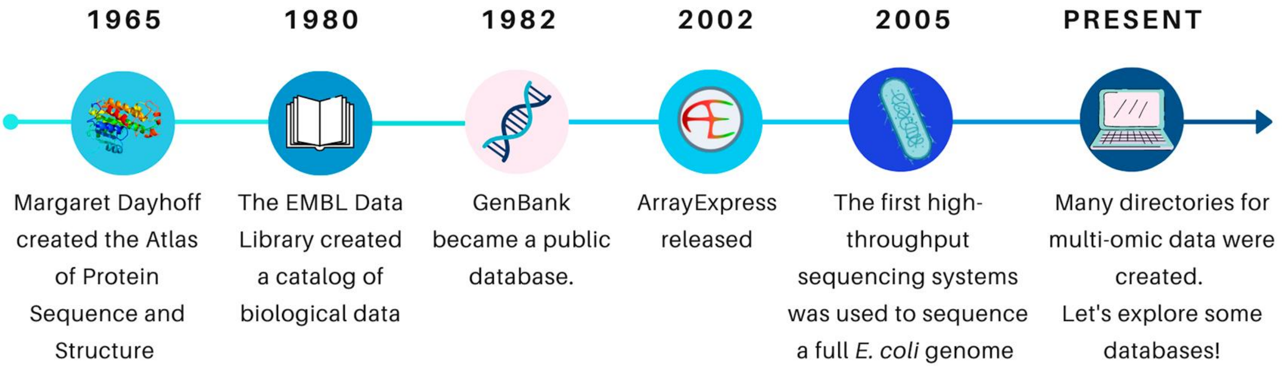

1965
1980
1982
2002
2005
PRESENT
Margaret Dayhoff created the Atlas of Protein Sequence and Structure
The EMBL Data Library created a catalog of biological data
GenBank became a public database.
ArrayExpress released
The first high-throughput sequencing systems was used to sequence a full *E. coli* genome
Many directories for multi-omic data were created.
Let's explore some databases!

Figure 13. A brief history of the biological databases (Villalba and Matte 2021).

2.3. Functional Roles and Applications of Omics Databases

The integration of high-throughput omics technologies—genomics, transcriptomics, proteomics, metabolomics, and epigenomics—has generated vast, complex datasets fundamental to understanding human health and disease. Omics databases serve as the critical infrastructure for storing, organizing, and providing access to this molecular data. They function as centralized hubs that enable researchers to share, cross-reference, and analyze large-scale biological information, thereby facilitating the identification of disease biomarkers, therapeutic targets, and regulatory networks (Vitorino 2024).

The primary utility of these databases lies in their application. Bioinformatic tools and algorithms are designed specifically to interface with these repositories, allowing for sophisticated analysis such as differential gene expression, protein-protein interaction prediction, and metabolic pathway mapping. This synergy between databases and analytical tools is transformative, powering advancements in personalized medicine, drug discovery, and our understanding of disease mechanisms (Ogunjobi *et al.* 2024).

A core principle of these resources is the promotion of open data access and collaboration, often incorporating data from major public initiatives like The Human Genome Project and The Genotype-Tissue Expression (GTEx) Project. This ensures data reproducibility and accelerates scientific discovery. Furthermore, the continuous curation and integration of new data types enhance their value and reliability for the research community.

The indispensable role of omics databases and their interrelationship with analytical tools in driving biomedical research is summarized in **Fig. 14** (Kaithal *et al.* 2024).

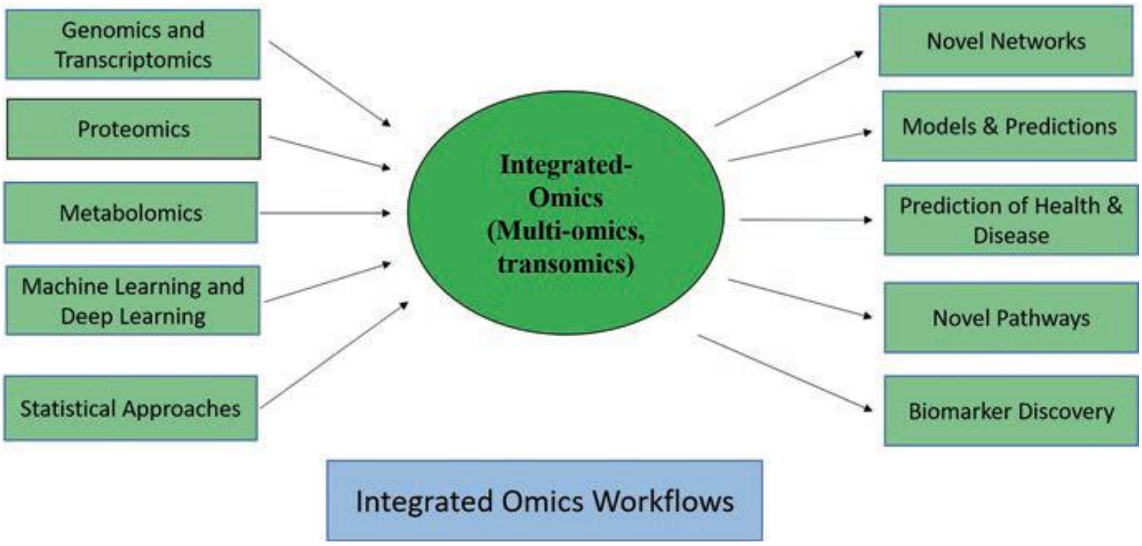

Genomics and Transcriptomics
Proteomics
Metabolomics
Machine Learning and Deep Learning
Statistical Approaches
Integrated-Omics (Multi-omics, transomics)
Novel Networks
Models & Predictions
Prediction of Health & Disease
Novel Pathways
Biomarker Discovery
Integrated Omics Workflows

Figure 14. Integrated omics workflows (Kaithal *et al.* 2024).

2.4. Types of Biological Databases

2.4.1. Sequence and Structure Databases

2.4.1.1. Nucleic Acid Databases

All publicly available DNA and RNA sequences are archived in three major international repositories that form the **International Nucleotide Sequence Database Collaboration (INSDC)**:

- **GenBank** (USA, maintained by NCBI),
- **EMBL-EBI's European Nucleotide Archive (ENA)** (UK), and
- **DDBJ** (the DNA Data Bank of Japan, maintained by the National Institute of Genetics).

These databases are synchronized daily, meaning that a sequence submitted to one is automatically shared with the others. Each sequence is assigned a unique accession number and version, which remain consistent across all three platforms.

Since its establishment in 1982, GenBank (**Figure 15**) has grown exponentially and now contains over 2.1 billion nucleotide sequences. Its current doubling time of approximately 20.8 months is comparable to that predicted by Moore's law. The databases accept direct submissions from researchers and include a wide range of data types such as:

(i) Raw sequencing reads

(ii) Assembled genomes

(iii) Expressed Sequence Tags (ESTs) – short mRNA fragments with high error rates

(iv) Genome Survey Sequences (GSS) – low-coverage genomic fragments often used for gene discovery (Tiwary 2022).

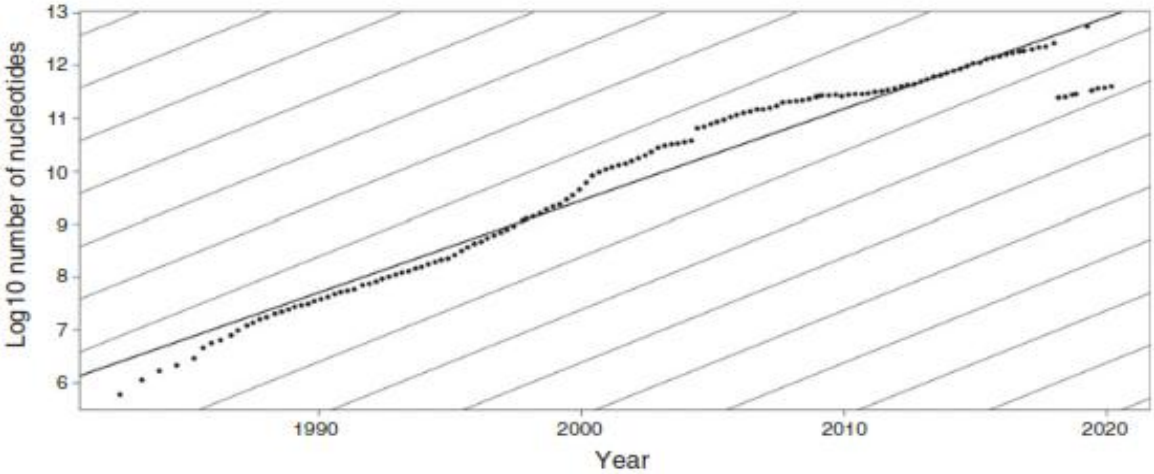
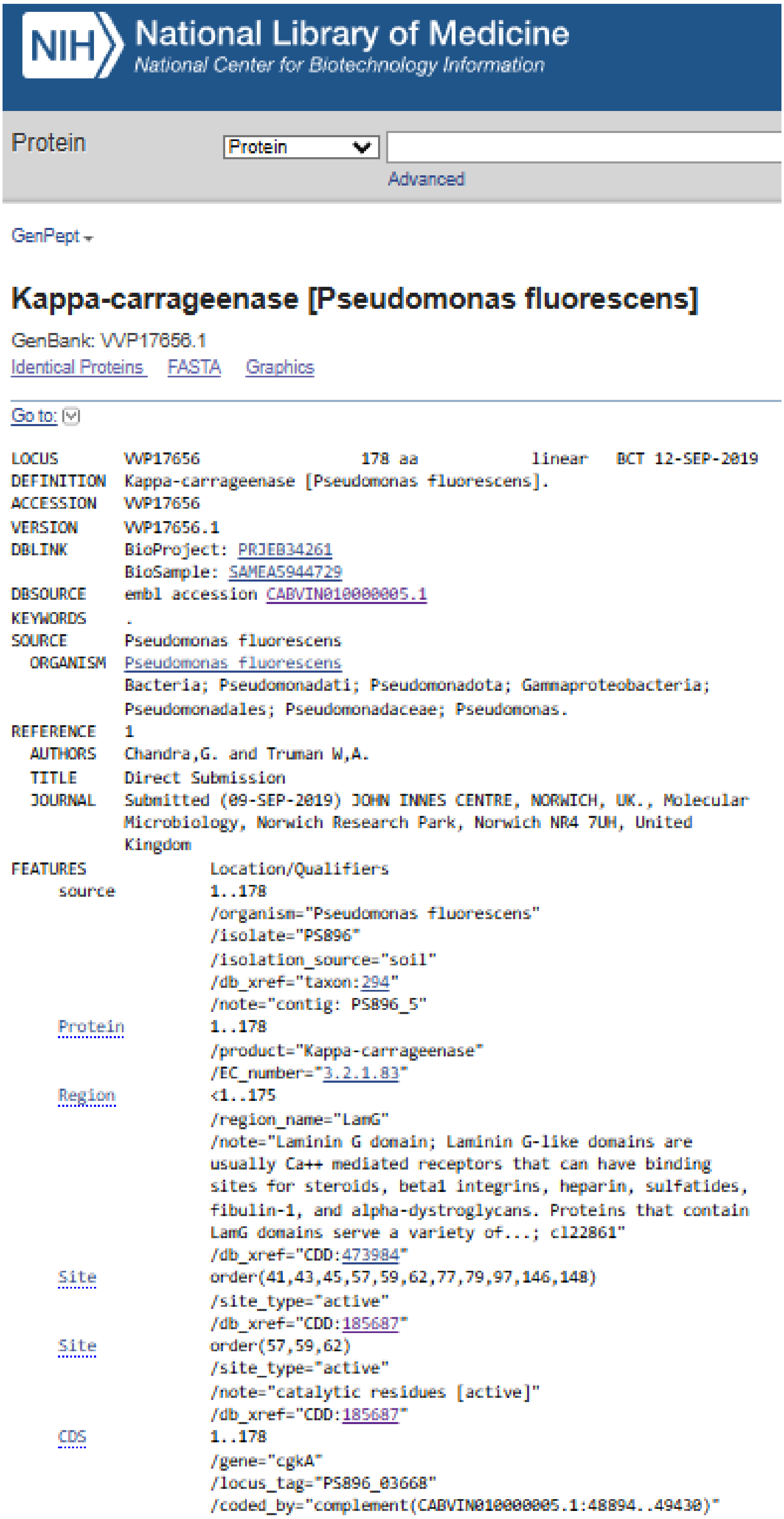


Figure 15. Exponential growth of GenBank. The growth of GenBank sequences from 1982 to 2020, plotted on a logarithmic scale. The central line represents the best-fit observed doubling time of 20.8 months. The outer lines indicate the previously projected 18-month doubling time for comparison (Tiwary 2022).

GenBank is accessed via NCBI’s Entrez system (**Figure 16**), while EMBL and DDBJ are searchable through SRS (Sequence Retrieval System) servers.



NIH National Library of Medicine
National Center for Biotechnology Information

Protein | Protein | Advanced

GenPept

Kappa-carrageenase [Pseudomonas fluorescens]

GenBank: VVP17656.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       VVP17656                 178 aa            linear   BCT 12-SEP-2019
DEFINITION  Kappa-carrageenase [Pseudomonas fluorescens].
ACCESSION   VVP17656
VERSION     VVP17656.1
DBLINK      BioProject: PRJEB34261
            BioSample: SAMEA5944729
DBSOURCE    embl accession CABVIN010000005.1
KEYWORDS    .
SOURCE      Pseudomonas fluorescens
  ORGANISM  Pseudomonas fluorescens
            Bacteria; Pseudomonadati; Pseudomonadota; Gammaproteobacteria;
            Pseudomonadales; Pseudomonadaceae; Pseudomonas.
REFERENCE   1
  AUTHORS   Chandra,G. and Truman W,A.
  TITLE     Direct Submission
  JOURNAL   Submitted (09-SEP-2019) JOHN INNES CENTRE, NORWICH, UK., Molecular
            Microbiology, Norwich Research Park, Norwich NR4 7UH, United
            Kingdom
FEATURES             Location/Qualifiers
     source          1..178
                     /organism="Pseudomonas fluorescens"
                     /isolate="PS896"
                     /isolation_source="soil"
                     /db_xref="taxon:294"
                     /note="contig: PS896_5"
     Protein         1..178
                     /product="Kappa-carrageenase"
                     /EC_number="3.2.1.83"
     Region          <1..175
                     /region_name="LamG"
                     /note="Laminin G domain; Laminin G-like domains are
                     usually Ca++ mediated receptors that can have binding
                     sites for steroids, beta1 integrins, heparin, sulfatides,
                     fibulin-1, and alpha-dystroglycans. Proteins that contain
                     LamG domains serve a variety of...; cl22861"
                     /db_xref="CDD:473984"
     Site            order(41,43,45,57,59,62,77,79,97,146,148)
                     /site_type="active"
                     /db_xref="CDD:185687"
     Site            order(57,59,62)
                     /site_type="active"
                     /note="catalytic residues [active]"
                     /db_xref="CDD:185687"
     CDS             1..178
                     /gene="cgkA"
                     /locus_tag="PS896_03668"
                     /coded_by="complement(CABVIN010000005.1:48894..49430)"
```

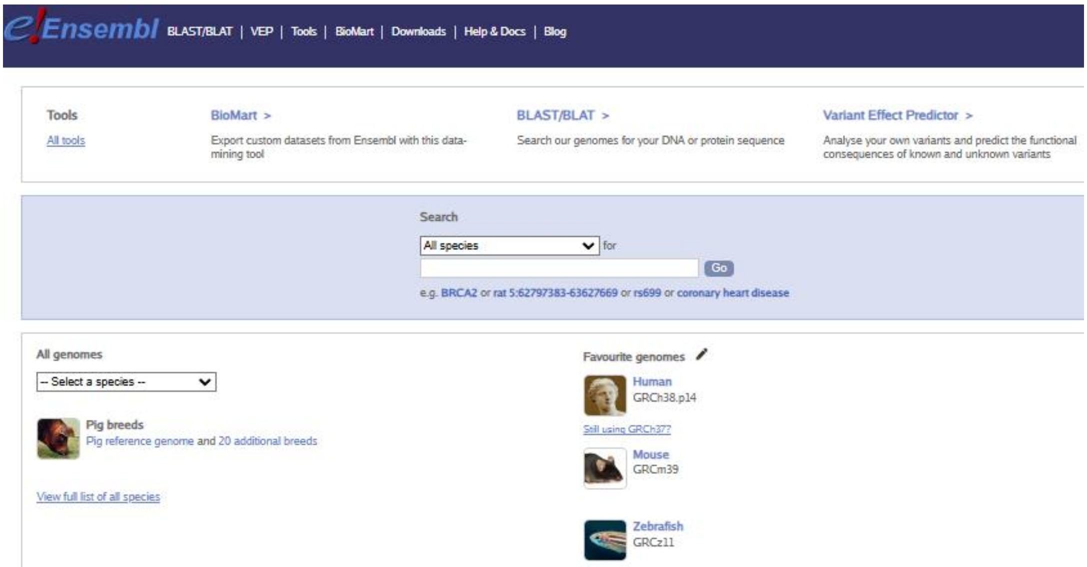
Figure 16. Partial screenshot of GenBank home page showing information on Kappa-carrageenase from *Pseudomonas fluorescens*.

In addition to these core repositories, several specialized databases support more focused aspects of nucleotide sequence analysis, such as genome annotation, variant detection, and non-coding RNA research. These databases include resources like RefSeq**,** dbSNP**,** Ensembl**,** UCSC Genome Browser**,** RNAcentral**,** DIANA-TarBase, and others. To help researchers quickly identify the purpose and access points of these databases, a summary of the most important platforms is provided in **Table 2** (Tiwary 2022).

Table 2. Nucleotide databases (Tiwary 2022).

Database	**Description**	**URL**
NCBI GenBank	Genetic sequence database	www.ncbi.nlm.nih.gov/genbank
Ensembl	A genome browser of vertebrates	www.ensembl.org
NCBI Refseq	A collection of non-redundant and well-annotated genomic sequences, transcripts and proteins	www.ncbi.nlm.nih.gov/refseq
UCSC genome browser	Interactive genome visualization browser	https://genome.ucsc.edu/
1000 genomes	A catalogue of human genomic variation	www.internationalgenome.org
GeneCards	An integrative database of predicted and annotated human genes	www.genecards.org
lncRNAdb	Database containing annotated long non-coding RNAs in eukaryotes	https://ngdc.cncb.ac.cn/databasecommons/database/id/23
miRBase	Database of published miRNA sequences along with annotation	www.mirbase.org
DIANATarBase	A database of experimentally supported miRNA targets	www.microrna.gr/tarbase

For instance, RefSeq provides curated reference sequences, particularly for human biomedical studies, while dbSNP catalogs known human genetic variations. In terms of genome annotation, Ensembl (**Figure 17**) offers an integrated platform for exploring gene structures, genomic variants, and regulatory elements across more than 227 vertebrate and model species. It also supports comparative genomics and features a dedicated browser for tracking SARS-CoV-2 genomic mutations (Goldfarb *et al.* 2025).



Export custom datasets from Ensembl with this data-mining tool

Search our genomes for your DNA or protein sequence

Analyse your own variants and predict the functional consequences of known and unknown variants

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

Pig reference genome and 20 additional breeds

Figure 17. Partial Screenshot of Ensembl genome database.

The UCSC Genome Browser (**Fig. 18**) is a web tool for viewing and studying genomic data. It includes information from over 4,000 different organisms. Since its launch in 2001, it has become a vital resource for genetics and bioinformatics research (Perez *et al.* 2025).

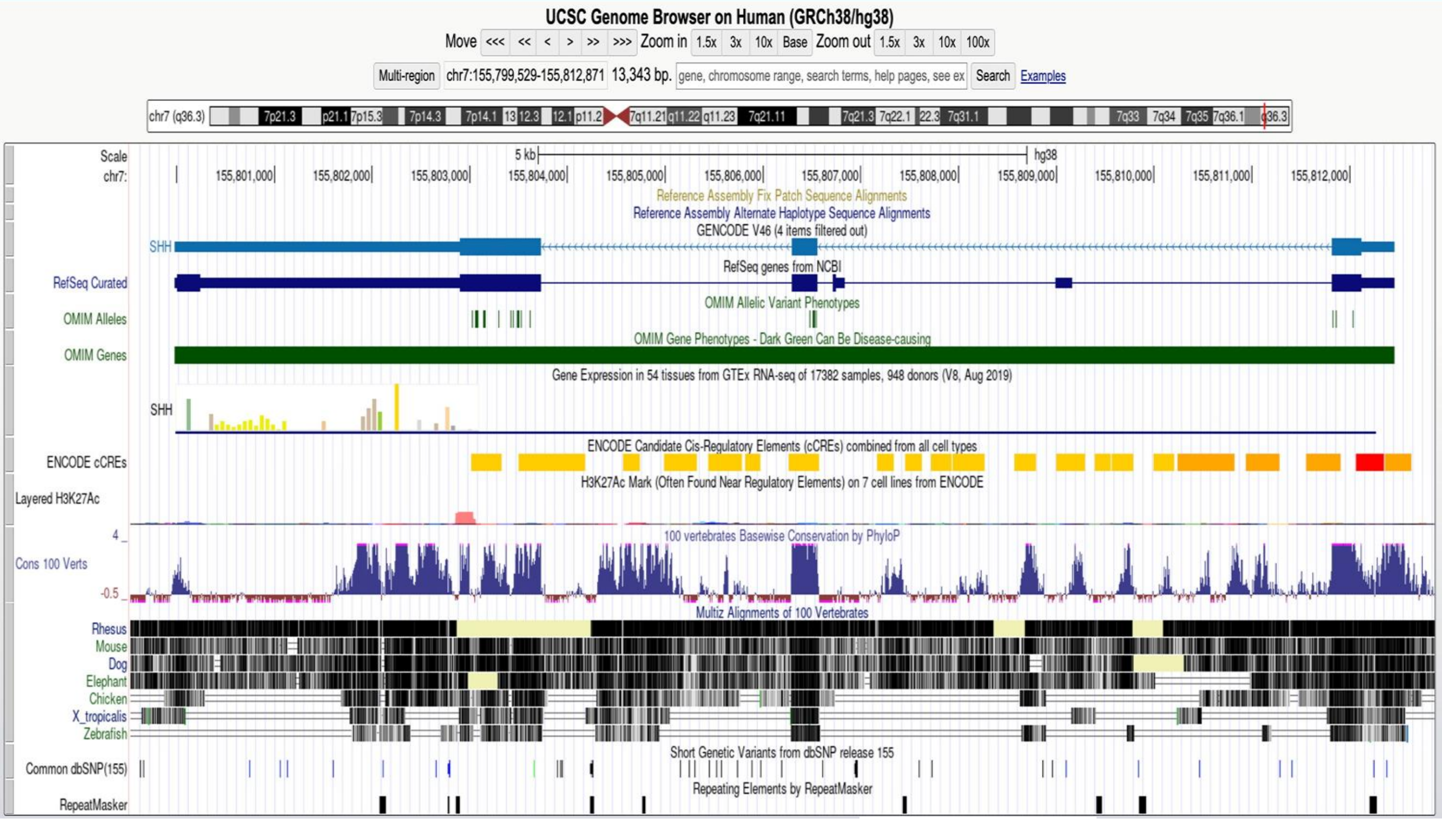

UCSC Genome Browser on Human (GRCh38/hg38)
Move <<< << < > >> >>> Zoom in 1.5x 3x 10x Base Zoom out 1.5x 3x 10x 100x
Multi-region chr7:155,799,529-155,812,871 13,343 bp. gene, chromosome range, search terms, help pages, see ex Search Examples
chr7 (q36.3) 7p21.3 p21.1 7p15.3 7p14.3 7p14.1 13 12.3 12.1 p11.2 7q11.21 q11.22 q11.23 7q21.11 7q21.3 7q22.1 22.3 7q31.1 7q33 7q34 7q35 7q36.1 q36.3
Scale
chr7:
5 kb
hg38
155,801,000 155,802,000 155,803,000 155,804,000 155,805,000 155,806,000 155,807,000 155,808,000 155,809,000 155,810,000 155,811,000 155,812,000
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V46 (4 items filtered out)
SHH
RefSeq genes from NCBI
RefSeq Curated
OMIM Allelic Variant Phenotypes
OMIM Alleles
OMIM Gene Phenotypes - Dark Green Can Be Disease-causing
OMIM Genes
Gene Expression in 54 tissues from GTEx RNA-seq of 17382 samples, 948 donors (V8, Aug 2019)
SHH
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
ENCODE cCREs
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
4
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
-0.5
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Short Genetic Variants from dbSNP release 155
Common dbSNP(155)
Repeating Elements by RepeatMasker
RepeatMasker

Figure 18. Partial view of UCSC Genome browser (Perez *et al.* 2025).

In addition, RNAcentral aggregates non-coding RNA sequences, including miRNAs and lncRNAs, providing unified access across multiple species and sources. DIANA-TarBase serves as a reference for experimentally validated miRNA–gene interactions, offering insights into gene regulation at the post-transcriptional level (Allmer 2023). Finally, high-throughput sequencing data are accessible through the ENA and the DDBJ Sequence Read Archive (DRA), which support both raw read datasets and assembled genomes derived from next-generation sequencing (NGS) platforms (Jain *et al.* 2024).

2.4.1.2. Protein Databases

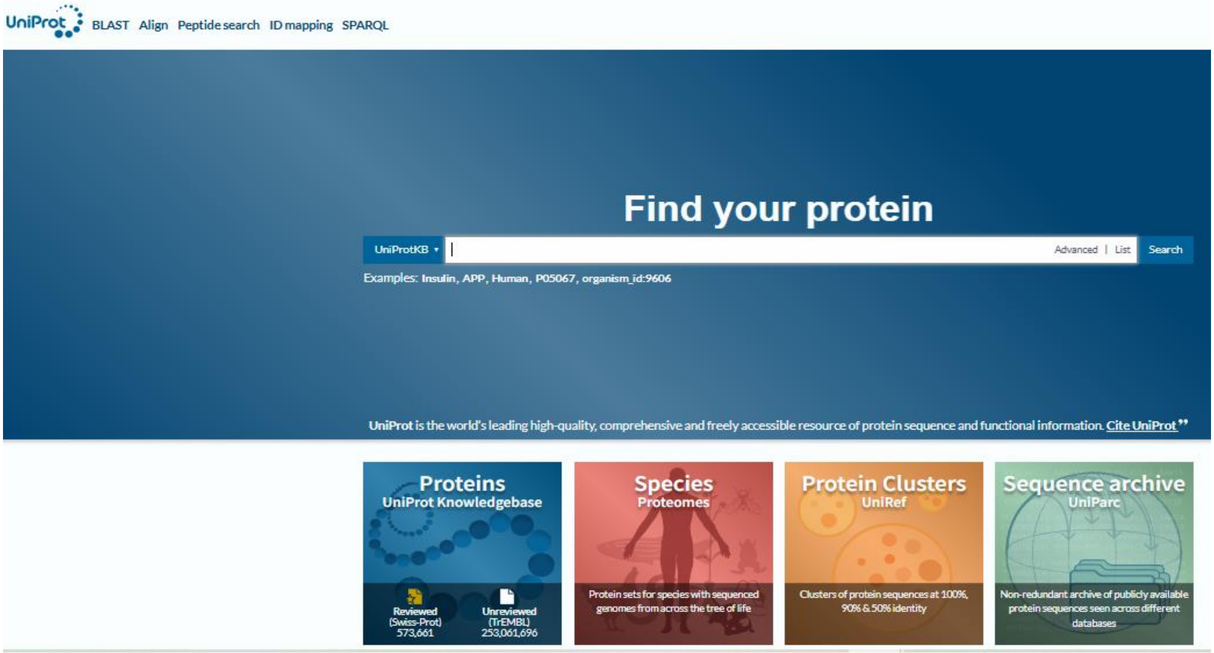
Protein sequences found in specialized databases (see **Table 3**) originate from multiple experimental and computational sources. Direct protein sequencing methods such as Edman degradation and mass spectrometry provide foundational data, while three-dimensional structures from X-ray crystallography and NMR spectroscopy offer additional sequence validation. A substantial proportion of protein data is derived computationally through translation of coding regions from DNA and RNA sequences deposited in nucleotide databases (Tiwary 2022).

Table 3. Protein databases (Tiwary 2022).

Database	Description	URL
UniProt	A public database of protein sequences with their functional information	www.uniprot.org
CATH-Gene3D	Database of protein classification and prediction of domain structure	www.cathdb.info
GenPept	Translated coding sequences from GenBank	www.ncbi.nlm.nih.gov/protein
DIP	A database of experimentally determined protein–protein interactions	http://dip.doembi.ucla.edu
HPRD	Integrated platform depicting various information regarding human proteome	www.hprd.org
InterPro	Integrated database providing functional analysis of proteins	www.ebi.ac.uk/interpro
ModBase	Database containing theoretically calculated protein structure models	https://modbase.compbio.ucsf.edu/
RCSB protein data Bank	Public database containing	www.rcsb.org

(PDB)	experimentally determined protein and other macromolecule structures	
Pfam	Database of large collection of protein families	http://pfam.xfam.org/
PROSITE	Database of protein domains, families and functional sites	https://prosite.expasy.org/
ProteomicsDB	A multi-omics database including proteomics, transcriptomics and cell line viability data	https://www.proteomicsdb.org/
CoV3D	A database of coronavirus protein structures	https://cov3d.ibbr.umd.edu/
STRING	A database of protein–protein interactions	https://string-db.org/

The Universal Protein Resource (UniProt, see **Fig. 19**) represents the central hub for protein sequence information, integrating data from multiple sources into a unified knowledgebase. UniProt Knowledgebase (UniProtKB) contains over 189 million sequences ranging from experimentally verified proteins to computationally predicted translations. The database's sophisticated annotation system includes functional descriptions, domain architectures, post-translational modifications, and cross-references to specialized resources. UniProt operates through two complementary divisions: Swiss-Prot, featuring expert manual curation with literature-derived annotations, and TrEMBL, providing automated annotations for sequences awaiting full curation (Consortium 2024).



Find your protein

Examples: Insulin, APP, Human, P05067, organism_id:9606

UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. Cite UniProt

Protein sets for species with sequenced genomes from across the tree of life

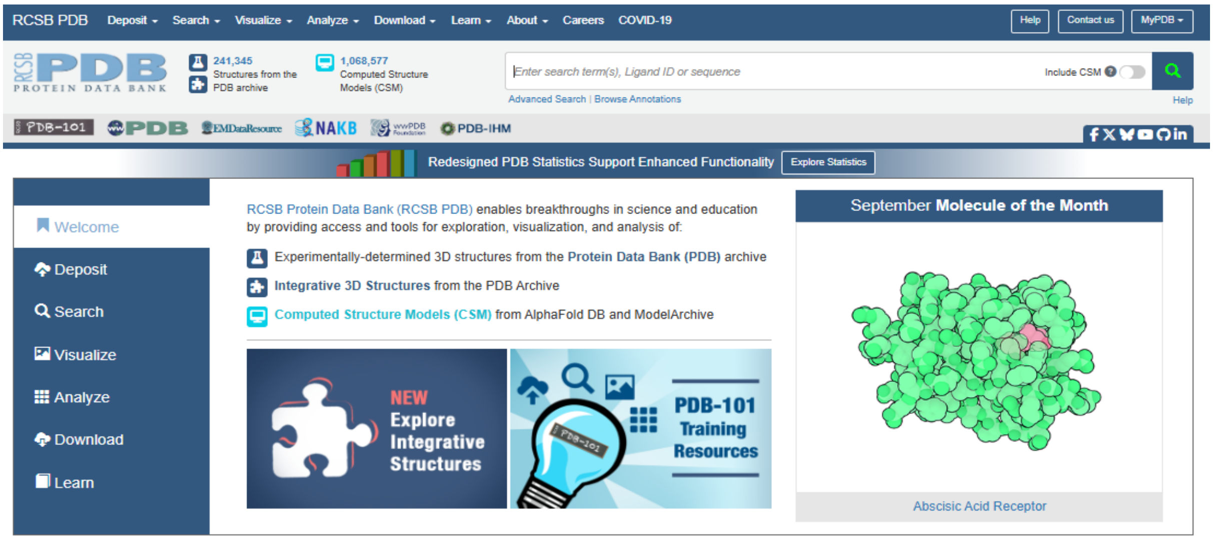
Clusters of protein sequences at 100%, 90% & 50% identity

Non-redundant archive of publicly available protein sequences seen across different databases

Figure 19. Partial screenshot of UniProt home page.

The protein database ecosystem extends beyond UniProt to include several specialized resources. CATH-Gene3D provides structural classification of protein domains into evolutionary superfamilies, enabling functional inferences through structural comparisons. GenPept and DAD serve as complementary resources providing automated translations of coding sequences from GenBank and DDBJ respectively (Tiwary 2022). For interaction studies, the Database of Interacting Proteins (DIP) documents experimentally verified protein-protein interactions, while the Human Protein Reference Database (HPRD) offers integrated information on human proteins including domain architecture, post-translational modifications, and disease associations. Pfam remains the primary resource for protein family and domain classification using hidden Markov models (Hollander *et al.* 2021).

Structural biology depends heavily on the Protein Data Bank (PDB), an international consortium maintaining the global archive for three-dimensional macromolecular structures (see **Fig. 20** for organizational structure). This resource, established in 1971, now contains over 150,000 structures determined primarily by X-ray crystallography, NMR spectroscopy, and increasingly by cryo-electron microscopy. The PDB continues growing at a rate of approximately 12,000 new structures annually, providing essential data for understanding protein function and facilitating drug discovery (Berman and Burley 2025).



241,345 Structures from the PDB archive

1,068,577 Computed Structure Models (CSM)

Enter search term(s), Ligand ID or sequence

Redesigned PDB Statistics Support Enhanced Functionality

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:

- Experimentally-determined 3D structures from the Protein Data Bank (PDB) archive
- Integrative 3D Structures from the PDB Archive
- Computed Structure Models (CSM) from AlphaFold DB and ModelArchive

Figure 20. A partial screenshot of RCSB-PDB home page.

Modern proteomics research relies on resources like ProteomicsDB, which integrates mass spectrometry-based quantitative data across human tissues, cell lines, and body fluids. This platform enables real-time exploration of protein abundance patterns and visualization of drug-target interactions (Jiang *et al.* 2025).

For interaction network analysis, the STRING database provides comprehensive coverage of both known and predicted protein-protein interactions, incorporating physical interactions, functional associations, and pathway information across more than 14,000 organisms. The continuous development of these resources reflects the ongoing expansion of proteomic data and its crucial role in biomedical research (Szklarczyk *et al.* 2025).

2.4.2. Expression Databases

Several key databases are dedicated to storing gene expression data for functional genomics studies (summarized in **Table 4**). The two primary public repositories for data from microarray and RNA-Seq experiments are the Gene Expression Omnibus (GEO) at the NCBI and ArrayExpress at the EBI. A third major repository, the Genomic Expression Archive (GEA), is hosted by the DDBJ (Tiwary 2022).

These core repositories follow international data standards, ensuring consistency and reliability. In addition to these large archives, several specialized databases provide curated, analysis-ready information. These include resources like the Expression Atlas for gene and protein expression across species and conditions, the Human Protein Atlas for the spatial localization of proteins in human tissues, and Oncomine for cancer-specific expression profiles (Wolde *et al.* 2025).

Other specialized resources focus on areas such as Tissue-specific Gene Expression and Regulation (TiGER), host-pathogen interactions during infection (DualSeqDB), and the expression of long non-coding RNAs (LncExpDB) (Tiwary 2022).

Table 4. Expression databases (Tiwary 2022).

Database	**Description**	**URL**
NCBI GEO	Gene expression data from microarray and sequencing	https://www.ncbi.nlm.nih.gov/geo/
EBI ArrayExpress	Gene expression data from microarray and sequencing	https://www.ebi.ac.uk/biostudies/arrayexpress
Expression atlas	Gene and protein expression data from various species and conditions	https://www.ebi.ac.uk/gxa/home
Gene expression archive	Gene expression data from microarray and sequencing	https://www.ddbj.nig.ac.jp/gea/index-e.html
Human protein atlas	A map of all human proteins in cells, tissues and organs	https://www.proteinatlas.org/
ONCOMINE	Gene expression and sample data from different cancer types	https://www.oncomine.com/
TiGER	Tissue-specific gene expression profile and gene regulation data	http://bioinfo.wilmer.jhu.edu/tiger/
DualSeqDB	A host-bacterial pathogen RNA-sequencing database having combined gene expression data during infection process	https://dualseqdb.tartaglialab.com/
LncExpDB	An expression database of human long non-coding RNAs	https://bigd.big.ac.cn/lncexpdb

2.4.3. Pathway Databases

Pathway databases provide organized maps that connect various biological molecules and their interactions, helping researchers understand complex biological processes (summarized in **Table 5**). These resources are essential for interpreting omics data and modeling cellular functions (Tiwary 2022).

Key resources include comprehensive databases like KEGG and MetaCyc, which catalog metabolic pathways across all organisms. Specialized databases include Reactome for human biological pathways, HMDB for human metabolites, and PathBank for pathway information in model organisms (Hasan *et al.* 2025).

Additional resources like BiGG Models provide computational models of metabolism, Plant Reactome focuses on plant pathways, and Ingenuity Pathway Analysis (IPA) offers commercial pathway analysis tools (Hasan *et al.* 2025).

Table 5. Metabolic pathway databases (Tiwary 2022).

Database	Description	URL
KEGG	Database of biological systems including drugs and diseases	https://www.genome.jp/kegg/
MetaCyc	Reference database of metabolic pathways	https://metacyc.org/
PathBank	Metabolic and signalling pathway database of model organisms	https://pathbank.org/
Reactome	Curated pathway database	https://reactome.org/
HMDB	Human metabolome database	https://www.hmdb.ca/
BiGG models knowledge base	A database of genome-scale metabolic models	http://bigg.ucsd.edu/
Plant Reactome	Comparative plant pathway database	https://www.gramene.org/pathways
Ingenuity pathway analysis (IPA)	Commercial pathway database developed by Qiagen	https://www.qiagen.com/us/products/

2.4.4. Disease Databases

2.4.4.1. Overview of Disease Databases

Disease databases represent a critical infrastructure for biomedical research, consolidating curated information on genetic diseases, molecular mechanisms, and associated phenotypes. Key resources include the Online Mendelian Inheritance in Man (OMIM), which offers comprehensive details on human genes and genetic disorders; The Cancer Genome Atlas (TCGA), hosting multi-omics data across 33 cancer types; and the Human Gene

Mutation Database (HGMD), cataloging published disease-causing mutations (Pettini *et al.* 2021). Other notable databases encompass GWAS Central for genome-wide association studies, canSAR for integrative cancer research and drug discovery, and DisGeNET for gene-disease associations. Specialized platforms such as HbVar (hemoglobinopathies), miR2Disease (microRNA-disease interactions), CNCDatabase (non-coding cancer drivers), and PAGER-CoV (COVID-19 genomics) address niche research domains (Beck *et al.* 2020; Hu *et al.* 2025).

The breadth and applications of these resources are summarized in **Table 6**, highlighting their pivotal role in advancing mechanistic insights and therapeutic development.

Table 6. Disease databases (Tiwary 2022).

Database	**Description**	**URL**
Online Mendelian Inheritance in Man (OMIM)	Comprehensive compendium of human genes and genetic phenotypes	https://www.omim.org/
The Cancer genome atlas (TCGA)	A cancer database having genomic, transcriptomic, proteomic and epigenomic data	https://www.cbioportal.org/
Human gene mutation database (HGMD)	Database of all published gene mutations involved in human inherited diseases	https://www.hgmd.cf.ac.uk/
GWAS central	Comprehensive repository of genome-wide association study data	https://mart.gwascentral.org/
HbVar	Database of human haemoglobin variants and thalassemia mutations	https://globin.bx.psu.edu/hbvar/menu.html
MalaCards	Integrated database of human diseases	https://www.malacards.org/
miR2Disease	A comprehensive database on miRNA-disease relationship	http://www.mir2disease.org/
DisGeNET	Database of genes and variants associated with human diseases	https://disgenet.com/

STAB	A cell atlas providing cellular landscape of human brain and neuropsychiatric diseases	https://compsysbio.org/
canSAR	A knowledge base for cancer translational research and drug discovery	https://cansar.ai/
CNCDatabase	Cornell non-coding Cancer driver database is a manually curated database regarding non-coding cancer drivers	https://cncdatabase.med.cornell.edu/
PAGER-CoV	Pathways, annotated gene-lists and gene signatures electronic repository for Corona virus	http://discovery.informatics.uab.edu/PAGER-CoV/

2.4.4.2. Cancer Pharmacoinformatics: A Case Study

A diverse array of cancer databases and data analysis tools exists, with only a subset being directly applicable to anti-cancer drug discovery, though all contribute broadly to health science. These resources can be broadly categorized as follows (**Fig. 21**):

(i) cancer multi-omics and immunology databases (encompassing genomics, transcriptomics, proteomics, and metabolomics);

(ii) cancer drug discovery databases; and

(iii) cancer therapy response predictors, among others. This article highlights these three key categories, with each subsequent section detailing their conceptual foundation, a curated list of representative databases, and major bioinformatics resources (Paananen and Fortino 2020).

Advances in high-throughput technologies—including next-generation sequencing, microarrays, and mass spectrometry—have enabled the generation of proteomic, transcriptomic, genomic, and other omics data at unprecedented resolution (Vitorino 2024).

These data are critical for deciphering the molecular mechanisms of cancer and for informing therapeutic strategies. Omics databases support the identification of novel drug targets, elucidation of drug mechanisms of action, and assessment of treatment-related toxicities. Furthermore, such resources are foundational to the development of precision medicine, allowing clinicians to correlate genetic variants with drug efficacy or toxicity in molecularly stratified patient cohorts (Vitorino 2024).

Systematic integration of disease molecular profiles and therapy-driven omics data has the potential to significantly accelerate the drug discovery and development pipeline (Kamble *et al.* 2024).

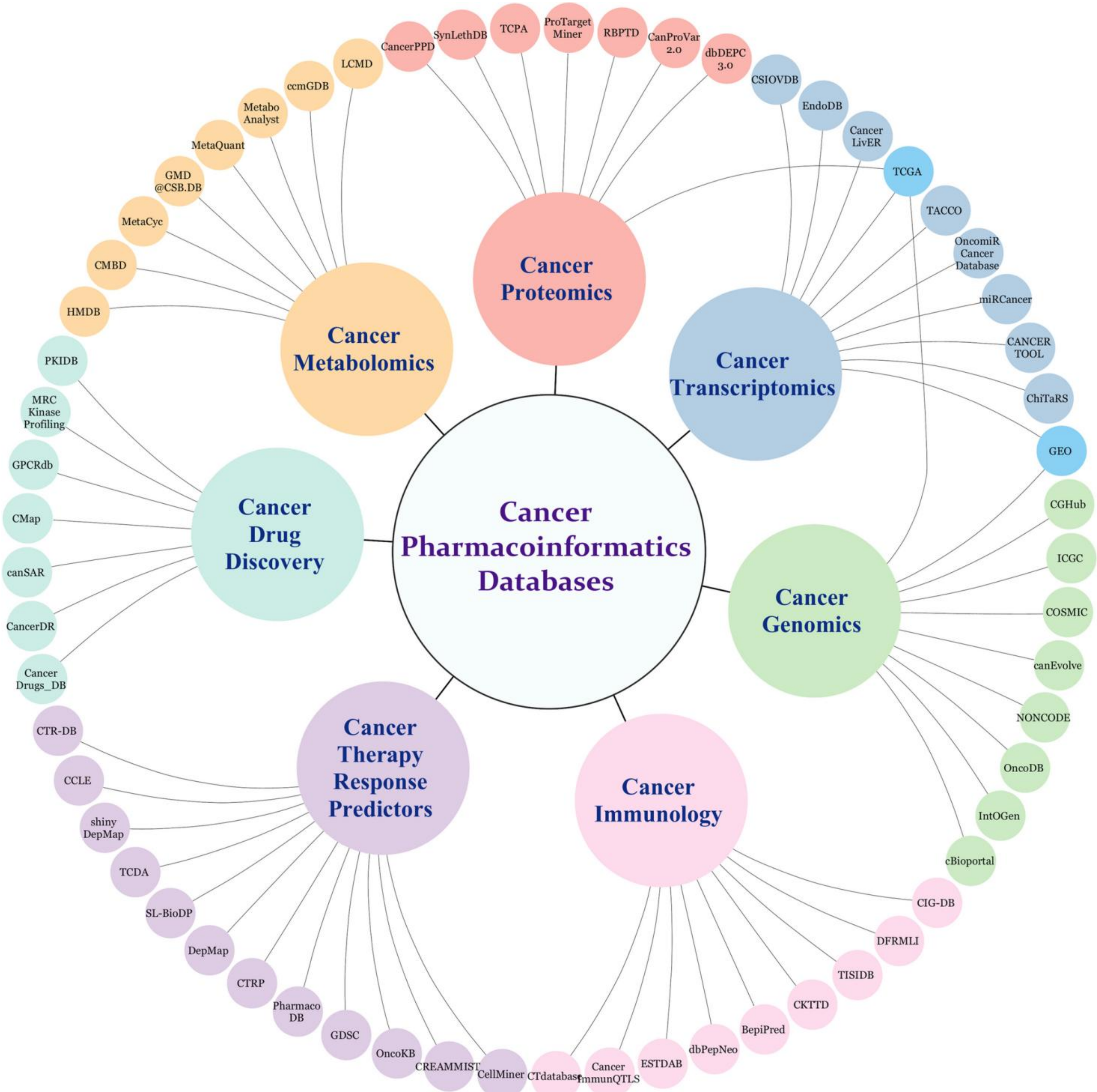


Figure 21. Overview of cancer pharmacoinformatics databases (Kamble *et al.* 2024**).**

Most cancer omics databases and tools are highly interconnected, frequently sharing foundational data sources. To visualize these relationships within cancer pharmacoinformatics, a network was constructed using Cytoscape. Core data repositories—such as TCGA, ICGC, COSMIC, CPTAC, GEO, GTEx, and scientific literature—form the central blue nodes, supplying raw multi-omics data for anti-cancer research (Kamble *et al.* 2024) (**Fig. 22**).

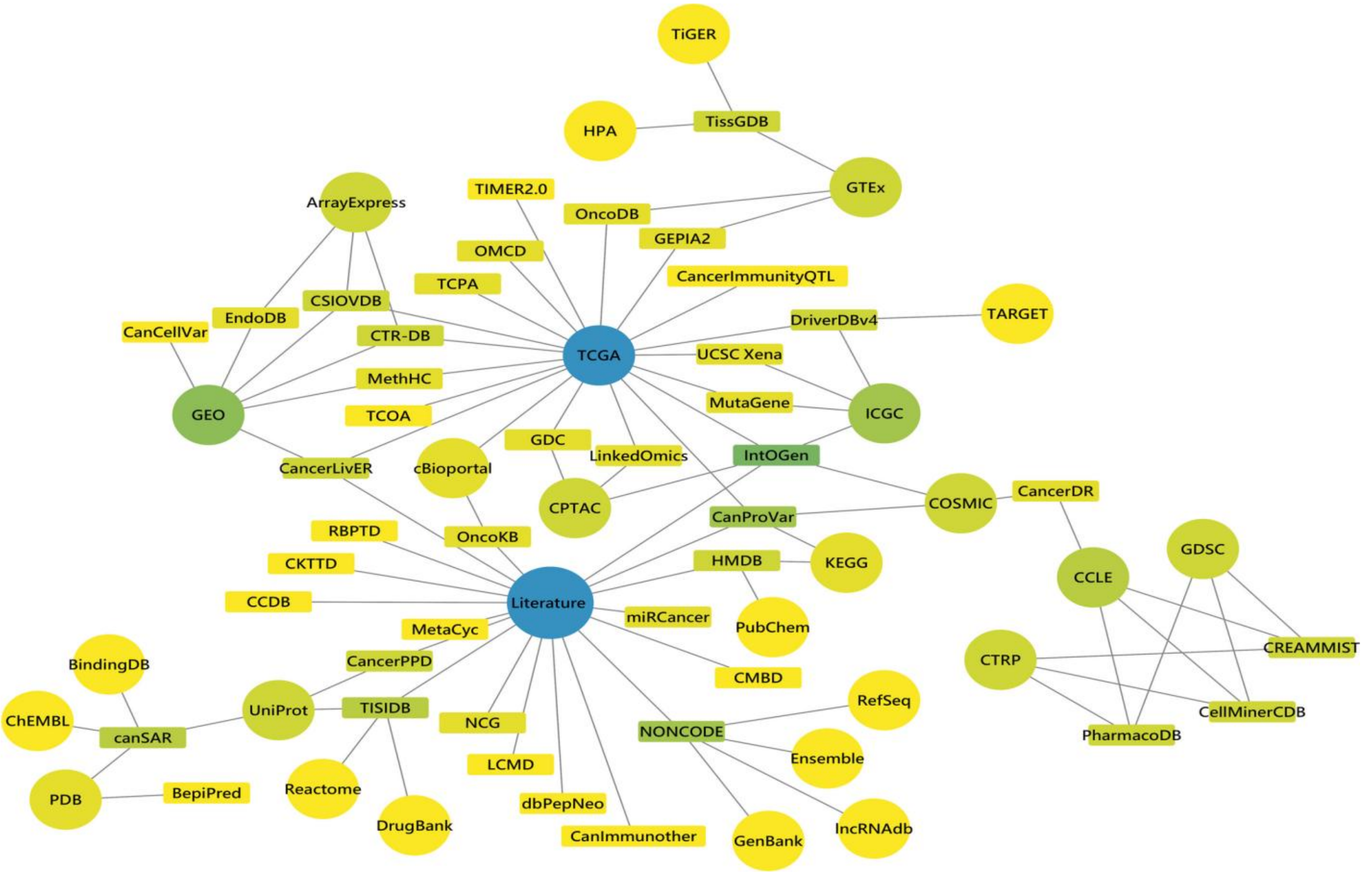


Figure 22. Interdependency network of cancer pharmacoinformatics databases and tools. The network maps the relationships between foundational data sources (circular nodes) and the computational tools (rectangular nodes) developed from them. Edges connect databases to tools that utilize their data. Node color indicates the number of tools derived from each source, ranging from green (most tools) to yellow (fewer tools). The blue highlighting of TCGA and literature nodes denotes their role as the primary sources for tool development in the field (Kamble *et al.* 2024).

Downstream platforms like cBioPortal, GDC, and the UCSC Xena Browser integrate primary data for analysis and visualization. Tools such as GEPIA2 and LinkedOmics specialize in TCGA data exploration, while cBioPortal enables multi-omics visualization across sources including TCGA and ICGC (Kamble *et al.* 2024). In drug discovery, knowledgebases like canSAR merge chemical and protein data from ChEMBL, PDB, and UniProt. Resources including GDSC, CTRP, and CCLE offer cell line drug sensitivity data, supporting predictive tools such as PharmacoDB and CREAMMIST for therapy response modeling. This illustrates a flow from data generation to application (Kamble *et al.* 2024).

2.4.5. Organism-Specific and Virus Databases

Scientists have created specialized databases for the genes of important animals, plants, and viruses. You can find a list of some key examples in **Table 7**. These resources are essential tools for modern biological research, agriculture, and medicine (Tiwary 2022).

Table 7. Organism-specific and Virus databases (Tiwary 2022).

Database	Description	URL
WormBase	Database of experimental data of nematode *C. elegans*	https://wormbase.org/
SilkDB	Database of silkworm genome	https://silkdb.bioinfotoolkits.net/main/species-info/-1
MBKbase-rice	Integrated omics database of rice	https://www.mbkbase.org/rice
Bovine genome database (BGD)	Bovine genomics database	https://bovinegenome.elsiklab.missouri.edu/
Pig genome database (PGD)	Pig genomics database	www.animalgenome.org/pig/genome/db
Zebrafish information network (ZFIN)	Zebrafish genomics database	https://zfin.org/
GISAID	A global database of influenza viruses and SARS-CoV-2 virus	www.gisaid.org
GESS	A database of global evaluation of SARSCoV-2/hCoV-19 sequences	https://ngdc.cncb.ac.cn/databasecommons/database/id/7279

For example, the Alliance of Genome Resources (https://www.alliancegenome.org/) brings together genetic data from many species used to study human biology. One member, WormBase, contains a vast amount of information on nematode worms, compiled from over 1,400 labs worldwide. In agriculture, the silkworm is crucial for making silk. Its complete genetic blueprint is available in the SilkDB database. Similarly, the MBKbase is a knowledge bank for crop breeding, holding detailed genetic information for rice, soybean, wheat, and maize (Tiwary 2022).

There are also important databases for farm animals. The Bovine Genome Database (BGD) lets scientists explore and analyze cow genes. The Pig Genome Database (PGD) is a rich source of information on pig genetics, including traits linked to specific genes. For research, the zebrafish is a common model animal, and its genetic data is stored in the Zebrafish Information Network (ZFIN) (Swargam and Kumari 2023).

Understanding the genes of viruses and bacteria that cause disease is critical for public health. Genomic databases help us track outbreaks and understand how pathogens evolve. For instance, during the COVID-19 pandemic, the GISAID database shared the genetic sequences of the SARS-CoV-2 virus with researchers globally. Scientists used tools like GESS to analyze these sequences and track new variants as the virus spread. Furthermore, knowing the 3D shape of viral proteins is key to designing drugs and vaccines. The CoV3D database provides detailed models of coronavirus proteins (Cheng *et al.* 2023). These essential resources, which include the examples mentioned above, can be systematically organized by their function into a cohesive workflow, as shown in **Figure 23**. This framework divides the tools into four key categories: databases for storing information, annotation tools for labeling genetic features, genomic analysis platforms for comparing sequences, and variant tracking systems for monitoring the global spread of new strains (Cheng *et al.* 2023).

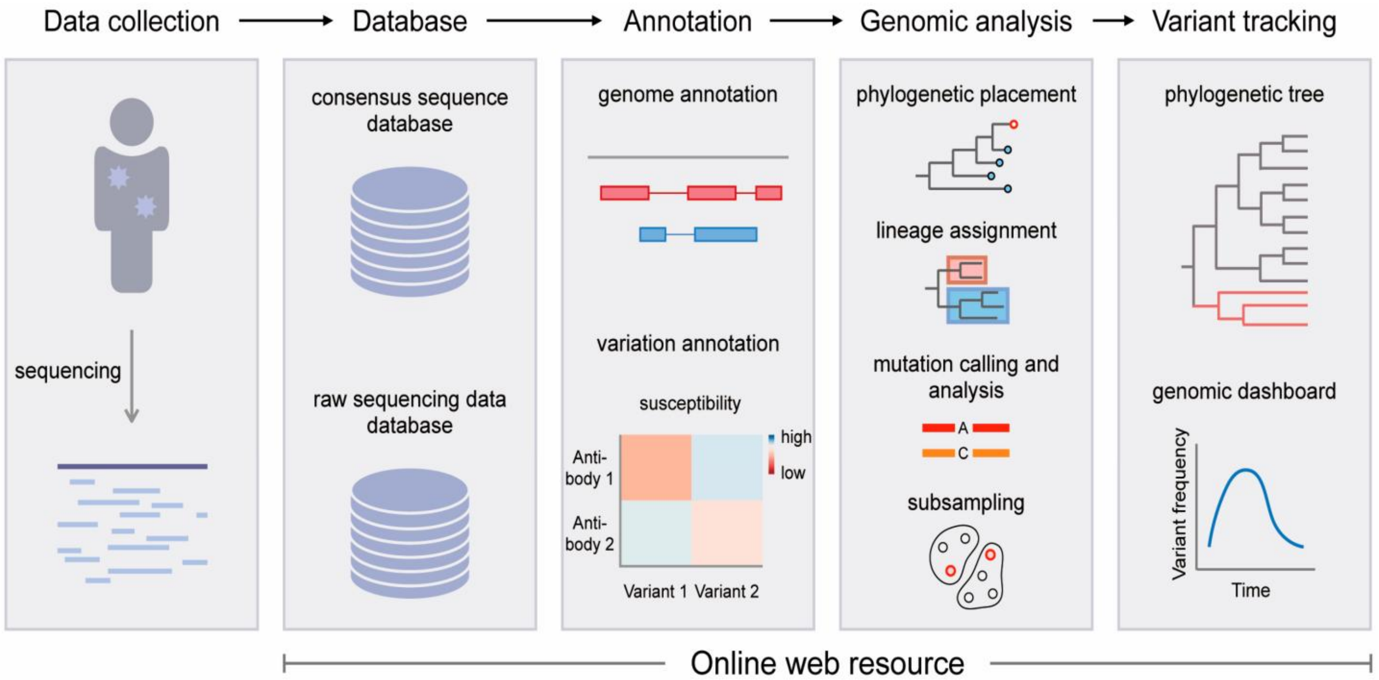
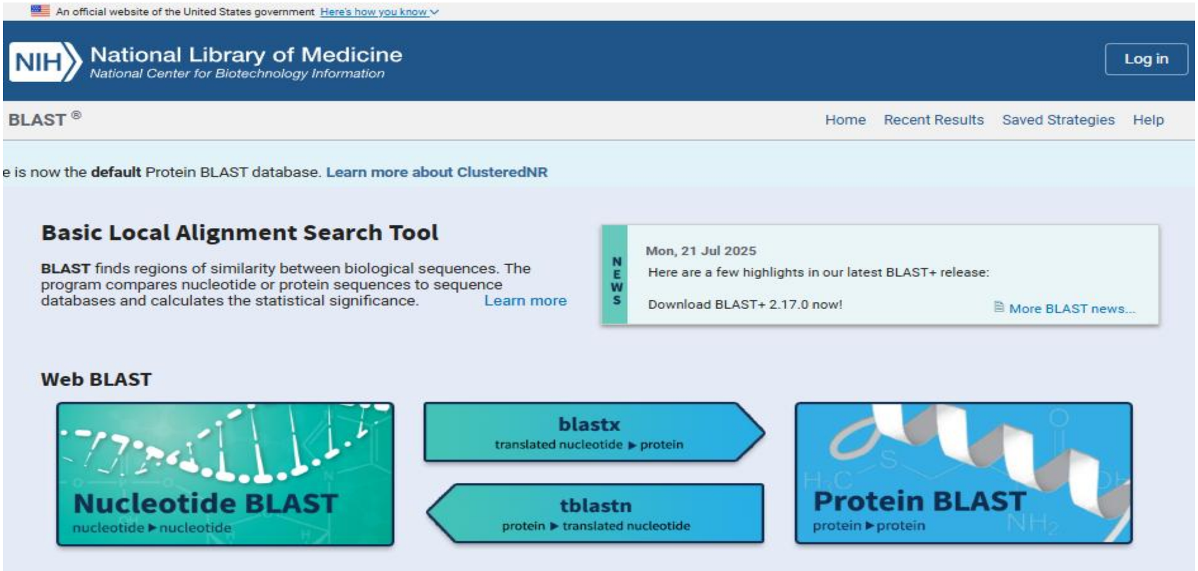


Figure 23. A Step-by-Step Plan for Studying the SARS-CoV-2 Genome Using Free Online Tools. The online tools for this research fall into four main groups: (1) databases to get the data, (2) annotation tools to label the virus's parts, (3) analysis tools to compare genomes, and (4) variant trackers to monitor new virus strains (Cheng *et al.* 2023).

2.5. Database Searching

To find specific information in biological databases, scientists use specialized search tools. The most common program for this is the Basic Local Alignment Search Tool (BLAST). Developed by the NCBI, BLAST acts like a "search engine for biological sequences," allowing researchers to take a single DNA or protein sequence (the "query") and quickly find similar sequences in a massive database (Tiwary 2022). Other tools also exist for sequence comparison. FASTA is a similar search program whose name is also used for a standard sequence file format. BLAT is another alignment tool optimized for finding very similar sequences (Lu *et al.* 2024). A typical BLAST search involves three key steps: providing an input sequence, selecting the appropriate target database, and choosing the correct BLAST program (Tiwary 2022).

This process is facilitated by a web interface, as shown in the **figure 24** below. The specific BLAST program to use and the detailed interpretation of the results, including statistical scores, are covered in the chapter on sequence alignment and comparison.



An official website of the United States government Here's how you know

National Library of Medicine
National Center for Biotechnology Information

e is now the **default** Protein BLAST database. Learn more about ClusteredNR

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

Mon, 21 Jul 2025

Here are a few highlights in our latest BLAST+ release:

Download BLAST+ 2.17.0 now!

Web BLAST

Figure 24. Partial screenshot of NCBI BLAST web interface.

2.6. Integration of Biological Databases

The integration of biological databases is critical for advancing research in the life sciences. As biological data continues to grow exponentially, researchers increasingly rely on the combined use of multiple databases to obtain a holistic understanding of complex biological systems. Integrated databases enable connections across various types of biological information, including nucleotide sequences, protein structures, genomic annotations, gene expression profiles, and metabolic pathways (Danielewski *et al.* 2025).

For instance, linking protein databases such as UniProt with structural databases like the Protein Data Bank (PDB) and domain databases such as Pfam allows researchers to relate protein sequences to their three-dimensional structures and functional domains. This integrated view significantly enhances the understanding of protein function and evolutionary relationships (Danielewski *et al.* 2025).

Similarly, combining genomic databases like Ensembl and the UCSC Genome Browser with pathway databases such as KEGG and interaction databases like BioGRID helps researchers analyze how genetic variations influence cellular processes and contribute to disease mechanisms. Such integration is invaluable for identifying potential therapeutic targets (**Fig. 25**) (Singh 2025b).

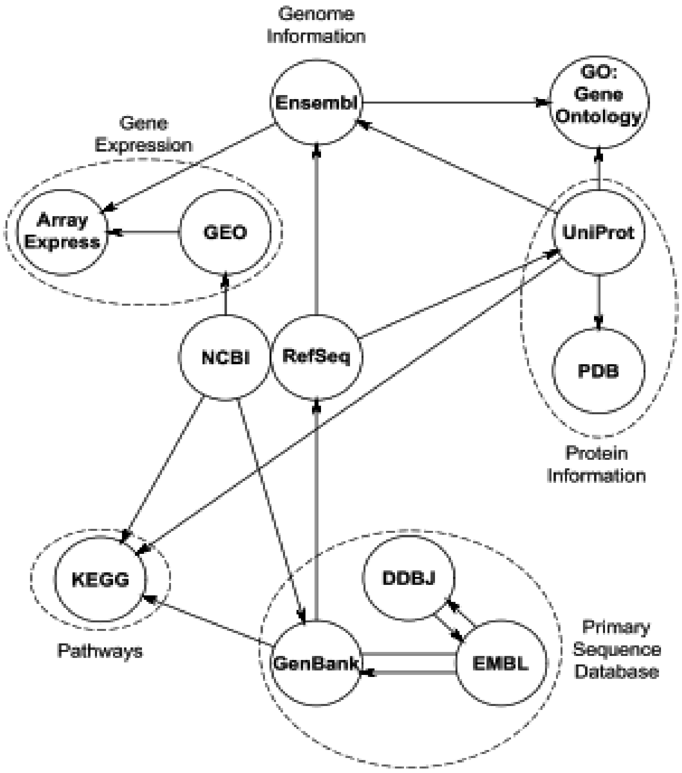


Figure 25. Integration of biological databases (Singh 2025b).

Beyond enhancing research, database integration promotes global data standardization and sharing. Initiatives like the International Nucleotide Sequence Database Collaboration (INSDC) ensure consistency across major databases including GenBank, EMBL-EBI, and DDBJ. Likewise, gene expression databases such as GEO and ArrayExchange are interlinked, facilitating cross-study comparisons and re-analysis. Furthermore, this interoperability supports developing advanced bioinformatics tools that can process, analyze, and visualize complex biological data. These capabilities are essential for driving discovery in fields like genomics, proteomics, and systems biology (Singh 2025b).

Chapter 3. Use of Appropriate Software

3.1. Introduction

Bioinformatics relies on two distinct approaches: web-based tools and command-line tools (**Fig. 26**). Web-based resources—often called "point-and-click" tools—are easy to use and require no programming experience. In contrast, command-line tools provide greater flexibility, precision, and power for handling large datasets, and support more reproducible workflows through explicit documentation of analytical steps (Pevsner 2015).

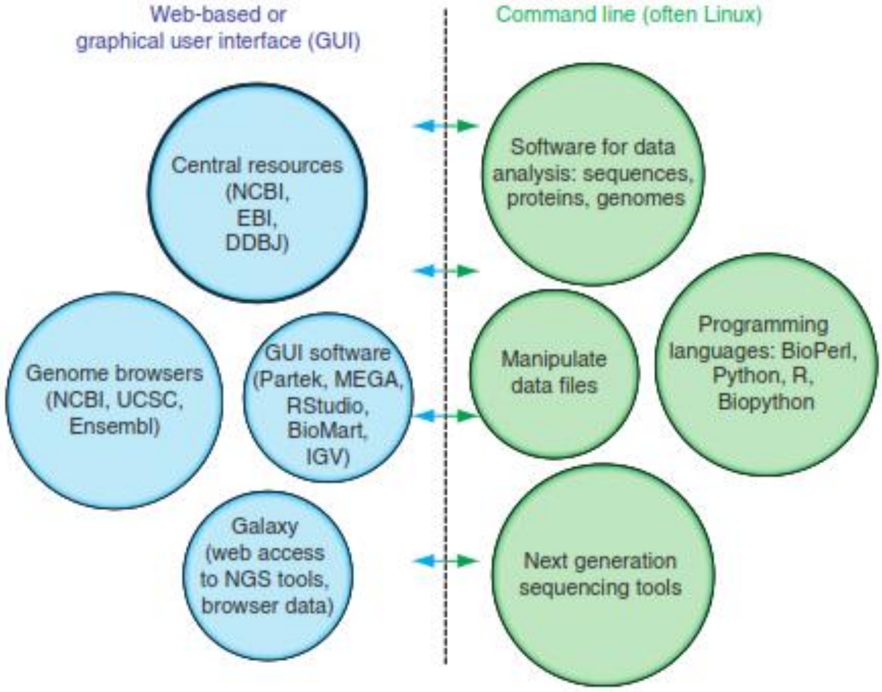

Web-based or graphical user interface (GUI)
Command line (often Linux)
Central resources (NCBI, EBI, DDBJ)
Software for data analysis: sequences, proteins, genomes
Genome browsers (NCBI, UCSC, Ensembl)
GUI software (Partek, MEGA, RStudio, BioMart, IGV)
Manipulate data files
Programming languages: BioPerl, Python, R, Biopython
Galaxy (web access to NGS tools, browser data)
Next generation sequencing tools

Figure 26. Bioinformatics resources. User-friendly, web-based tools are listed on the left side, featuring popular platforms such as the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute, genome browsers like Ensembl and UCSC, and other specialized databases. On the right side, command-line tools are shown, which include programming languages like Biopython, BioPerl, and R, as well as software tools most commonly used via the Linux operating system (Pevsner 2015).

3.2. Web-Based Software Approach

Bioinformatics makes extensive use of web-based software tools that are accessible through internet browsers. These tools allow researchers to analyze biological data without needing to install complex programs or use command-line interfaces. The internet serves as a central platform for accessing data, analyzing sequences, and integrating different types of biological information (Ma *et al.* 2024).

Some of the most important web resources include:

- The National Center for Biotechnology Information (NCBI) (http://ncbi.nlm.nih.gov), which provides GenBank and many analysis tools
- The European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk), which offers numerous databases and analysis services
- Ensembl (http://www.ensembl.org), which features powerful genome browsers and annotation resources
- The UCSC Genome Browser (http://genome.ucsc.edu), which enables visualization and exploration of genomic data

Throughout this book, we will introduce nearly 1,000 additional specialized websites for various bioinformatics applications. Web-based tools are particularly valuable because they are:

- Easily accessible from any computer with internet connection
- Frequently updated with new features and data
- Designed with user-friendly interfaces
- Available without requiring programming knowledge
- Community-supported with documentation and help resources

These characteristics make web tools ideal for students and researchers who are new to bioinformatics or who need to perform analyses without extensive computational training.

For comprehensive information about available biological databases, consult the annual database issue of Nucleic Acids Research (http://nar.oxfordjournals.org).

3.3. Command-Line Software Approach

Command-line tools offer distinct, critical advantages for bioinformatics research. High-throughput approaches to biology result in the creation of both large and small datasets that require sophisticated analyses. We can think about command-line software in several ways.

The operating system is often Linux (a Unix-like environment). The Mac OS is compatible with Linux as well (and is POSIX-compliant). POSIX (Portable Operating System Interface) offers standards for maintaining compatibility between operating systems. However, while Windows-type operating systems are popular, they are not appropriate for the majority of command-line programs. For Windows users, solutions like Cygwin or Windows Subsystem for Linux (WSL) can provide Unix-like functionality. Many researchers also access Linux servers remotely using tools like PuTTY (Brandies and Hogg 2021).

Programming languages are commonly used in bioinformatics. Examples are Perl (and its relative BioPerl), Python (as well as Biopython), and R to manipulate data. Learning such languages is important as it is extremely useful to be able to write scripts and thus accomplish a broad range of tasks. The BioConductor project currently includes over 1,000 packages that are useful for solving many tasks. While acquiring knowledge of R has a steep learning curve, it is possible to use R packages without being an expert user (Das *et al.* 2024).

The command line of Unix systems offers Bash, a default shell for Linux and Mac OS X operating systems. Bash includes a series of utilities that can accomplish tasks such as sorting a table of data, transposing it, counting the numbers of rows and columns, merging data, or working with regular expressions (Yingngam 2024).

We may further distinguish between using command-line software and using a programming language. Learning Perl, Python, or other languages offers tremendous benefits. However, even if you do not program, you still should learn basic information about how to acquire, store, manipulate, and explore large files. Many files used in bioinformatics and genomics are simply too large to be handled efficiently by web-based or GUI-based software. Many files that are generated by software tools require some level of restructuring to be further studied. For many students, it has become essential to learn techniques to manipulate files on the command line (Karabayev *et al.* 2021).

3.4. Bridging the Two Approaches

Bioinformatics provides multiple pathways to access and analyze data, ranging from user-friendly web interfaces to powerful command-line tools. As shown in **Table 8**, major bioinformatics resources typically offer both approaches to accommodate users with different levels of computational experience and research needs (Pevsner 2015).

Table 8. Overview of web-based (GUI) and command-line software tools in bioinformatics (Pevsner 2015).

Topic	Web-based or GUI software	Command-line software
Access to information	BioMart Genome Workbench	EDirect
Pairwise alignment	BLAST	BLAST+, Biopython, needle (EMBOSS), water (EMBOSS)
BLAST	BLAST	BLAST+
Database searching	DELTA-BLAST, Megablast	HMMER
Multiple alignment	Pfam, MUSCLE	MAFFT
Phylogeny	MEGA	MrBayes
Chromosomes	Galaxy	geecee (EMBOSS) isochore (EMBOSS)
Next-generation sequencing	Galaxy, SIFT, PolyPhen2	SAMTools, tabix, VCFtools
RNA	RNAfam, tRNAscan	
RNAseq	Galaxy	affy (R package), RSEM
Proteomics	ExPASy	pepstats (EMBOSS)
Protein structure	Cn3D, PyMol	psiphi (EMBOSS)
Functional genomics	FLink, Cytoscape	
Tree of life		Velvet (assembly)
Viruses		MUMmer (alignment)
Bacteria and archaea	MUMmer	GLIMMER (gene-finding)
Fungi	YGOB	Ensembl (variants)
Human genome		PLINK
Human disease	OMIM, BioMart	EDirect, MitoSeek

The National Center for Biotechnology Information (NCBI) exemplifies this dual approach through its Entrez system, which provides a web-based search interface for biological databases, alongside EDirect, a suite of command-line programs that enable automated access and data retrieval. Similarly, the European Bioinformatics Institute (EBI) and Ensembl offer both web interfaces and comprehensive application programming interfaces (APIs) for programmatic access using various programming languages (Kans 2024).

Galaxy represents a particularly valuable bridge between these approaches, providing web-based access to hundreds of tools that would normally require command-line expertise. This platform allows users to build complex analysis workflows through a graphical interface while essentially running command-line tools in the background (Bogaerts *et al.* 2025).

The choice between web-based and command-line approaches depends on several factors:

- The scale and complexity of the research project
- The need for reproducibility and automation
- The researcher's computational background and willingness to learn new skills
- Specific requirements for data processing and analysis

For large-scale projects like next-generation sequencing, command-line tools are essential for handling big data. Beginners can transition gradually by starting with web platforms like Galaxy to learn core concepts before advancing to the command line (Bogaerts *et al.* 2025).

The bioinformatics community actively develops integrated solutions that connect these approaches. Resources like BioMart provide a web-based interface for querying interconnected databases, while packages like biomaRt enable the same functionality through programming environments like R. This dual access strategy ensures that both computational biologists and bench scientists can access the same data using tools appropriate to their skills and needs (Pevsner 2015).

3.5. Ensuring Reproducibility in Bioinformatics

Scientific research depends on reproducibility—the ability to confirm and build upon previous findings. In bioinformatics, this requires careful documentation and organization of all research components, whether using web-based or command-line approaches. Key practices include:

3.5.1. Documentation and Organization

- Maintain detailed records of all analytical steps, including commands, parameters, and software versions
- Use electronic lab notebooks alongside traditional notebooks to capture computational workflows
- Implement consistent file organization systems following established guidelines

3.5.2. Data and Metadata Management

- Deposit research data in appropriate public repositories:
  - **a.** Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) at NCBI
  - **b.** ArrayExpress and European Nucleotide Archive (ENA) at EBI
- Include comprehensive metadata—essential information about experimental conditions, sample characteristics, and processing methods
- Document database versions and access dates, as biological databases frequently update their contents

3.5.3. Software and Workflow Preservation

- Record exact software versions and analysis parameters
- Use version control systems like GitHub to share and preserve analysis code
- Provide sufficient detail to enable other researchers to exactly replicate your analyses

These practices ensure that bioinformatics research remains transparent, verifiable, and building toward reliable scientific knowledge. By carefully documenting workflows and sharing data and code, researchers contribute to the cumulative progress of science while maintaining the integrity of their work (Ziemann *et al.* 2023).

3.6. Machine Learning as a Bioinformatics Tool

Machine learning (ML) has emerged as a fundamental category of software tools within the bioinformatics ecosystem, designed to extract meaningful patterns and build predictive models from the complex and high-dimensional data typical of modern biology. The power of ML lies in its ability to learn directly from data without being explicitly programmed for a specific task, making it uniquely suited for uncovering hidden relationships in genomic sequences, protein structures, microscopic images, and biological networks. Selecting the appropriate ML tool is a critical software decision, as its effectiveness is contingent on a strong alignment between the algorithm's architecture and the inherent structure of the biological data (Crossa *et al.* 2025).

The foundational principles of using any bioinformatics software, particularly those concerning reproducibility and rigorous validation as discussed in section 3.4, are paramount when employing ML tools. This ensures that models are not "black boxes" but are instead transparent, documented, and their results can be independently verified. The core challenge lies in choosing the right algorithmic approach. For instance, Convolutional Neural Networks (CNNs) are the appropriate tool for analyzing data with spatial or image-like properties, such as microscopy images or sequences represented as spectrograms. In contrast, Recurrent Neural Networks (RNNs) or Transformers are specifically designed for sequential data like nucleotide or protein sequences. For data representing relationships and interactions, such as molecular structures or protein-protein interaction networks, Graph Neural Networks (GCNs) are the most suitable software tool (Crossa *et al.* 2025).

This strategic selection process, which matches the software's algorithmic strength to the data's structure, is fundamental to generating robust and biologically insightful results. The following **table 9** provides a practical guide for this software selection process across common biological data types (Greener *et al.* 2022).

Table 9. Recommendations for the use of machine learning strategies for different biological data types (Greener *et al.* 2022).

input data	example prediction tasks	recommended models	challenges
Gene sequence	DNA accessibility 3D genome organization Enhancer–promoter interactions	1D CNNs RNNs Transformers	Repetitive regions in genome Sparse regions of interest Very long sequences
Protein sequence	Protein structure Protein function Protein–protein interaction	2D CNNs and residual networks using co-variation data Multilayer perceptrons with windowing Transformers	Metagenome data stored in many places and therefore hard to access Data leakage (from homology) can make validation difficult
Protein 3D structure	Protein model refinement Protein model quality assessment Change in stability upon mutation1	GCNs using molecular graph 3D CNNs using coordinates Traditional methods using structural features Clustering	Lack of data, particularly on protein complexes Lack of data on disordered proteins
Gene expression	Intergenic interactions or co-expression Organization of transcription machinery	Clustering CNNs Autoencoders	Unclear link between co-expression and function High dimensionality High noise
Mass spectrometry	Detecting peaks in spectra Metabolite annotation	CNNs using spectral data Traditional methods using derived features	Lack of standardized benchmarks Normalizationation required between different datasets
Images	Medical image recognition Cryo-EM image reconstruction RNA-sequencing profiles	2D CNNs and residual networks Autoencoders Traditional methods using image features	Systematic differences in data collection affect prediction Hard to obtain large datasets of consistent data
Molecular structure	Antibiotic activity Drug toxicity Protein-ligand docking Novel drug generation	GCNs using molecular graph Traditional methods or multilayer perceptrons using molecular properties RNNs using text-based representations of molecular structure such as SMILES Autoencoders	Experimental data available for only a tiny fraction of possible small molecules
Protein–protein interaction network	Polypharmacology side effects Protein function	GCNs Graph embedding	Interaction networks can be incomplete Cellular location affects whether proteins interact High number of possible combinations

Biological data types each present distinct prediction tasks where machine learning has proven effective, along with appropriate model choices and specific analytical challenges. Common issues like high dimensionality affect most biological data. CNN, convolutional neural network; cryo-EM, cryogenic electron microscopy; GCN, graph convolutional network; RNN, recurrent neural network. Normalization involves rescaling variables across datasets to ensure comparable weighting and ranges, typically through mean subtraction and standard deviation division (standardization), which is essential for mitigating systematic differences between experimental protocols and instrumentation.

Chapter 4. Biological Sequences Analysis

4.1. Short History of Sequence Analysis

The field of sequence analysis began with the sequencing of the insulin protein. This work earned Frederick Sanger his first Nobel Prize and marked the start of modern molecular biology by providing the first precise molecular dataset. In the 1960s, new protein sequences were discovered very slowly. Scientists analyzed them using manual methods, like writing sequences on paper strips and moving them around on lab walls to find matches. This slow pace turned out to be useful, as the computers needed to analyze this data were not yet available. When the first large computers arrived, scientists began programming these manual comparison methods into them. This was a brand-new challenge that required researchers to invent techniques from scratch, leading to the creation of a new field: bioinformatics, the computer-based analysis of biological sequences (Claverie and Notredame 2011).

The field grew quickly in the 1980s with the rise of DNA sequencing. The first specialized software appeared, such as the GCG Suite for large mainframe computers and DNASTAR for personal computers. This period also saw the creation of the three major international genetic databases: GenBank in the USA, EMBL in Europe, and DDBJ in Japan. Shortly after, in 1986-1987, these groups agreed to share their data, forming the International Nucleotide Sequence Database Collaboration (INSDC), a global partnership that remains essential to research today (Tiwary 2022).

4.2. Advances in Biological Sequence Analysis

Progress in biotechnology and information technology has created a new era of biological big data. This includes data about molecular sequences, structures, functions, and properties. The amount of available biological sequence data has grown tremendously, along with information about what these sequences do and how they are structured (**Fig. 27A**). Analyzing this biological big data helps us better understand how living systems work and provides important insights into disease mechanisms. Current approaches for analyzing biological sequences fall into three main categories: methods using patterns and statistics from sequences, methods leveraging large biological datasets, and techniques adapted from natural language processing (**Fig. 27B**). These analysis methods have become valuable tools for medical applications (**Fig. 27C**). The continued development of sequence analysis techniques has enhanced our understanding of biological processes and significantly advanced biomedical research (Wei *et al.* 2024).

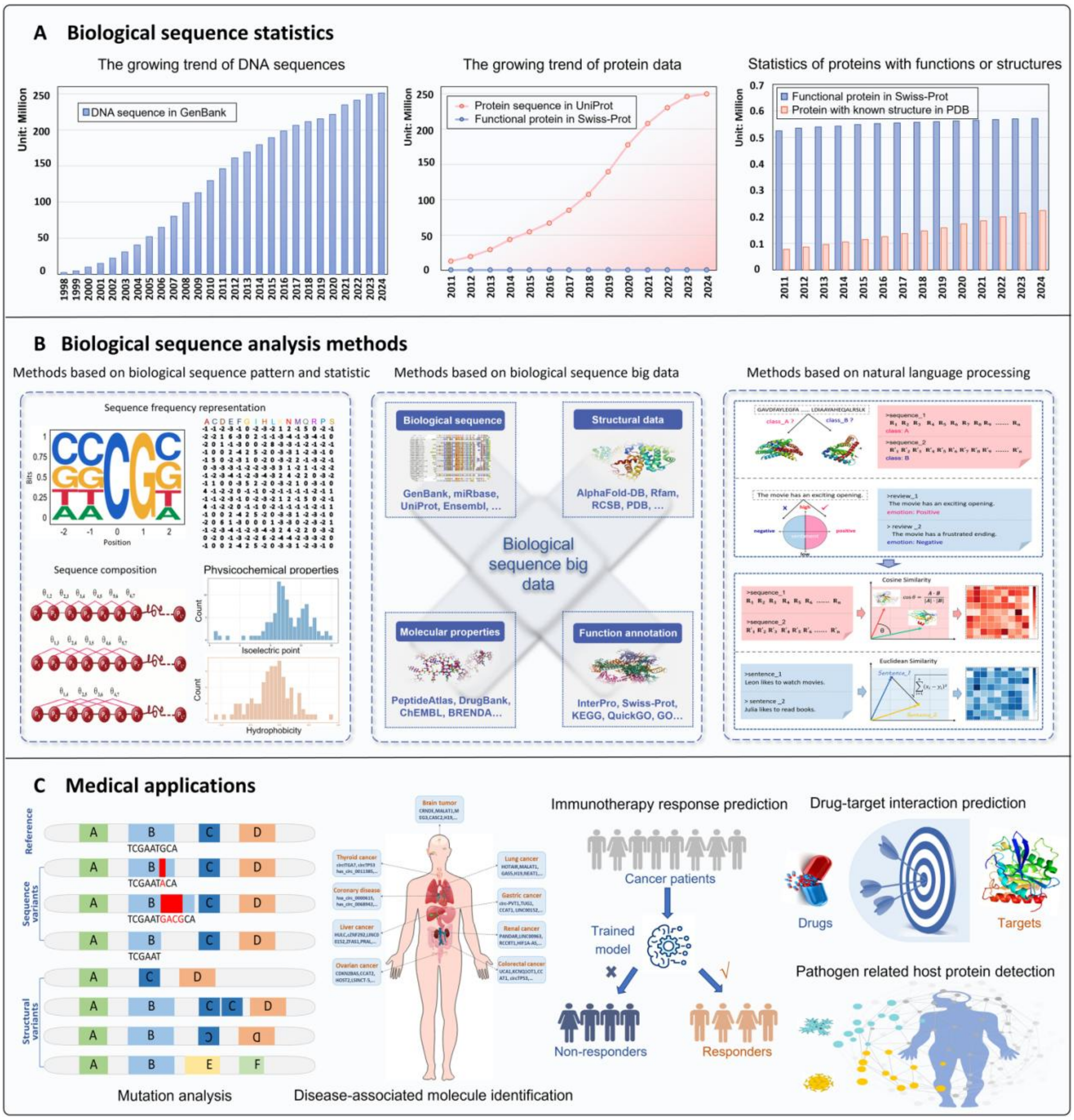

A Biological sequence statistics
The growing trend of DNA sequences
Unit: Million
DNA sequence in GenBank
The growing trend of protein data
Protein sequence in UniProt
Functional protein in Swiss-Prot
Statistics of proteins with functions or structures
Functional protein in Swiss-Prot
Protein with known structure in PDB
B Biological sequence analysis methods
Methods based on biological sequence pattern and statistic
Methods based on biological sequence big data
Methods based on natural language processing
Sequence frequency representation
Sequence composition
Physicochemical properties
Isoelectric point
Hydrophobicity
Biological sequence
GenBank, miRbase, UniProt, Ensembl, ...
Structural data
AlphaFold-DB, Rfam, RCSB, PDB, ...
Biological sequence big data
Molecular properties
PeptideAtlas, DrugBank, ChEMBL, BRENDA...
Function annotation
InterPro, Swiss-Prot, KEGG, QuickGO, GO...
Cosine Similarity
Euclidean Similarity
C Medical applications
Reference
Sequence variants
Structural variants
Mutation analysis
Disease-associated molecule identification
Immunotherapy response prediction
Cancer patients
Trained model
Non-responders
Responders
Drug-target interaction prediction
Drugs
Targets
Pathogen related host protein detection

Figure 27. Advancements in biological sequence data, analysis methods, and medical applications (Wei *et al.* 2024).

4.3. Retrieving Protein Sequences

Following the historical development of sequence databases, the next practical step is retrieving sequences for analysis. After using PubMed for literature searches, a common task is to obtain the corresponding protein sequences.

Two of the most important resources for this are the **NCBI Protein** database and the **Universal Protein Resource (UniProt)**. Both are essential for researchers, but they have different strengths.

- **NCBI Protein** is fully integrated with other NCBI resources like GenBank and PubMed. This makes it powerful for finding protein data linked to genetic and bibliographic information.
- **UniProt** is a specialized protein database known for its high-quality, detailed annotations. It is collaboratively maintained by institutes in Europe, Switzerland, and the United States. UniProt is updated frequently and offers specialized tools for protein analysis.

These databases provide sequences in standard formats for use in software. The most common format is **FASTA**, which is simple and consists of a definition line starting with a > symbol, followed by the sequence itself using one-letter codes. Using the correct format is essential for most bioinformatics tools (Consortium 2024; Goldfarb *et al.* 2025).

4.4. Retrieving DNA Sequences

Working with DNA sequences is more complex than working with proteins. Protein sequences are relatively simple; they have a fairly consistent size and clear boundaries. DNA sequences, especially in higher organisms like animals and plants, are much more variable and complex.

A gene is not just the segment of DNA that codes for a protein. It includes several parts:

- **Regulatory regions** that control the gene's activity.

- **The coding region** itself, which is often split into segments called exons.
- **Non-coding segments** called introns that are removed before the protein is made.

Because of this complexity, the link between a protein and its DNA sequence in databases is not always straightforward. A single protein can be linked to several different DNA records, such as the full gene (including introns), the final processed mRNA, or just the coding segment.

While it's often easier to work directly with protein sequences, sometimes you need to find the original DNA sequence. A common reason is to clone a gene using a technique like PCR. To find the DNA sequence for a protein, you can use database cross-references. For example, after finding your protein of interest in a database like UniProt, look for a "Cross-References" section. This section provides links to other databases like GenBank or EMBL, which will contain the relevant DNA sequence information for that protein (Claverie and Notredame 2011).

4.5. Sequence Homology

4.5.1. Definition and Core Concepts

Sequence homology is a fundamental concept in bioinformatics. It refers to the similarity between DNA, RNA, or protein sequences that results from their shared evolutionary ancestry. When two sequences are homologous, it means they descended from a common ancestor sequence. This principle allows us to use sequence similarity to predict the function of a new gene, understand evolutionary relationships between species, and infer the structure of proteins (Singh 2025d).

It's crucial to distinguish homology from two related terms:

- **Identity** measures the exact percentage of matching positions in an alignment.
- **Similarity** quantifies how closely related sequences are, including biochemically similar residues.

Homology itself has two main subtypes:

- **Orthologs** are homologous sequences in different species that arose from a common ancestral gene during speciation.
- **Paralogs** result from gene duplication within a species and may evolve new functions.

While homologous sequences often share significant similarity and related structures, some show minimal sequence identity despite their common origin (Pevsner 2015).

4.5.2. Visualizing Similarity: Dot Plots

A simple way to visualize sequence similarity is with a dot plot. This is a grid where one sequence is on the x-axis and the other on the y-axis. A dot is placed where the characters in both sequences match.

- A solid diagonal line indicates strong similarity or identical sequences (**Figure 28A**).
- A broken or shifted diagonal line reveals differences like mutations or insertions.

Figure 28B shows a real example: a dot plot comparing two prion gene sequences. The clear diagonal line confirms they are very similar. This method provides a quick, visual way to compare sequences (Singh 2025d).

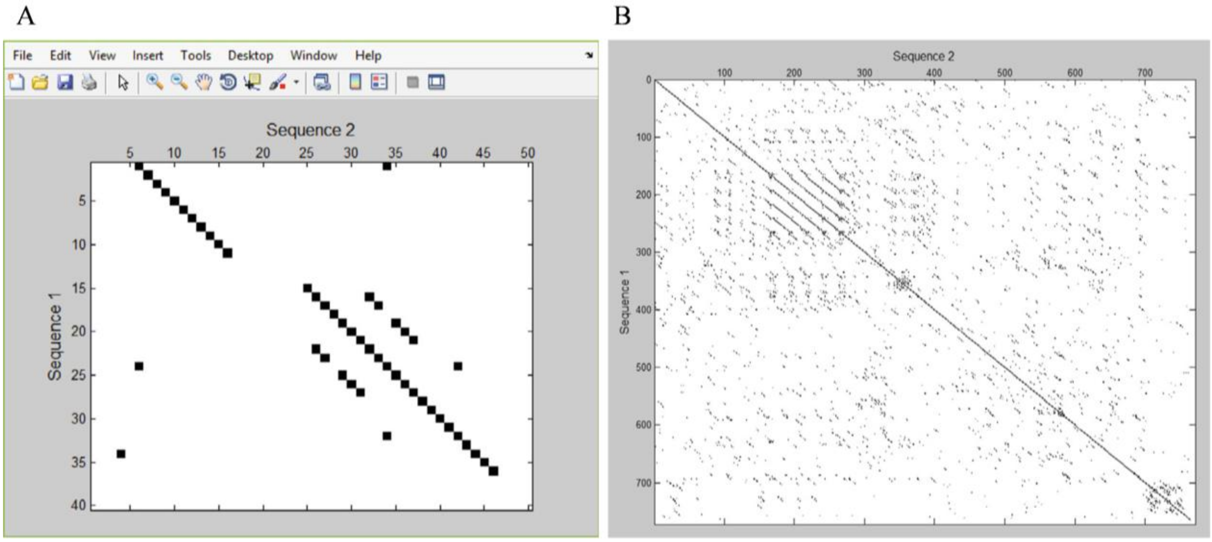


Figure 28. Dot plot analysis. A. Dot plot for two DNA sequences. B. Dot plot for DNA sequences of two prion proteins downloaded from GenBank (Singh 2025d).

4.5.3. Methodologies for Detecting Homology

To systematically find homologous sequences, scientists use specialized tools and methods:

a. Sequence Alignment Tools (e.g., BLAST): These are the most widely used tools. They perform fast, sensitive searches of databases to find sequences with significant similarity to a query sequence, suggesting homology.

b. Pattern-Based Methods: These methods compare a query sequence to known patterns or motifs associated with specific protein families to detect homology.

c. Advanced Algorithms (e.g., COMER, Markov Models): More sophisticated tools offer increased sensitivity and accuracy for detecting distant evolutionary relationships that simpler tools might miss (Dapkūnas and Margelevičius 2022).

4.5.4. Applications of Sequence Homology

Sequence homology is a powerful tool with many applications:

a. Gene Function Assignment: If a new gene is homologous to a gene with a known function (e.g., a digestive enzyme), we can hypothesize that it has a similar function.

b. Protein Structure Prediction: Homologous sequences often fold into similar 3D structures.

c. Evolutionary Studies: By comparing homologous sequences from different species, we can reconstruct evolutionary trees and understand how organisms are related (Dapkūnas and Margelevičius 2022).

4.5.5. Challenges and Limitations

While incredibly useful, sequence homology has limitations that require careful interpretation:

a. Short matches can be misleading: A short, identical stretch between two sequences may not always indicate true homology or similar function.

b. Complex Scenarios: In cases like detecting circular RNAs, homologous sequences can sometimes lead to false positives, complicating analysis.

c. Distant Relationships: Detecting homology becomes very challenging when sequences have diverged significantly over evolutionary time (Sayed *et al.* 2018).

In short, sequence homology is essential in bioinformatics, but we must be aware of both its strengths and limitations.

4.6. Sequence Analysis Tools

Sequence analysis tools are essential in bioinformatics for studying DNA, RNA, and protein sequences. Each tool has a distinct purpose, facilitating various stages of genomic data analysis, from assembly and alignment to gene prediction. Understanding their roles is crucial for researchers.

These tools can be broadly categorized by their primary function:

4.6.1. Fundamental Sequence Alignment Tools

- **BLAST (Basic Local Alignment Search Tool):** BLAST is used to compare a query nucleotide or protein sequence against large databases to find similar sequences. This helps infer gene function, predict gene families, and explore evolutionary relationships. A key limitation is that its extensive output often requires additional software for analysis.
- **ClustalW:** This tool performs multiple sequence alignments, which are vital for identifying conserved regions across sequences and understanding phylogenetic relationships. It is often used with other tools for comprehensive analysis.

4.6.2. Genome Assembly Tools

- **Velvet:** Velvet is a *de novo* assembler designed for short-read data from next-generation sequencing. It is highly effective for assembling a single genome but less suited for mixed-species samples (metagenomics), leading to the development of MetaVelvet for that purpose.
- **SPAdes:** SPAdes is another genome assembler that excels at handling data from both single-cell and standard multicell samples. It is particularly useful for assembling genomes of uncultivable bacteria and is designed to address challenges like uneven read coverage and sequencing errors.

4.6.3. Gene Prediction Tools

- **GeneMark:** This tool identifies protein-coding regions within DNA sequences, a critical step in annotating newly sequenced genomes. It is frequently used alongside other annotation software for comprehensive results.
- **AUGUSTUS:** AUGUSTUS is an *ab initio* gene prediction tool that predicts gene structures, including the boundaries between exons and introns. It is valuable in evolutionary genomics and is used in automated analysis pipelines.

While these tools are powerful, researchers face challenges such as their complexity to install and the need to integrate them into automated workflows for efficient data processing. The choice of tool depends on specific research needs, like the type of sequencing data available. As bioinformatics evolves, developing more integrated and user-friendly tools will be essential to manage the growing volume and complexity of genomic data (Ishengoma and Rhode 2022; Abdi *et al.* 2024).

4.7. Biological Sequences and Association Analysis

4.7.1. Biological Sequence Pattern Mining

Patterns in biological sequences often signal important functional or structural elements. For example, conserved patterns within a protein family can determine the family's overall structure and function. Evolutionary processes also generate repetitive sequences that may lead to new gene formation. Additionally, regulatory sequences like transcription factor binding sites help control gene expression (Chen 2024). Because these patterns are biologically significant, identifying them is a key task in bioinformatics. Pattern mining algorithms generally fall into two categories: those that find frequent patterns within a single sequence, and those that find patterns repeated across multiple sequences. The latter is often more relevant in biology, where comparing across sequences is essential (Chen 2024).

Several algorithms have been developed for multi-sequence mining. Early algorithms like GSP generated many candidate patterns, requiring high computational resources. PrefixSpan improved efficiency by using a divide-and-conquer approach without generating candidates, reducing memory and time requirements (Chen 2024). However, biological data poses unique challenges due to its size and complexity. Specialized algorithms like BioPM and MBioPM have been developed for biological sequences, but issues around efficiency, memory use, and scalability remain. Methods such as SUA_SATR and REPuter use advanced data structures to find repeats quickly but may struggle under certain conditions (Chen 2024).

4.7.2. Pattern Recognition of Biological Sequence

The Human Genome Project was a massive international scientific project. Its main goal was to map the entire human genome by sequencing all 3 billion DNA base pairs in our chromosomes. This meant creating a complete genetic map, identifying every gene, and ultimately decoding all human genetic information. When we have a genome sequence, the most important task is to find the genes within it and to understand how those genes are controlled. This process of finding and interpreting these genetic instructions is known as pattern recognition (Oyelade *et al.* 2020).

4.7.2.1. Gene Recognition

Gene recognition is the process of finding genes within a long DNA sequence. This work has two main parts. The first part is to identify special sequence signals that are related to genes, like promoters and start codons. Finding these signals helps scientists roughly locate the area where a gene begins. The second part is to accurately predict the exact location of the gene's coding region, which is the part that carries the instructions to make a protein. To do this, bioinformaticians use methods like training artificial neural networks and performing evolutionary analyses. The final result of this analysis helps us understand the gene's function by revealing its structure and how it is regulated (Koonin and Galperin 2003).

4.7.2.2. Regulation of Gene Expression

Gene expression regulation is the process cells use to control when and how a gene is used to make a protein. This complex process involves turning the genetic information stored in DNA into a functional protein molecule (Blanco and Blanco 2022).

4.7.2.2.1. Regulatory Elements

Regulatory elements are short DNA sequences, often called motifs, that control gene expression. They are usually found in the region just before a gene (the upstream region). These elements work by being recognized and bound by special proteins called transcription factors. This binding acts like a switch to turn genes on or off. Identifying these elements is a crucial step in understanding the genome. While they can be found through lab experiments, that process is very slow. Therefore, scientists now primarily use computational bioinformatics methods to quickly find these patterns and guide experiments (Narlikar and Ovcharenko 2009).

4.7.2.2.2. Computation Methods for Studying Regulatory Elements

There are two main types of computational methods used to find regulatory elements, each with a different strategy. The first type uses heuristic search algorithms, like Gibbs sampling or hidden Markov models. These methods use smart guesses to find a good solution quickly and are efficient for long sequences. However, they might not find the absolute best pattern and can sometimes get stuck on a good-but-not-perfect answer. The second type uses exhaustive search algorithms. These methods check every single possible pattern to guarantee they find the very best one. The downside is that they are very slow and can only be used for very short sequences (Chen 2024).

4.7.2.2.3. General Steps for the Regulatory Element Recognition

Although different algorithms are used, the process for finding regulatory elements usually follows four general steps. The first step is Selection, which means choosing the most representative sequence fragments to study. The second step is Classification, which is grouping those sequences into meaningful categories. The third step is Alignment, where scientists line up the sequences to find a common pattern or motif that they all share. The final step is Search, where this common pattern is used to scan the entire genome to find all other locations where it appears (Chen 2024).

4.7.2.3. Biological Sequence Pattern Recognition

Biological sequence pattern recognition is the fundamental task of finding important patterns within DNA or protein sequences. Whether the goal is to find a gene, identify a regulatory motif, or understand a protein's function, it all depends on this ability to recognize meaningful patterns. This makes pattern recognition a central and essential part of bioinformatics research (Chen 2024).

Chapter 5. Sequence Alignment and Comparison

5.1. Introduction

Sequence alignment is a foundational bioinformatics method used to compare DNA, RNA, or protein sequences. By identifying similarities between sequences, researchers can infer evolutionary relationships, predict protein functions, and discover conserved domains. As genomic data continues to grow, alignment methods remain essential for understanding biological systems (Zhang *et al.* 2024).

5.2. Protein Alignment: More Revealing than DNA

Protein alignment is generally more informative than DNA alignment for sequence comparison. This is because protein sequences filter out silent DNA mutations and can detect similarities between amino acids with related properties, making them superior for identifying distant evolutionary relationships. Therefore, analyzing the protein product is usually the best strategy when studying a coding gene, using tools like TBLASTN to search translated DNA databases. However, DNA alignment remains crucial for specific applications, including gene confirmation, polymorphism detection, analysis of cloned fragments, and studying non-coding regulatory regions (Pevsner 2015).

5.3. Protein Sequence Alignment

Protein sequencing has evolved from a final characterization step to a primary tool for functional prediction. The Human Genome Project enabled this shift, allowing researchers to sequence genes first and infer protein function through alignment methods. High-performance comparison tools now enable discoveries through computational analysis alone. For instance, computational methods revealed that a human tumor suppressor gene relates to DNA repair enzymes in yeast and *E. coli*, providing crucial insights into oncogene mutations. As genomic data expands, protein sequence comparison becomes increasingly essential for understanding biological function. This approach works because evolution conserves key sequence information. Proteins can be traced back billions of years and compared with homologous relatives sharing common ancestors. These homologous proteins typically maintain similar 3D structures, active sites, binding domains, and functions. Protein sequence alignment serves two main purposes: creating accurate sequence alignments and identifying functions of new proteins by finding similar known proteins in databases (Singh 2025c).

5.3.1. Scoring Matrices for Protein Alignment

Protein sequence alignment relies on scoring matrices to evaluate amino acid substitutions. The PAM (Point Accepted Mutation) and BLOSUM (Blocks Substitution Matrix) matrices serve this purpose but differ significantly in their design and applications (Trivedi and Nagarajaram 2020).

5.3.1.1. PAM Matrices

a. Evolutionary Foundation: PAM matrices model evolutionary changes, with PAM1 representing 1% amino acid change and higher numbers (e.g., PAM250) indicating greater evolutionary distances

b. Construction Method: Derived from closely related proteins, using mathematical extrapolation to model longer evolutionary timescales

c. Typical Use: Best for aligning evolutionarily related sequences, with higher PAM numbers for more divergent sequences

5.3.1.2. BLOSUM Matrices

a. Empirical Basis: Built from observed substitutions in conserved protein blocks without evolutionary modeling assumptions

b. Construction Method: Uses sequence blocks from the BLOCKS database, with numbers (e.g., BLOSUM62) indicating the minimum percentage identity of included sequences

c. Typical Use: Preferred for general alignment tasks, with lower numbers for more divergent sequences

5.3.2. Matrix Selection Guidelines

5.3.2.1. Comparative Applications

PAM matrices suit evolutionary studies, while BLOSUM matrices work better for practical alignment of sequences with unknown divergence. PAM may not handle highly divergent sequences well, while BLOSUM doesn't provide evolutionary distance information.

5.3.2.2. Practical Considerations

Understanding these differences ensures selecting the right matrix for specific alignment challenges. BLOSUM matrices are generally preferred for most database searches and routine alignments, while PAM matrices remain valuable for evolutionary studies and phylogenetic analysis (Trivedi and Nagarajaram 2020).

5.4. Pairwise Sequence Alignment

5.4.1. Fundamental Concepts and Definition

Pairwise sequence alignment is a fundamental bioinformatics technique used to compare two biological sequences (DNA, RNA, or proteins) to identify regions of similarity. This method helps researchers understand functional, structural, and evolutionary relationships between sequences. The alignment process involves arranging sequences to maximize matches while accounting for gaps and mismatches (Sofi *et al.* 2022).

5.4.2. Types of Pairwise Alignment

5.4.2.1. Global Alignment

Global alignment compares sequences from end to end using algorithms like Needleman-Wunsch. It works best when sequences are similar in length and highly conserved throughout. This method is ideal for comparing closely related genes or proteins that maintain similar structures and functions (Sofi *et al.* 2022).

5.4.2.2. Local Alignment

Local alignment, implemented through Smith-Waterman algorithm, identifies regions of high similarity within larger sequences. This approach is valuable for finding conserved domains in otherwise divergent sequences or detecting shared functional motifs in different proteins (Saif *et al.* 2023).

5.4.2.3. Semi-Global Alignment

Semi-global alignment focuses on aligning internal regions without penalizing end gaps. This method is particularly useful when comparing sequences with important central domains but variable terminal regions, such as in certain enzyme families (Pevsner 2015).

5.4.3. Algorithms and Methods

5.4.3.1. Classical Dynamic Programming Algorithms

The Needleman-Wunsch and Smith-Waterman algorithms form the foundation of sequence alignment. Using dynamic programming, they guarantee optimal alignments but require substantial computational resources, particularly for long sequences (DineshDarsi *et al.* 2023).

5.4.3.2. Advanced Computational Approaches

Recent developments include heuristic methods like BLAST and FASTA for rapid database searches, plus machine learning approaches that optimize scoring parameters. These methods sacrifice guaranteed optimality for significant speed improvements (DineshDarsi *et al.* 2023).

5.4.3.3. Emerging Technologies

Quantum computing algorithms show promise for exponential speedup in alignment tasks. Parallel computing approaches using graphics processing units (GPU) acceleration and distributed systems enable handling of large-scale genomic comparisons (Schmidt and Hildebrandt 2024).

5.4.4. Applications in Biological Research

Pairwise alignment enables evolutionary studies through phylogenetic analysis, functional annotation of unknown genes, mutation detection in disease research, and structural prediction. It forms the basis for database searching and comparative genomics (Saif *et al.* 2023).

5.4.5. Current Challenges and Future Directions

Key challenges include handling ultra-long sequences (e.g., complete chromosomes), improving accuracy for distantly related sequences, and developing faster algorithms for growing databases. Future work focuses on integrating AI methods, enhancing quantum algorithms, and developing specialized hardware for sequence analysis (DineshDarsi *et al.* 2023).

5.5. Basic Local Alignment Search Tool (BLAST)

The Basic Local Alignment Search Tool (BLAST) is the foundational algorithm for rapid sequence similarity searching in bioinformatics. While the previous chapter introduced its role in database querying, this section details its algorithmic principles, variants, and the interpretation of its statistical output. BLAST uses a heuristic approach to accelerate the search process, making it practical to find local alignments in massive databases. This method prioritizes speed by first identifying short, high-scoring matches ("seeds") between the query and database sequences, which it then extends to find significant alignments, rather than comparing every single residue exhaustively (Ali *et al.* 2024; Pevsner and Safari 2009).

5.5.1. BLAST Variants and Algorithmic Specificity

The core BLAST algorithm has been specialized into several programs, each tailored for a specific type of sequence comparison. The choice of program is critical for a successful search, as using an inappropriate variant can lead to missed matches or misinterpreted results. The main variants are summarized in **Table 10**.

Table 10. Variants of the NCBI-BLAST Program (Tiwary 2022; Singh 2025a).

BLAST program	Query Type	Database Type	Description
BLASTN	Nucleotide	Nucleotide	Compares a nucleotide query sequence against a nucleotide sequence database.
BLASTP	Protein	Protein	Compares an amino acid query sequence against a protein sequence database.
BLASTX	Nucleotide (translated)	Protein	Translates a nucleotide query sequence in all six reading frames and compares it against a protein sequence database.
TBLASTN	Protein	Nucleotide (translated)	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames.
TBLASTX	Nucleotide (translated)	Nucleotide (translated)	Translates both the nucleotide query and the nucleotide database in all six reading frames and compares the resulting protein sequences.
MEGABLAST	Nucleotide	Nucleotide	Suitable for finding alignments between closely related nucleotide sequences.
PSI-BLAST	Protein	Protein	Position-Specific Iterative BLAST. Suitable for searching remote homologs or members of a protein family.
PHI-BLAST	Protein	Protein	Pattern Hit Initiated BLAST. Suitable for finding protein sequences with a specific pattern in the database.

5.5.2. Statistical Significance of BLAST Results

The output of a BLAST search is not just a list of matches, but a statistically ranked list. Two key concepts are essential for interpretation:

- **High-Scoring Segment Pairs (HSPs):** These are the aligned regions between the query and a database sequence that lack long gaps and have a high similarity score based on the chosen substitution matrix and gap penalties.
- **E-value (Expect Value):** This is the most important statistical measure. It estimates the number of HSPs with a given score (or better) that one would expect to see by chance alone in the searched database.

A lower E-value indicates greater statistical significance. A common threshold for considering a match biologically relevant is 10^{-5}, meaning there is a 1 in 100,000 probability that the alignment occurred by random chance (Lu *et al.* 2024).

5.5.3. Comparative Tools: FASTA and BLAT

While BLAST is the most widely used tool, other algorithms offer different trade-offs between speed and sensitivity.

- **FASTA:** A similar search program that uses a different heuristic strategy. It is often cited as being more sensitive but slower than the standard BLASTN for some types of DNA-DNA comparisons.
- **BLAT (BLAST-Like Alignment Tool):** Optimized for scenarios where both the query and the database sequences are from the same or very closely related species (e.g., >95% identity). BLAT is dramatically faster than BLAST for mapping sequences like ESTs or mRNAs to a finished genome but is less sensitive for detecting more divergent sequences (Lu *et al.* 2024).

5.5.4. Interpreting BLAST Output

Understanding BLAST results requires analyzing key metrics to distinguish significant matches from random ones. The most important parameters include the E-value, which indicates statistical significance; percent identity, which shows sequence similarity; and query coverage, which reveals how much of your sequence aligns. Visualization tools can help manage large datasets and complex alignments (Lu *et al.* 2024).

The following figures illustrate how to read key sections of a BLAST report.

Figure 29 shows the search parameters, including the program and database used.

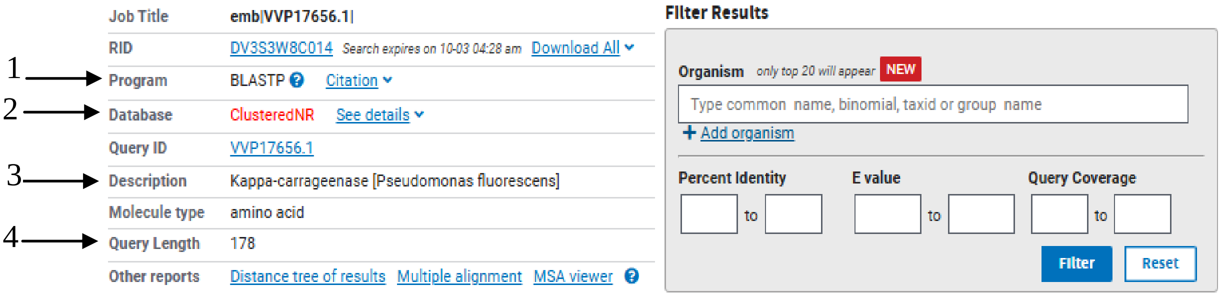
Figure 30 shows conserved domains and the distribution of homologous sequences.

Figure 31 displays the ranked list of matches with their statistical scores.

Figure 32 provides details of the pairwise sequence alignment. Together, these components enable comprehensive interpretation of BLAST results.

Another important result is the taxonomic distribution of the matches. As shown in **Figure 33**, BLASTP can group its highest-scoring hits by species, genus, and family. This provides a quick visual summary of which organisms contain proteins most similar to your query, offering immediate insight into its evolutionary relationships.

Together, these components enable a comprehensive interpretation of BLAST results, from statistical significance to biological context.



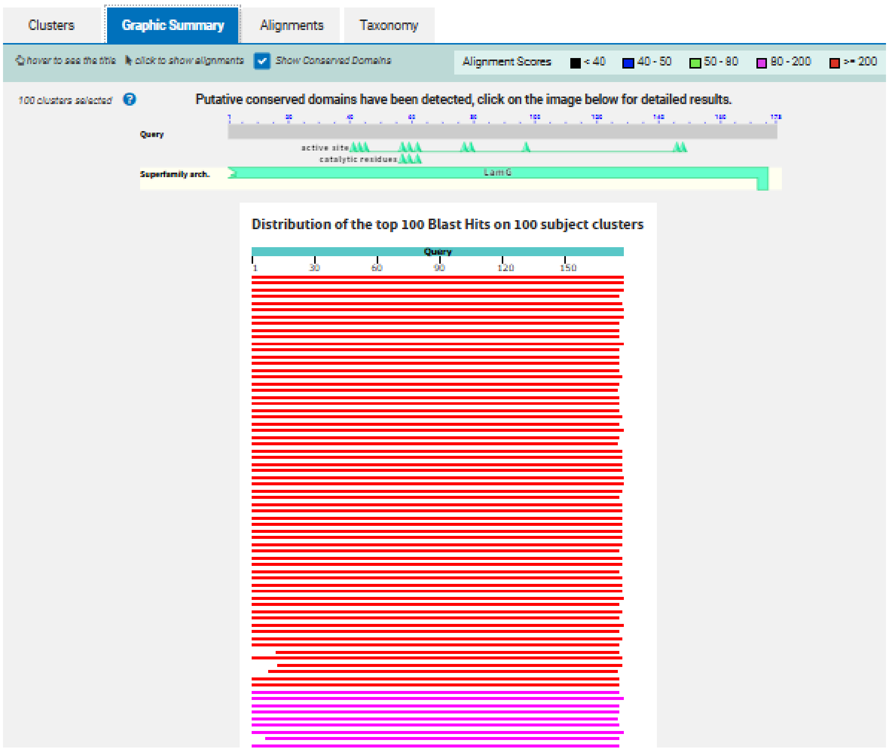
Job Title	emb|VVP17656.1|
RID	DV3S3W8C014 *Search expires on 10-03 04:28 am* Download All
Program	BLASTP Citation
Database	ClusteredNR See details
Query ID	VVP17656.1
Description	Kappa-carrageenase [Pseudomonas fluorescens]
Molecule type	amino acid
Query Length	178
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism *only top 20 will appear* NEW

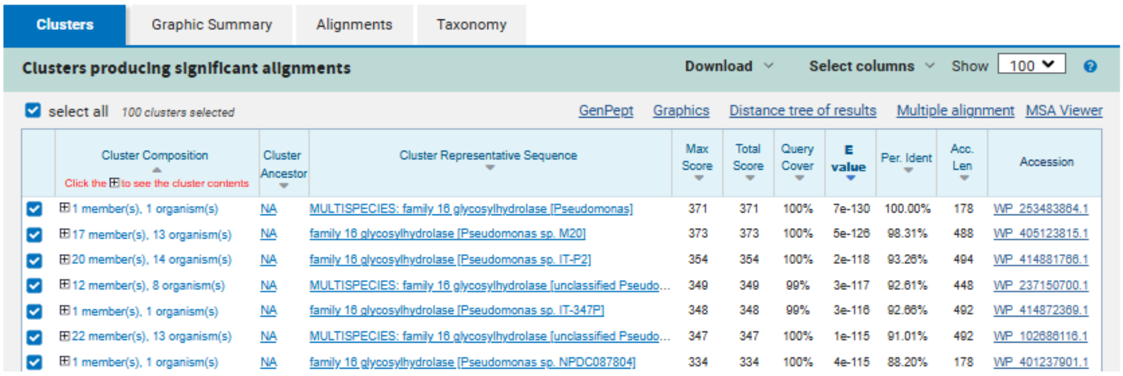
Type common name, binomial, taxid or group name

Figure 29. The section of a BLAST result for a κ-carrageenase from *Pseudomonas fluorescens*, showing key search parameters: the program used (arrow 1), database searched (arrow 2), query sequence (arrow 3), and query length (arrow 4).



Putative conserved domains have been detected, click on the image below for detailed results.

Figure 30. BLAST results summary showing conserved hits domains matching a κ-carrageenase from *Pseudomonas fluorescens* and the distribution of homologous sequences, color-coded by alignment score along the 178-amino acid query.

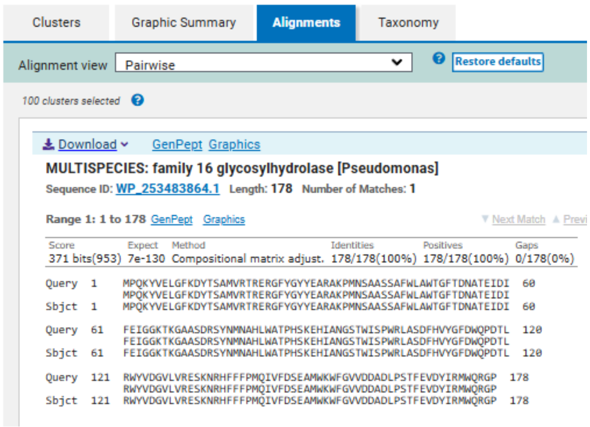


Clusters producing significant alignments

☑ select all 100 clusters selected

	Cluster Composition Click the ⊞ to see the cluster contents	Cluster Ancestor	Cluster Representative Sequence	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	⊞1 member(s), 1 organism(s)	NA	MULTISPECIES: family 16 glycosylhydrolase [Pseudomonas]	371	371	100%	7e-130	100.00%	178	WP_253483864.1
☑	⊞17 member(s), 13 organism(s)	NA	family 16 glycosylhydrolase [Pseudomonas sp. M20]	373	373	100%	5e-126	98.31%	488	WP_405123815.1
☑	⊞20 member(s), 14 organism(s)	NA	family 16 glycosylhydrolase [Pseudomonas sp. IT-P2]	354	354	100%	2e-118	93.26%	494	WP_414881766.1
☑	⊞12 member(s), 8 organism(s)	NA	MULTISPECIES: family 16 glycosylhydrolase [unclassified Pseudo...	349	349	99%	3e-117	92.61%	448	WP_237150700.1
☑	⊞1 member(s), 1 organism(s)	NA	family 16 glycosylhydrolase [Pseudomonas sp. IT-347P]	348	348	99%	3e-116	92.66%	492	WP_414872369.1
☑	⊞22 member(s), 13 organism(s)	NA	MULTISPECIES: family 16 glycosylhydrolase [unclassified Pseudo...	347	347	100%	1e-115	91.01%	492	WP_102686116.1
☑	⊞1 member(s), 1 organism(s)	NA	family 16 glycosylhydrolase [Pseudomonas sp. NPDC087804]	334	334	100%	4e-115	88.20%	178	WP_401237901.1

Figure 31. BLASTP result hits list of clusters showing matching sequences from the database, ranked by similarity. Each entry includes: a link to its database file, a link to the detailed alignment with your query, a bit score (higher is better), query coverage (higher is better), and an E-value (lower is better). The best matches appear first with the highest bit scores and lowest E-values.



Alignment view: Pairwise

100 clusters selected

MULTISPECIES: family 16 glycosylhydrolase [Pseudomonas]
Sequence ID: WP_253483864.1 Length: 178 Number of Matches: 1

Range 1: 1 to 178 GenPept Graphics

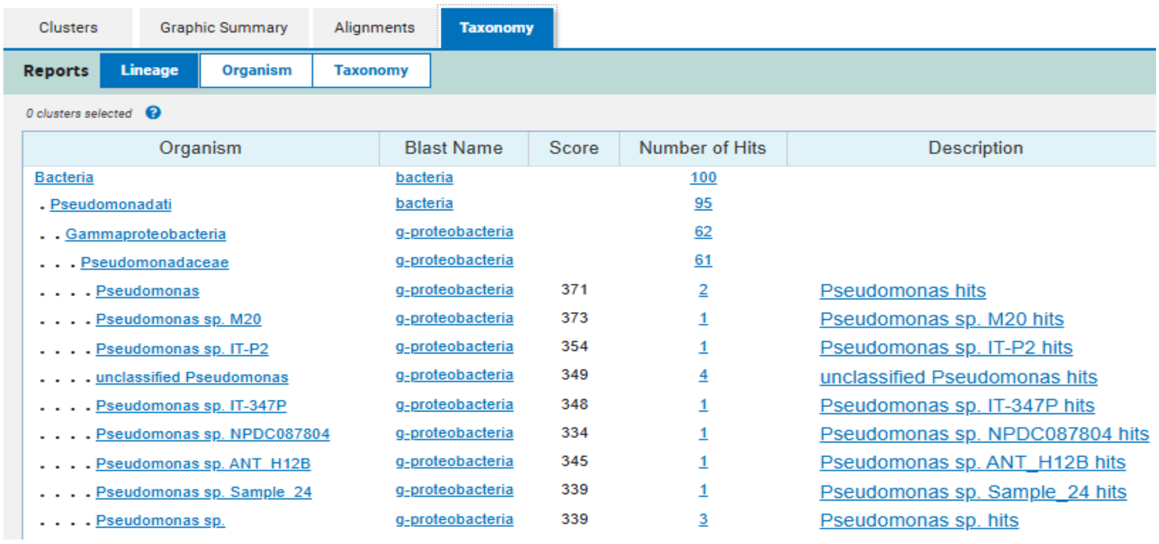
```
Score           Expect   Method                         Identities     Positives      Gaps
371 bits(953)   7e-130   Compositional matrix adjust.   178/178(100%)  178/178(100%)  0/178(0%)

Query  1    MPQKYVELGFKDYTSAMVRTRERGFYGYYEARAKPMNSAASSAFWLAWTGFTDNATEIDI  60
            MPQKYVELGFKDYTSAMVRTRERGFYGYYEARAKPMNSAASSAFWLAWTGFTDNATEIDI
Sbjct  1    MPQKYVELGFKDYTSAMVRTRERGFYGYYEARAKPMNSAASSAFWLAWTGFTDNATEIDI  60

Query  61   FEIGGKTKGAASDRSYNMNAHLWATPHSKEHIANGSTWISPWRLASDFHVYGFDWQPDTL  120
            FEIGGKTKGAASDRSYNMNAHLWATPHSKEHIANGSTWISPWRLASDFHVYGFDWQPDTL
Sbjct  61   FEIGGKTKGAASDRSYNMNAHLWATPHSKEHIANGSTWISPWRLASDFHVYGFDWQPDTL  120

Query  121  RWYVDGVLVRESKNRHFFFPMQIVFDSEAMWKWFGVVDDADLPSTFEVDYIRMWQRGP  178
            RWYVDGVLVRESKNRHFFFPMQIVFDSEAMWKWFGVVDDADLPSTFEVDYIRMWQRGP
Sbjct  121  RWYVDGVLVRESKNRHFFFPMQIVFDSEAMWKWFGVVDDADLPSTFEVDYIRMWQRGP  178
```

Figure 32. BLASTP pairwise sequence alignments of hits with key parameters including Score (bits), Expect value (E-value), Method (Compositional matrix adjust), Percent Identities (%), and Gaps (%).



0 clusters selected

Organism	Blast Name	Score	Number of Hits	Description
Bacteria	bacteria		100	
. Pseudomonadati	bacteria		95	
. . Gammaproteobacteria	g-proteobacteria		62	
. . . Pseudomonadaceae	g-proteobacteria		61	
. . . . Pseudomonas	g-proteobacteria	371	2	Pseudomonas hits
. . . . Pseudomonas sp. M20	g-proteobacteria	373	1	Pseudomonas sp. M20 hits
. . . . Pseudomonas sp. IT-P2	g-proteobacteria	354	1	Pseudomonas sp. IT-P2 hits
. . . . unclassified Pseudomonas	g-proteobacteria	349	4	unclassified Pseudomonas hits
. . . . Pseudomonas sp. IT-347P	g-proteobacteria	348	1	Pseudomonas sp. IT-347P hits
. . . . Pseudomonas sp. NPDC087804	g-proteobacteria	334	1	Pseudomonas sp. NPDC087804 hits
. . . . Pseudomonas sp. ANT_H12B	g-proteobacteria	345	1	Pseudomonas sp. ANT_H12B hits
. . . . Pseudomonas sp. Sample_24	g-proteobacteria	339	1	Pseudomonas sp. Sample_24 hits
. . . . Pseudomonas sp.	g-proteobacteria	339	3	Pseudomonas sp. hits

Figure 33. Taxonomic distribution of BLASTP matches, showing the species and higher-level groups of the most significant hits, which helps infer the evolutionary context of the query sequence.

5.6. Multiple Sequence Alignment

Multiple sequence alignment (MSA) is a key tool in bioinformatics. It is used to line up and compare several biological sequences, like DNA, RNA, or protein sequences. MSA is important because it helps scientists find parts of a sequence that are the same across different species (conserved regions). It also shows where sequences have added (insertions) or lost (deletions) material over time. By revealing these patterns, MSA helps us understand how the structure, function, and evolution of these sequences are related (Ali *et al.* 2024).

5.6.1. Multiple Sequence Alignment Techniques

Multiple Sequence Alignment (MSA) places related sequences together to find common areas and see which parts match up across different sequences. It starts by comparing sequences in pairs. Two main methods for this are Needleman–Wunsch and Smith-Waterman. They use a scoring system to find the best match by checking for similarities, differences, and gaps.

Progressive alignment methods, like ClustalW, use these pairwise comparisons to build a full multiple alignment. ClustalW first compares all sequences to build a family tree. It then aligns the most similar sequences first and slowly adds the others, inserting gaps where needed to improve the match. This creates a final "global" alignment.

Newer tools improve on this process:

- **MUSCLE**: Builds better trees and refines the alignment in repeating steps
- **MAFFT**: Uses a fast method to find similar regions quickly
- **T-Coffee**: Combines results from many different alignments
- **Clustal Omega**: Uses information from families of related sequences

Choosing the right tool depends on your data. If the sequences are very different, modern tools like MAFFT usually work better than older global aligners like ClustalW. Also, working with a lot of long sequences takes more computing power. Finally, the type of sequence you have, like specific protein parts, guides the best settings for penalties and scoring to get the best alignment (Ali *et al.* 2024).

5.6.2. Multiple Sequence Alignment Visualization

5.6.2.1. Importance of Visualization

Visualizing multiple sequence alignments (MSAs) is a critical step that transforms raw sequence data into interpretable graphics, enabling the identification of conserved regions, variable sites, and evolutionary patterns that are essential for understanding gene function and structure (Zhou *et al.* 2022).

5.6.2.2. Foundational Methods: Sequence Logos

Sequence logos provide a powerful graphical summary of conservation, where the stack height indicates a position's conservation and individual symbol heights represent residue frequency, as shown in the sequence logo visualization (**Figure 34**).

While tools like WebLogo make this technique accessible, traditional logos often oversimplify heterogeneous datasets by assuming sequence uniformity (Schneider and Stephens 1990).

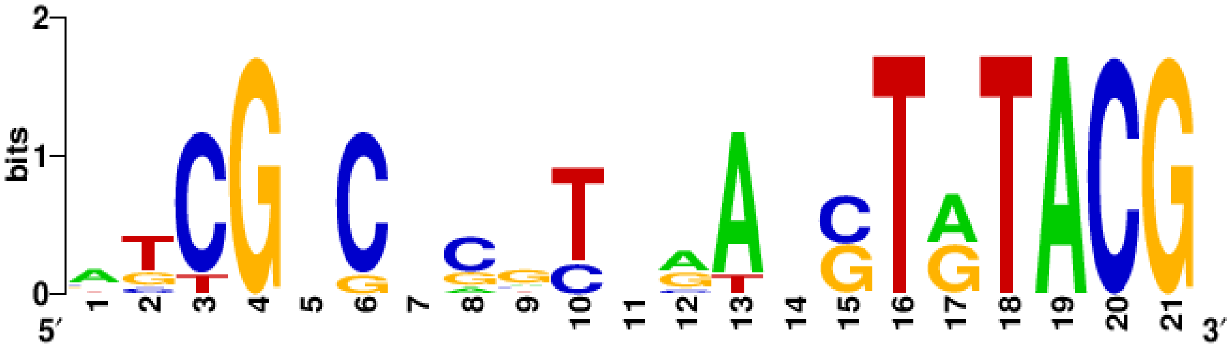


Figure 34. Sequence logo of the alignment. This visualization highlights the most conserved DNA bases at each position.

5.4.2.3. Advanced Logo-Based Tools

Next-generation tools address the limitations of traditional logos:

- **MetaLogo** clusters sequences into phylogenetic or functional groups before generating logos, effectively visualizing heterogeneity across sequence types
- **CodonLogo** extends the concept by treating codons as single units, particularly useful for analyzing mRNA regulatory signals
- **Seq2Logo** incorporates sequence weighting and pseudo-counts to correct for data redundancy and provide more accurate representations

5.6.2.4. Alternative Visualization Approaches

Sequence Bundles maintain a direct correspondence between sequences and their graphical representation, making them ideal for identifying covariant sites and sequence motifs. The Alvis platform implements this method with interactive tools for joint analysis of MSAs and phylogenetic trees (Schwarz *et al.* 2016).

5.6.2.5. Integrated Visualization with ggmsa

The ggmsa R package represents a comprehensive approach by integrating multiple visualization methods including sequence logos, bundles, and stacked alignments. This toolkit addresses the critical need to identify conserved and variable regions in MSAs, transforming sequence features into understandable visual representations. It enables researchers to explore sequence conservation patterns while correlating MSA data with phenotypic traits, molecular structures, and functional annotations, thereby supporting the investigation of sequence-structure-function relationships.

The package's utility is demonstrated in its application to bacterial kinase domains and phenylalanine hydroxylase (**Figure 35**). As a freely available, open-source Bioconductor package, ggmsa provides scalable visualizations that broaden the scope of sequence investigation and assist researchers in discovering MSA insights (Zhou *et al.* 2022).

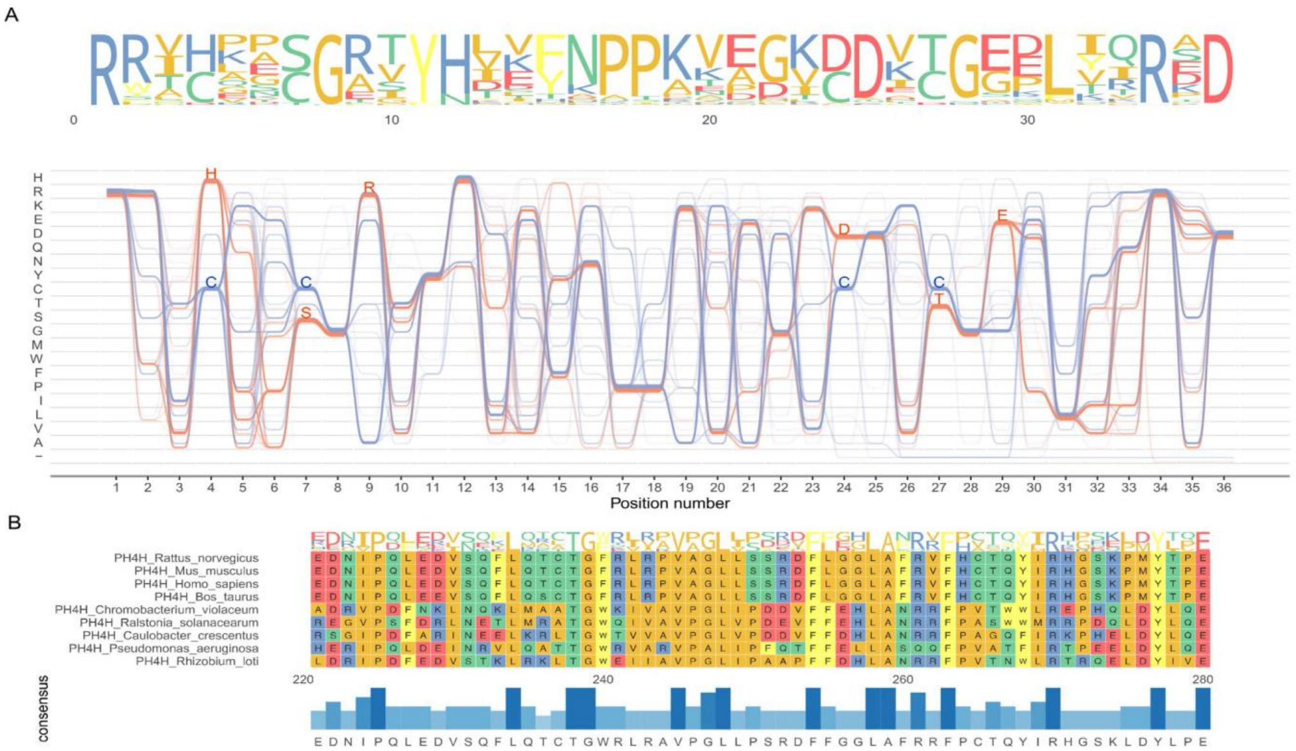


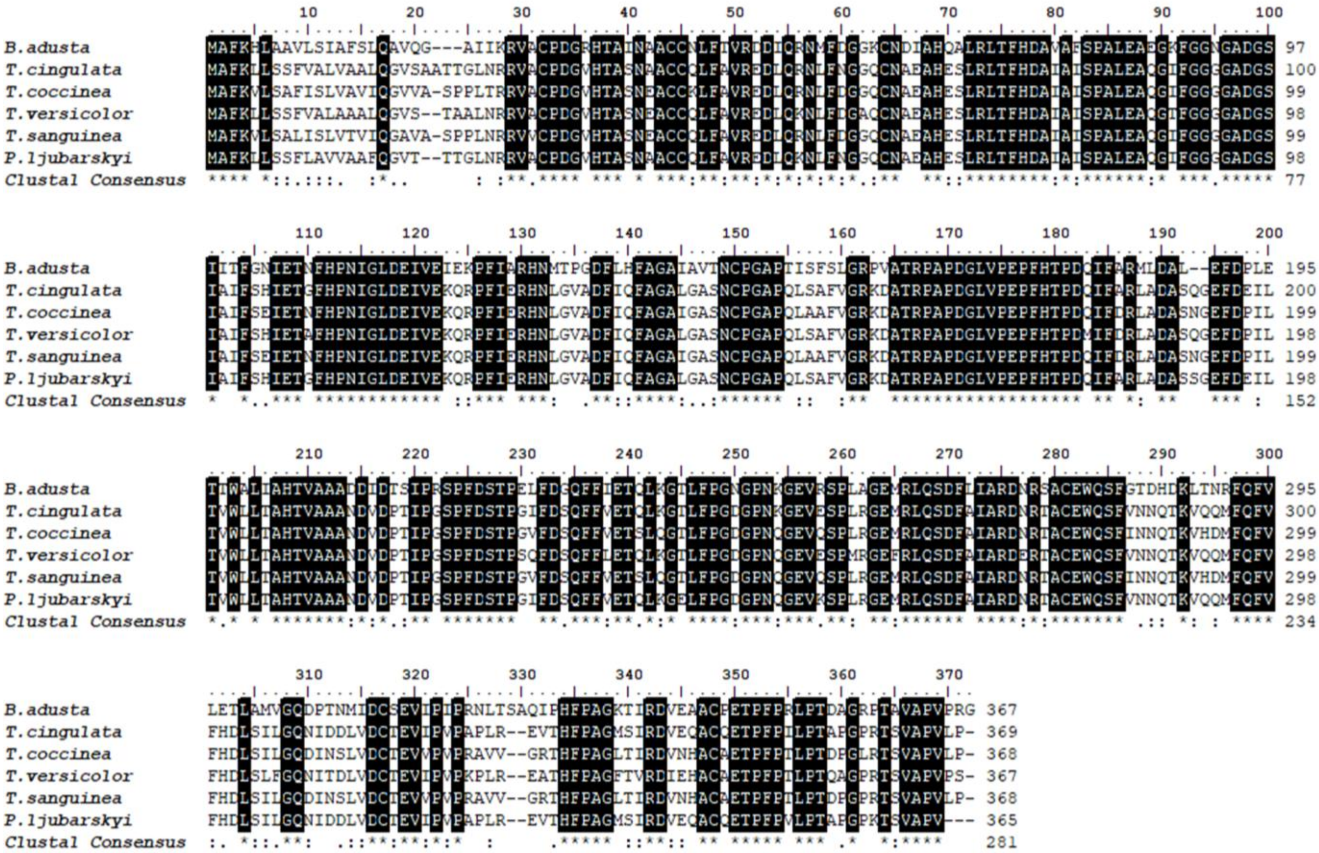
Figure 35. Visualizing multiple sequence alignments with ggmsa. A. Sequence logos and bundles of the Adenylate Kinase Lid Domain (AKL) reveal distinct conservation patterns between gram-negative (orange) and gram-positive (purple) bacteria, highlighting group-specific residue relationships critical for structural stability. B. Stacked MSA plot of phenylalanine hydroxylase (PH4H) with chemical coloring and annotations (sequence logos and residue frequency bars) demonstrates ggmsa's integrated approach to exploring sequence conservation and variation (Zhou *et al.* 2022).

5.6.2.6. Large-Scale MSA Visualization

For genomic-scale analyses, NX4 provides specialized web-based visualization of alignments containing thousands of sequences. It employs color-blind friendly palettes and a focus-plus-context approach to efficiently navigate and identify regions of high genetic diversity (Solano-Roman *et al.* 2019).

5.6.2.7. Practical Application

An MSA of lignin peroxidase from *Bjerkandera adusta* fungi was generated using ClustalW with its homologs identified via BLAST. The resulting alignment, visualized with BioEdit, reveals extensive conservation across these fungal sequences (**Figure 36**), highlighting functionally critical regions through conserved amino acid patterns (Bakli *et al.* 2025).



```
                            10        20        30        40        50        60        70        80        90       100
B.adusta           MAFKHLAAVLSIAFSLQAVQG---AIIKRVACPDGRHTAINAACCNLFTVRDDIQRNMFDGGKCNDIAHQALRLTFHDAVAFSPALEAEGKFGGNGADGS 97
T.cingulata        MAFKLLSSFVALVAALQGVSAATTGLNRRVACPDGVHTASNAACCQLFAVREDLQRNLFNGGQCNAEAHESLRLTFHDAIAISPALEAQGIFGGGGADGS 100
T.coccinea         MAFKVLSAFISLVAVIQGVVA-SPPLTRRVACPDGVHTASNEACCKLFAVREDLQRNLFDGGQCNAEAHESLRLTFHDAIAISPALEAQGIFGGGGADGS 99
T.versicolor       MAFKLLSSFVALAAALQGVS--TAALNRRVACPDGVHTASNEACCQLFAVREDLQKNLFDGAQCNAEAHESLRLTFHDAIAISPALEAQGIFGGGGADGS 98
T.sanguinea        MAFKVLSALISLVTVIQGAVA-SPPLNRRVVCPDGVHTASNEACCQLFAVREDLQRNLFDGGQCNAEAHESLRLTFHDAIAISPALEAQGIFGGGGADGS 99
P.ljubarskyi       MAFKLLSSFLAVVAAFQGVT--TTGLNRRVACPDGVHTASNAACCQLFAVREDLQKNLFNGGQCNAEAHESLRLTFHDAIAISPALEAQGIFGGGGADGS 98
Clustal Consensus  **** *::.:::.  :*..        : :**.**** *** * ***:**:**:*:*:*:*:*.:**  **::********:*:******:* ***.***** 77

                           110       120       130       140       150       160       170       180       190       200
B.adusta           IITFGNIETNFHPNIGLDEIVEIEKPFIARHNMTPGDFLHFAGAIAVTNCPGAPTISFSLGRPVATRPAPDGLVPEPFHTPDQIFARMLDAL--EFDPLE 195
T.cingulata        IAIFSHIETGFHPNIGLDEIVEKQRPFIERHNLGVADFIQFAGALGASNCPGAPQLSAFVGRKDATRPAPDGLVPEPFHTPDQIFARLADASQGEFDEIL 200
T.coccinea         IAIFSEIETNFHPNIGLDEIVEKQRPFIERHNLGVADFIQFAGAIGASNCPGAPQLAAFVGRKDATRPAPDGLVPEPFHTPDQIFDRLADASNGEFDPIL 199
T.versicolor       IAIFSHIETAFHPNIGLDEIVEKQRPFIERHNLGVADFIQFAGALGASNCPGAPQLSAFVGRKDATRPAPDGLVPEPFHTPDMIFDRLADASQGEFDPIL 198
T.sanguinea        IAIFSEIETNFHPNIGLDEIVEKQRPFIERHNLGVADFIQFAGAIGASNCPGAPQLAAFVGRKDATRPAPDGLVPEPFHTPDQIFDRLADASNGEFDPIL 199
P.ljubarskyi       IAIFSHIETGFHPNIGLDEIVEKQRPFIERHNLGVADFIQFAGALGASNCPGAPQLSAFVGRKDATRPAPDGLVPEPFHTPDQIFARLADASSGEFDEIL 198
Clustal Consensus  *  *..*** ************ ::*** ***:   .**::****:..:****** ::  :**  ***************** ** *: **    *** : 152

                           210       220       230       240       250       260       270       280       290       300
B.adusta           TIWALIAHTVAAADDIDTSIPRSPFDSTPELFDGQFFIETQLKGTLFPGNGPNKGEVRSPLAGEMRLQSDFLIARDNRSACEWQSFGTDHDKLTNRFQFV 295
T.cingulata        TVWLLTAHTVAAANDVDPTIPGSPFDSTPGIFDSQFFVETQLKGTLFPGDGPNKGEVESPLRGEMRLQSDFAIARDNRTACEWQSFVNNQTKVQQMFQFV 300
T.coccinea         TVWLLTAHTVAAANDVDPTIPGSPFDSTPGVFDSQFFVETSLQGTLFPGDGPNQGEVQSPLRGEMRLQSDFAIARDNRTACEWQSFINNQTKVHDMFQFV 299
T.versicolor       TVWLLTAHTVAAANDVDPTIPGSPFDSTPSQFDSQFFLETQLKGTLFPGDGPNQGEVESPMRGEFRLQSDFAIARDERTACEWQSFVNNQTKVQQMFQFV 298
T.sanguinea        TVWLLTAHTVAAANDVDPTIPGSPFDSTPGVFDSQFFVETSLQGTLFPGDGPNQGEVQSPLRGEMRLQSDFAIARDNRTACEWQSFINNQTKVHDMFQFV 299
P.ljubarskyi       TVWLLTAHTVAAANDVDPTIPGSPFDSTPGIFDSQFFVETQLKGELFPGDGPNQGEVKSPLRGEMRLQSDFAIARDNRTACEWQSFVNNQTKVQQMFQFV 298
Clustal Consensus  *.* * *******:*:*.:** *******  **.***:**.*:* ****:***:***.**: **:****** ****:*:******* .:: *: : **** 234

                           310       320       330       340       350       360       370
B.adusta           LETLAMVGQDPTNMIDCSEVIPIPRNLTSAQIPHFPAGKTIRDVEAACPETPFPRLPTDAGRPTAVAPVPRG 367
T.cingulata        FHDLSILGQNIDDLVDCTEVIPVPAPLR--EVTHFPAGMSIRDVEQACQETPFPILPTAPGPRTSVAPVLP- 369
T.coccinea         FHDLSILGQDINSLVDCTEVVPVPRAVV--GRTHFPAGLTIRDVNHACAETPFPTLPTDPGLRTSVAPVLP- 368
T.versicolor       FHDLSLFGQNITDLVDCTEVIPVPKPLR--EATHFPAGFTVRDIEHACAETPFPTLPTQAGPRTSVAPVPS- 367
T.sanguinea        FHDLSILGQDINSLVDCTEVVPVPRAVV--GRTHFPAGLTIRDVNHACAETPFPTLPTDPGPRTSVAPVLP- 368
P.ljubarskyi       FHDLSILGQNIDDLVDCTEVIPVPAPLR--EVTHFPAGMSIRDVEQACQETPFPVLPTAPGPKTSVAPV--- 365
Clustal Consensus  :. *::.**:  .::**:**:*:*  :     .***** ::**:: ** ***** *** .*  *:****      281
```

Figure 36. Multiple sequences alignment of lignin peroxidase (Gi: 588479560) from *Bjerkandera adusta* with and its five homologous enzymes. The indicated species of homologous enzymes are in the order : *Trametes cingulata*, *Trametes coccinea*, *Trametes versicolor*, *Trametes sanguinea*, and *Pilatotrama ljubarskyi*. Identical amino acids are shown with a black background. The Clustal consensus symbols indicate the amount of conservation ('*': Exact, ':': Conserved Substitution, '.': Semi-conserved substitution) (Bakli *et al.* 2025).

5.7. Phylogenetic Analysis

Phylogenetic analysis is the study of the evolutionary relationships between different organisms. It is based on the idea that all organisms share a common ancestor and that their differences are the result of changes accumulated over time. Phylogenetic trees are the primary graphical representation of these evolutionary relationships (**Fig. 37**). These branching diagrams illustrate how various organisms are related to one another, with the tree branches symbolizing the evolutionary events that have led to the diversification of life (Ali *et al.* 2024).

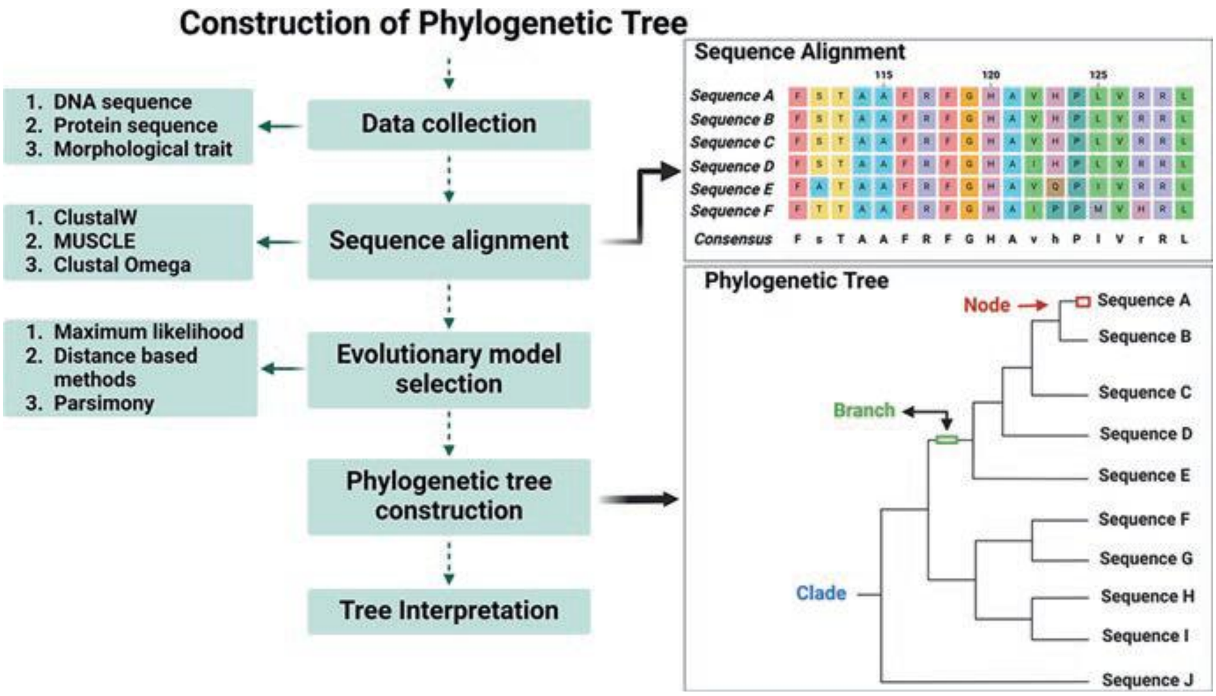

Construction of Phylogenetic Tree
1. DNA sequence
2. Protein sequence
3. Morphological trait
Data collection
1. ClustalW
2. MUSCLE
3. Clustal Omega
Sequence alignment
1. Maximum likelihood
2. Distance based methods
3. Parsimony
Evolutionary model selection
Phylogenetic tree construction
Tree Interpretation
Sequence Alignment
115
120
125
Sequence A F S T A A F R F G H A V H P L V R R L
Sequence B F S T A A F R F G H A V H P L V R R L
Sequence C F S T A A F R F G H A V H P L V R R L
Sequence D F S T A A F R F G H A I H P L V R R L
Sequence E F A T A A F R F G H A V Q P I V R R L
Sequence F F T T A A F R F G H A I P P M V H R L
Consensus F s T A A F R F G H A v h P I V r R L
Phylogenetic Tree
Node
Branch
Clade
Sequence A
Sequence B
Sequence C
Sequence D
Sequence E
Sequence F
Sequence G
Sequence H
Sequence I
Sequence J

Figure 37. Overview of the Methodology for Constructing a Phylogenetic Tree. The construction of a phylogenetic tree involves multiple key steps. Homologous sequences are aligned to identify common regions. An appropriate evolutionary model accounting for substitution patterns is selected. A tree topology is derived using neighbor-joining and likelihood optimization under the model parameters. The resulting tree visually depicts the evolutionary relationships hypothesized among taxa based on genetic/molecular data, offering insights into diversification (Ali *et al.* 2024).

There are four main types of phylogenetic trees. Unrooted trees depict the relationships among a group of organisms without identifying a single common ancestor. In contrast, rooted trees have a designated common ancestor and show the evolutionary history of the group from that starting point. Ultrametric trees have branches of equal length, which implies that evolutionary change has occurred at a constant rate since the common ancestor. Additive trees have branch lengths that are proportional to the amount of evolutionary change that has occurred along each branch (Nei 2019).

The construction of a phylogenetic tree typically involves several key steps. The first step is to collect data suitable for inferring evolutionary relationships, such as DNA sequences, protein sequences, or morphological characteristics. The next step is to align this data, which means arranging the sequences so that homologous, or corresponding, positions are matched. The final step is to construct the tree itself using computational methods. These methods are generally categorized as distance-based, character-based, or model-based. Distance-based methods, like the neighbor-joining algorithm, calculate the overall genetic distance between organisms and group the closest ones together iteratively. Character-based methods, such as maximum parsimony and maximum likelihood, analyze shared and derived characteristics. Maximum parsimony seeks the tree that requires the fewest evolutionary changes, while maximum likelihood finds the tree that is most probable given the observed data. Model-based methods, including Bayesian inference, use statistical models to estimate evolutionary relationships. Bayesian inference calculates the probability of each possible tree, allowing it to account for uncertainty and provide confidence estimates for the results (Zou *et al.* 2024).

The interpretation of phylogenetic trees relies on understanding specific terms. A clade is a group of organisms that includes a common ancestor and all its descendants. A branch is a line connecting two parts of the tree, representing a lineage. A node is a point where two or more branches meet, representing a common ancestor. By analyzing these elements, phylogenetic trees can be used to determine the order in which species evolved, identify common ancestors, and infer the functions of genes and proteins (Ali *et al.* 2024).

Chapter 6. Structural Bioinformatics

6.1. Introduction

Structural bioinformatics is a field of science that focuses on predicting and analyzing the 3D shapes of biological molecules like proteins, DNA, and RNA. Because a molecule's structure determines its function, understanding these shapes helps us answer fundamental questions about how life works (Schwalbe *et al.* 2024).

Researchers in this field use computers and advanced algorithms to build models of these complex molecules. These models allow them to simulate how proteins fold, predict how they will interact with each other, and identify important sites where drugs might bind. This makes structural bioinformatics a powerful tool for discovering new medicines (Mahanandia *et al.* 2025).

By comparing structures, scientists can also learn about evolutionary relationships between organisms and understand the rules that govern how molecules assemble. In short, structural bioinformatics combines data from experiments with computational power to generate new knowledge and drive innovation in biology and medicine (Israr *et al.* 2024).

6.2. Protein Structure

A fundamental concept in bioinformatics is that a protein’s specific three-dimensional shape is what allows it to perform its biological function. This direct link between form and role is known as the structure-function relationship. The precise folding of a protein’s amino acid chain creates unique shapes and surfaces that enable activities like catalyzing chemical reactions, recognizing other molecules, and binding to DNA or other proteins (Gromiha *et al.* 2025).

Proteins form local structural patterns such as alpha-helices, beta-sheets, and loops. These elements are stabilized by weak chemical interactions—particularly hydrogen bonds—that hold the overall structure together. These interactions can occur within a single protein chain (stabilizing its tertiary structure) or between different protein subunits (forming quaternary structure) (Israr *et al.* 2024).

Protein structure is organized into four hierarchical levels, as illustrated in the mind map (**Fig. 38**): primary (the amino acid sequence), secondary (local folding patterns like alpha-helices and beta-sheets), tertiary (the overall 3D shape of a single chain), and quaternary (the assembly of multiple protein subunits). While a protein's unique 3D structure is entirely determined by its amino acid sequence, the rules governing how it folds are not fully understood—making structure prediction one of the key challenges in bioinformatics (Zvelebil and Baum 2007).

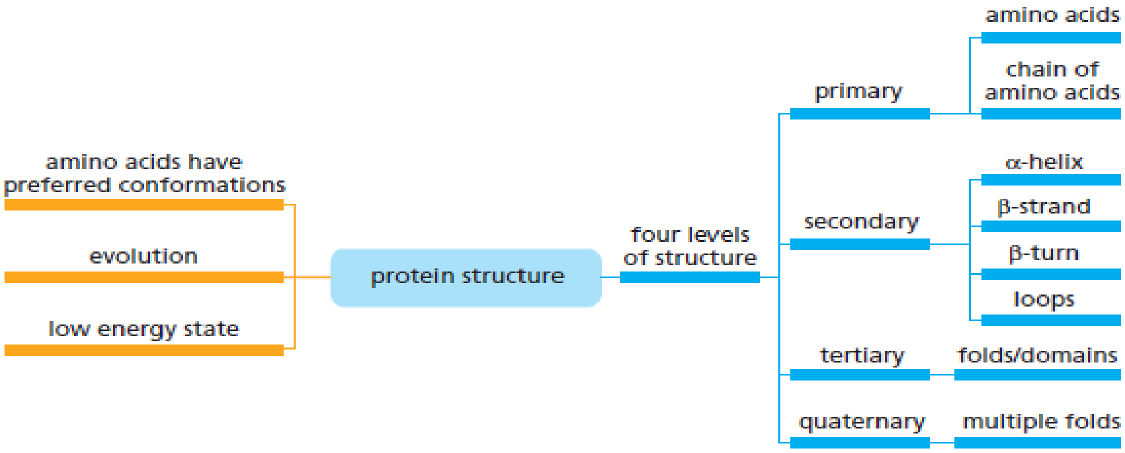
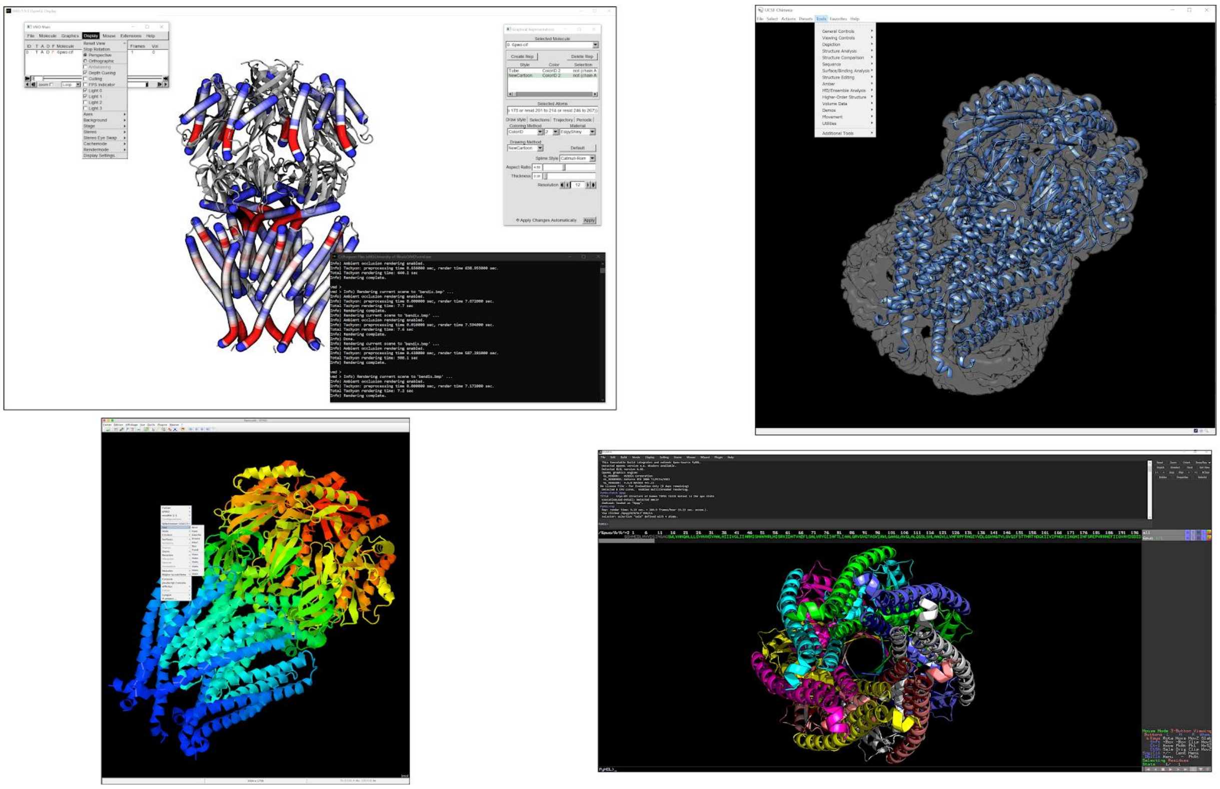


Figure 38. A mind map visualization of aspects of protein structure (Zvelebil and Baum 2007).

Structural bioinformatics focuses on analyzing these molecular architectures and interactions using computational tools. Researchers study how proteins fold, how they bind to other molecules, and how their physical organization supports their biological roles. By understanding these principles, scientists can better predict protein behavior and design interventions—for example, developing drugs that target a specific protein's shape (Robin *et al.* 2024).

6.3. Structure Visualization

In structural bioinformatics, visualizing protein structures is an essential skill. It helps researchers see and understand the complex three-dimensional shapes of proteins, which is key to figuring out how they work. Specialized software tools allow scientists to view molecules as both static images and dynamic simulations, helping them identify how different parts of a molecule interact. These visualizations make it possible to study and explain the molecular mechanisms behind biological activity (Martinez *et al.* 2020). veral standard visualization methods are commonly used, each highlighting different aspects of a protein's structure (See **Fig. 39**):



```
Info) Ambient occlusion rendering enabled.
Info) Tachyon: preprocessing time 0.656000 sec, render time 638.953000 sec.
Total Tachyon rendering time: 640.2 sec
Info) Rendering complete.

vmd >
vmd > Info) Rendering current scene to 'bendix.bmp' ...
Info) Ambient occlusion rendering enabled.
Info) Tachyon: preprocessing time 0.000000 sec, render time 7.672000 sec.
Total Tachyon rendering time: 7.7 sec
Info) Rendering complete.
Info) Rendering current scene to 'bendix.bmp' ...
Info) Ambient occlusion rendering enabled.
Info) Tachyon: preprocessing time 0.016000 sec, render time 7.594000 sec.
Total Tachyon rendering time: 7.6 sec
Info) Rendering complete.
Info) Done.
Info) Rendering current scene to 'bendix.bmp' ...
Info) Ambient occlusion rendering enabled.
Info) Tachyon: preprocessing time 0.438000 sec, render time 587.381000 sec.
Total Tachyon rendering time: 588.1 sec
Info) Rendering complete.

vmd >
vmd > Info) Rendering current scene to 'bendix.bmp' ...
Info) Ambient occlusion rendering enabled.
Info) Tachyon: preprocessing time 0.000000 sec, render time 7.172000 sec.
Total Tachyon rendering time: 7.2 sec
Info) Rendering complete.
```

Figure 39. A visual comparison of user interfaces and typical rendering features for VMD (top left), Chimera (top right), Jmol (bottom left) and PyMol (bottom right) software tools. A largely alpha-helical protein channel is used as example, highlighting features related to its secondary structure (Martinez *et al.* 2020).

a. Cartoon: This is one of the most popular styles. It simplifies the protein to emphasize its secondary structure. Alpha-helices are shown as spiral ribbons or coils, beta-strands are drawn as flat arrows, and connecting loops are represented by thin tubes or lines. This view helps quickly identify folding patterns and major structural features.

b. Lines: In this simple style, bonds between atoms are shown as thin lines. It is a basic and clear way to view the entire molecular skeleton without extra detail, making it useful for seeing the overall shape or preparing publication-quality images.

c. Surface: This representation displays the outer surface of the molecule, showing its overall shape and topography. It is especially useful for understanding how other molecules, such as drugs or binding partners, might interact with the protein, since it displays the accessible areas and physical contours.

d. Sticks: This method shows atoms and bonds as small sticks, highlighting the chemical structure and bonding between atoms. It is particularly helpful for examining specific interactions—such as hydrogen bonds or chemical contacts—between amino acids or between a protein and a small molecule. Each of these visualization styles serves a different purpose, and bioinformatics researchers often switch between them to analyze various aspects of a protein's structure and function (Israr *et al.* 2024). (see an exemple in **Figure 40**).

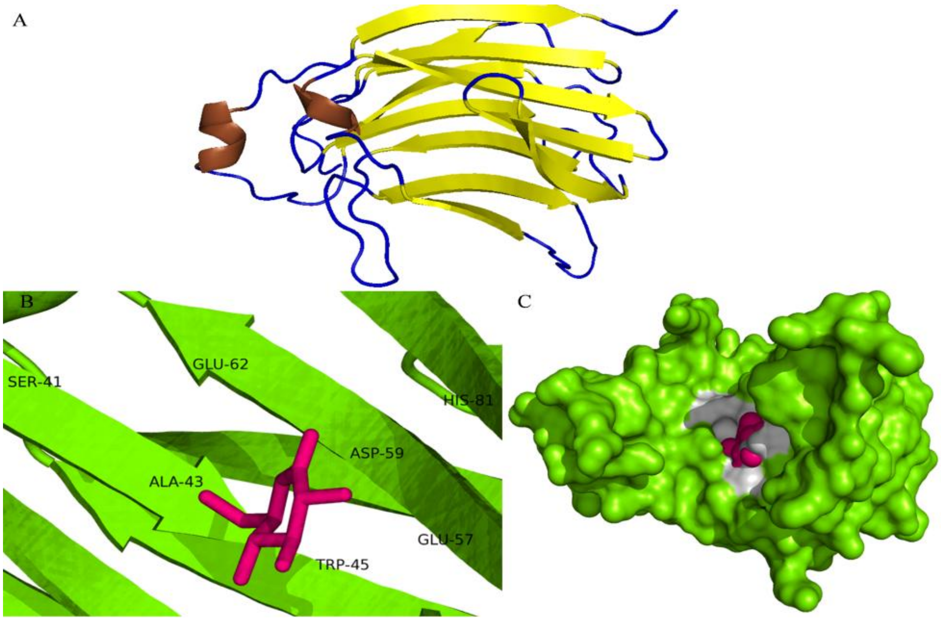


Figure 40. PyMOL visualizations of κ-carrageenase from Pseudomonas fluorescens. A. Cartoon representation (α-helices: brown, β-sheets: yellow, loops: blue). B. Ligand-binding pocket with key residues (grey) of the active site and ligand (pink line). C. Molecular surface (green) and ligand (pink spheres) (Bakli *et al.* 2022).

6.4. DNA Structure Background

The structure of DNA was first discovered by Watson and Crick, with important contributions from Rosalind Franklin. DNA is made of three main parts: a phosphate group, a sugar (deoxyribose), and one of four nitrogenous bases—adenine (A), thymine (T), cytosine (C), or guanine (G). These components form repeating units called nucleotides. The phosphate and sugar form the backbone of the DNA molecule, while the bases carry genetic information. DNA has a double-helix structure, like a twisted ladder. The bases pair up specifically: A always bonds with T, and C always bonds with G. These pairs are held together by hydrogen bonds, which help stabilize the double helix (Israr *et al.* 2024).

In structural bioinformatics, researchers often study how small molecules or drugs interact with DNA, which is important for understanding genetics and designing therapies (Zhang *et al.* 2025).

6.5. Molecular Interactions and Contact Analysis

Molecular interactions refer to the physical and chemical forces that occur between atoms and molecules. These interactions are essential for maintaining the three-dimensional structure of biological molecules and for enabling their functions. Common types of interactions include hydrogen bonds, electrostatic attractions, hydrophobic effects, and van der Waals forces. These forces play critical roles in processes such as protein folding, enzyme-substrate binding, and DNA-protein recognition (Lipomi and Ramji 2024).

In bioinformatics, the study of molecular interactions involves identifying and analyzing regions where molecules come into close proximity. One common approach is to calculate molecular "contacts" based on the distance between atoms. A widely used method involves measuring the Euclidean distance between atoms and applying a distance cutoff—typically between 3.5 and 5.0 Å—to determine whether two atoms are interacting. If the distance between them is less than the cutoff, they are considered to be in contact (Martins *et al.* 2018).

While distance-based methods are useful, they have limitations. For example, they may miss interactions in crowded molecular environments or fail to account for the geometry and chemical properties of atoms. To address these challenges, advanced techniques like Delaunay triangulation and machine learning algorithms are increasingly used. These methods incorporate additional factors such as atomic properties, spatial arrangements, and bond orientations to improve the accuracy of contact detection (Zhang *et al.* 2025).

Accurately identifying molecular contacts is vital for applications such as molecular docking, protein structure prediction, and drug design. It helps researchers understand how proteins interact with ligands, how mutations affect stability, and how to design molecules with desired binding properties. **Table 11** provides commonly used distance criteria for defining molecular contacts in structural bioinformatics (Israr *et al.* 2024).

Table 11. Distance criteria for contact definition (Israr *et al.* 2024).

Type	Max distance criteria
Hydrogen bond	3.9 Å
Hydrophobis interaction	5 Å
Ionic interaction	6 Å
Aromatic stacking	6 Å

6.6. Protein Data Bank (PDB) and Data Formats

The Protein Data Bank (PDB) is a global database that stores 3D structural information for important biological molecules, including proteins, DNA, and RNA. It is managed by the Worldwide Protein Data Bank (https://www.rcsb.org/), a collaboration of organizations such as RCSB PDB (USA), PDBe (Europe), PDBj (Japan), and BMRB. These groups work together to ensure that structural data are freely and publicly available to researchers around the world (Burley *et al.* 2024). The number of structures in the PDB has grown rapidly over the years, thanks to advances in experimental methods like X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy (Burley *et al.* 2024).

To store and share this structural data, two main file formats are used:

a. PDB Format: This is a traditional text-based format. It uses keywords (up to 6 letters) to label different types of data, such as atomic coordinates or secondary structure. However, this format has limitations. For example, it cannot represent molecules with more than 99,999 atoms or more than 62 chains.

b. PDBx/mmCIF Format: This is a more modern and flexible format. It uses a key-value system to store data, which allows for more detailed and accurate representation of large and complex molecules. Since 2014, this has been the standard format for the PDB archive. Files in this format use the extension « .cif ».

The shift to the mmCIF format has improved how researchers store, share, and analyze molecular structures, supporting better collaboration and discovery in bioinformatics (Israr *et al.* 2024).

6.7. Structure Comparison

6.7.1. Structural Alignment

Structural alignment is a method used to compare the 3D shapes of two or more proteins. Unlike simple sequence comparison, structural alignment looks at the overall folding pattern, which can reveal similarities even when the amino acid sequences are very different. This helps scientists understand evolutionary relationships and identify conserved functional regions (Ferrari and Guerra 2003).

To perform a structural alignment, one protein structure is rotated and moved until it matches another as closely as possible. This is usually done by aligning the positions of key atoms, such as those in the protein backbone. The quality of the alignment is measured using a value called RMSD (Root Mean Square Deviation), which calculates the average distance between matched atoms after alignment. A lower RMSD means the structures are more similar. RMSD is measured in Ångströms (Å), where 1 Å equals 10^{-10} meters. This technique is useful for studying how proteins evolve, how they function, and for identifying common structural motifs across different proteins (Sapundzhi *et al.* 2022).

6.7.2. Graph-Based Structural Signatures

Structural signatures, often called fingerprints, are simplified representations that capture key patterns in a protein’s 3D shape. These fingerprints allow researchers to efficiently compare large sets of proteins without relying on computationally expensive structural alignment methods like RMSD. Instead of comparing every atom, these approaches use graph-based techniques to encode meaningful structural information—such as distances between atom pairs or spatial relationships—into compact numerical vectors or patterns. This abstraction enables rapid comparison and analysis of protein structures at scale (Ferrari and Guerra 2003).

The integration of machine learning and linear algebra has further enhanced the utility of structural signatures. These techniques are now applied to diverse tasks such as grouping proteins with similar structures, identifying potential drug-binding molecules, predicting thermodynamic properties, and evaluating how genetic mutations might affect protein folding or function. By leveraging these efficient computational representations, researchers can extract meaningful biological insights from large structural datasets more effectively than with traditional alignment-based methods (Gaudelet *et al.* 2021).

6.8. Protein Structure Prediction

6.8.1. Protein Prediction Goal

The main goal of protein structure prediction is to determine the three-dimensional (3D) shape of a protein from its linear amino acid sequence. This is a fundamental challenge in biology because a protein's function is directly determined by its structure. This process, often called the "sequence-structure-function" paradigm, bridges a critical gap in our understanding of how genetic information leads to cellular activity and is summarized in **Fig 41** (Abdi *et al.* 2024)**.**

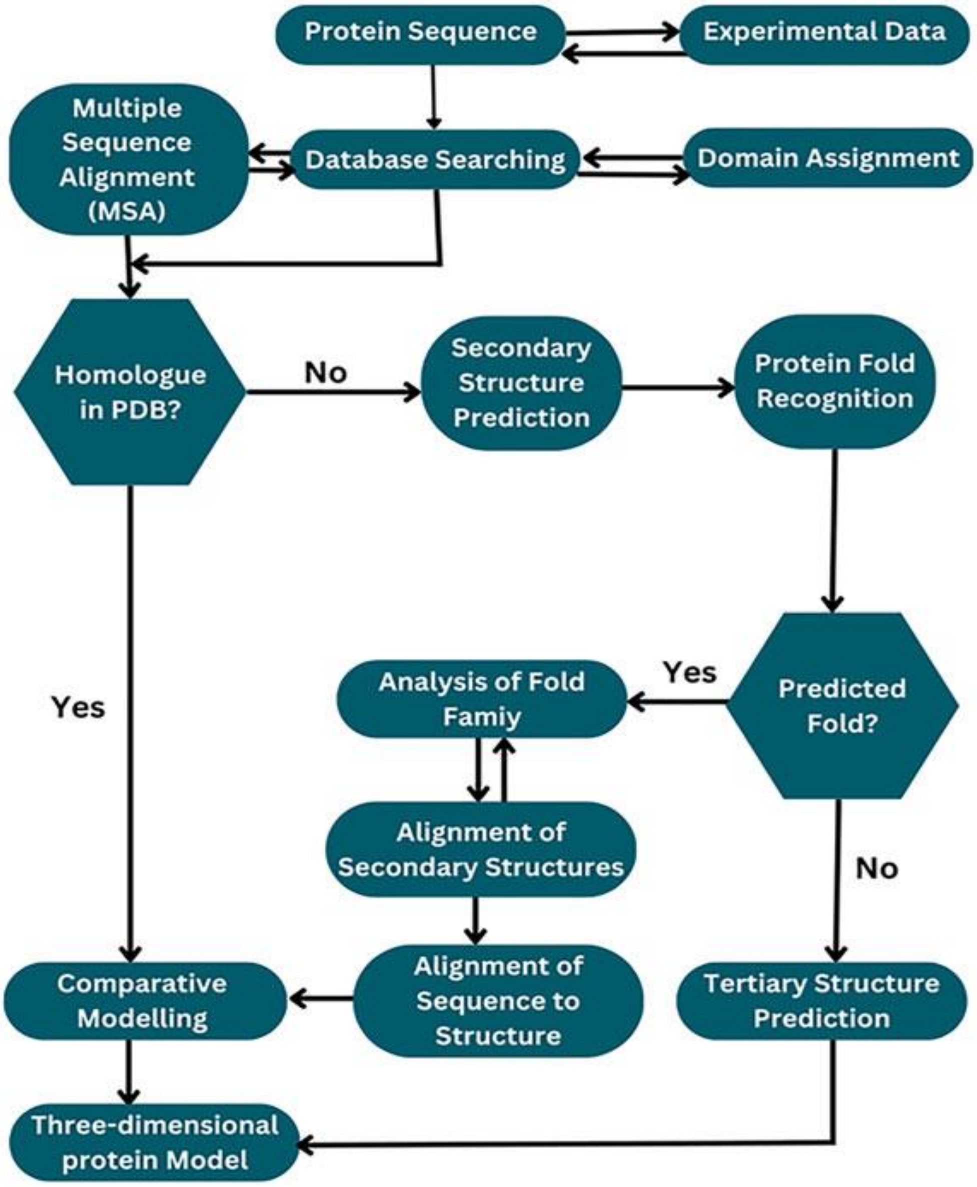

Protein Sequence
Experimental Data
Multiple Sequence Alignment (MSA)
Database Searching
Domain Assignment
Homologue in PDB?
No
Secondary Structure Prediction
Protein Fold Recognition
Yes
Analysis of Fold Famiy
Yes
Predicted Fold?
Alignment of Secondary Structures
No
Comparative Modelling
Alignment of Sequence to Structure
Tertiary Structure Prediction
Three-dimensional protein Model

Figure 41. Bioinformatics automation in protein structure prediction (Abdi *et al.* 2024)**.**

6.8.2. Multi-Step Bioinformatics Process

Solving this problem computationally involves a pipeline of interconnected steps. This pipeline is designed to extract the maximum amount of information from the sequence data itself. Key steps include identifying stable protein regions (domains), predicting local structures like helices and sheets (secondary structure), finding evolutionarily related sequences (similarity searches and multiple sequence alignment), and estimating which parts of the protein are exposed to water (solvent accessibility). The ultimate outputs are atomic-resolution 3D models, which are then checked for errors through validation (Edwards and Cottage 2003).

6.8.3. Challenges Driving Automation

Key challenges in protein structure prediction include a lack of similar templates, errors in sequence alignment, and difficulties modeling flexible regions. Furthermore, the massive scale of data from modern genomics makes manual analysis impossible. Because experimental methods for determining structure are slow and expensive, automating this process with bioinformatics tools is not just helpful—it is essential for large-scale progress (Kellici *et al.* 2024).

6.8.4. Computational Strategies

Automated tools typically follow one of two core strategies:

a. Template-Based Modeling: This strategy relies on finding evolutionarily related proteins with known structures (templates) in databases and using them as a guide to build the model. It is the most common and reliable approach when suitable templates exist.

b. Template-Free Modeling (*Ab Initio*): This strategy attempts to predict the structure from physical principles alone, without relying on a template. It is much more difficult and computationally intensive but is necessary for proteins with no evolutionary relatives.

These automated strategies power the various software tools that researchers use, many of which are discussed in the following sections. An overview of these tools is provided in **Table 12** (Abdi *et al.* 2024).

Table 12. Automation techniques in protein structure prediction (Abdi *et al.* 2024).

Software	Type
PSIPRED	Secondary structure of protein
PredictProtein	Secondary structure of protein and others
SABLE	Secondary structure of protein and solvent accessibility

SAM-T02	Secondary structure of protein
PONDR	Disordered region
PORTER	Secondary structure of protein
COILS	Coiled-coil region
GlobPlot	Disordered region
THHMM	Transmenmbrane domain
HHPred	Three-dimentional structure, homology modeling
SWISS-MODEL	Three-dimentional structure, homology modeling
FUGE	Three-dimentional structure, threading
HMMTOP	Transmenmbrane domain
MODELLER	Three-dimentional structure, homology modeling
Phyre	Three-dimentional structure, threading
Robetta	Three-dimentional structure, *ab initio*
SPARKS	Three-dimentional structure, threading

6.9. Evolution of Protein Structure Prediction Before AlphaFold

6.9.1. Historical Context

Before the revolutionary breakthrough of AlphaFold, predicting a protein's 3D structure was a complex, multi-stage process. Scientists relied heavily on methods like homology modeling, which builds a model for an unknown protein using the known structure of a related "template" protein. The field progressed through key innovations, summarized in **Figure 42**, that paved the way for modern AI-based prediction (Qiu *et al.* 2024).

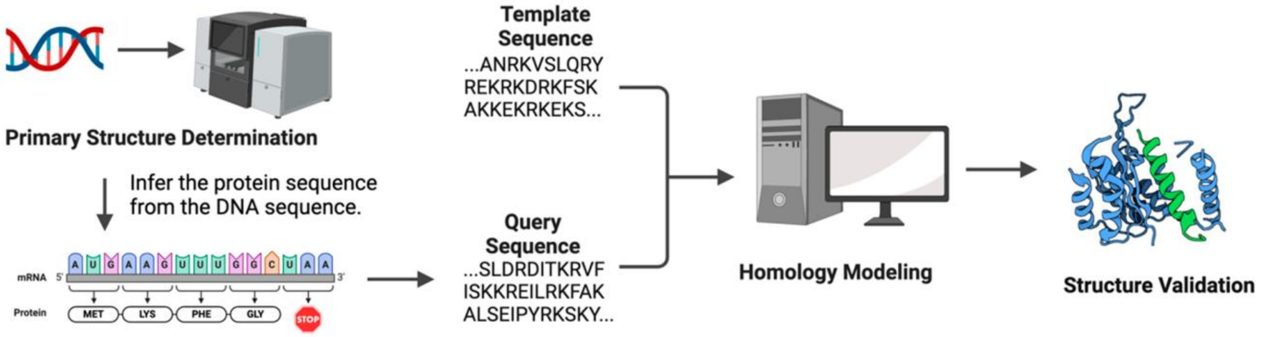


Figure 42. Stages of protein structure prediction. The foundational stage involves determining the DNA sequence that encodes the protein of interest. The next step is to infer the protein sequence from the DNA sequence. Homology modeling uses known protein structures as templates to predict the structure of a protein with an unknown structure but similar sequence. Lastly, validation of structure ensures the predicted structure's biological plausibility. This involves checks on stereochemical quality, energy evaluation, and comparison to known structural data (Qiu *et al.* 2024).

6.9.2. Key Advancements Pre-AlphaFold

Progress was driven by several critical developments:

a. Distant Homology Detection: Algorithms like HHpred improved the ability to find distant evolutionary relationships, providing better templates for modeling

b. Contact Map Predictions: Methods predicting which amino acids are in contact, based on evolutionary patterns, were enhanced by early deep learning applications

c. Hybrid Modeling Tools: By the late 2010s, tools like Rosetta and I-TASSER blurred the line between homology modeling and more advanced techniques

- Rosetta used a "fragment-based assembly" strategy combined with energy-based scoring.
- I-TASSER used iterative threading assembly refinement

Their accuracy began to approach methods that predict structure from scratch (*ab-initio*), directly leading to AlphaFold's development

6.9.3. Structure Validation

A critical step was (and remains) validation to ensure predicted models are accurate and physically possible. This involves

a. Geometry Checks: Tools like PROCHECK analyze bond geometry and dihedral angles.

b. Energy Assessment: Methods like ANOLEA evaluate the structure's potential energy to spot errors

c. Experimental Comparison: The ultimate validation is comparison with experimental data

For proteins with no known structure, validation often involves testing the predicted function in the lab (Qiu *et al.* 2024).

6.10. Homology Modeling

6.10.1. Definition

Homology modeling is a computational technique used to predict the three-dimensional (3D) structure of a protein. It is used when the structure has not been determined through experiments. The method is based on the principle that evolutionarily related proteins with similar sequences will fold into similar structures. This approach is also termed comparative modeling.

6.10.2. Methodological Framework

The process uses a protein with a known structure (a "template") to build a model for the unknown "target" protein. The key steps involved are:

a. Template Selection: Finding a suitable related protein structure in a database like the Protein Data Bank (PDB). This step relies on sequence similarity searches against structural databases

b. Sequence Alignment: Carefully aligning the amino acid sequences of the target and template proteins. Accurate alignment is critical for model quality

c. Model Building: Using software to construct a 3D model for the target based on the template's structure. Conserved regions are directly copied, while variable regions require modeling

d. Model Refinement: Improving the model's accuracy by correcting errors and relaxing the structure through methods like energy minimization. This step addresses steric clashes and geometric inaccuracies

e. Model Validation: Checking the final model's quality and reliability using various assessment tools. Common validation tools include MolProbity, PROCHECK, and QMEAN

6.10.3. Key Software Tools

Several important tools are used for this process:

a. SWISS-MODEL: An automated, web-based service that generates high-quality protein models

b. MODELLER: A popular software program that builds models based on comparative modeling

c. I-TASSER: A robust tool that uses a combination of homology modeling and other techniques to predict structure and function

d. AlphaFold: A revolutionary deep learning system known for predicting protein structures with exceptional, near-experimental accuracy

e. AlphaFold 3: An advanced version that can also predict how proteins interact with other molecules

f. Phyre^2: A widely used tool for protein structure prediction, especially for distant homologs

6.10.4. Importance and Limitations

This method is crucial for providing insights into protein function, predicting how mutations affect a protein, and facilitating drug discovery. It enables researchers to study proteins that are difficult to characterize experimentally. However, its accuracy can be limited by difficulties in modeling flexible loop regions and positioning side chains correctly. Models based on templates with <20% sequence identity may have significant errors. These inaccuracies often require further refinement using molecular dynamics simulations (Rout *et al.* 2025).

6.11. AlphaFold2 Revolution

6.11.1. Architecture Overview

DeepMind's AlphaFold2 was a revolutionary leap in predicting protein structures. It uses massive evolutionary data and machine learning to predict 3D shapes with astonishing accuracy, solving a problem that challenged scientists for decades. A high-level view of its architecture is shown in **Fig. 43** (Sibli *et al.* 2025).

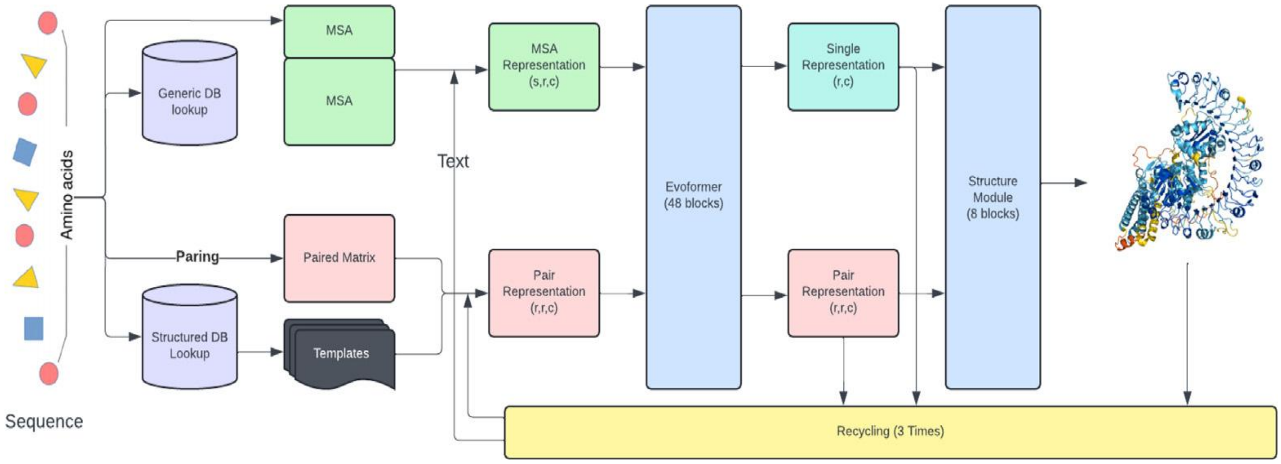

Amino acids
Sequence
Generic DB lookup
Structured DB Lookup
Paring
MSA
MSA
Paired Matrix
Templates
Text
MSA Representation (s,r,c)
Pair Representation (r,r,c)
Evoformer (48 blocks)
Single Representation (r,c)
Pair Representation (r,r,c)
Structure Module (8 blocks)
Recycling (3 Times)

Figure 43. The pipeline of AlphaFold2 architecture (Sibli *et al.* 2025).

The process starts by taking a protein sequence from a database like UniProt. This sequence is then aligned with similar ones from other species to find patterns vital for the protein's structure and function. AlphaFold2 also searches for existing structural templates in the Protein Data Bank (PDB). While a template can help, less than 0.1% of proteins have one, so AlphaFold2 is designed to work without them, relying on its powerful computational methods (Sibli *et al.* 2025).

The system is built on a neural network with two main parts :

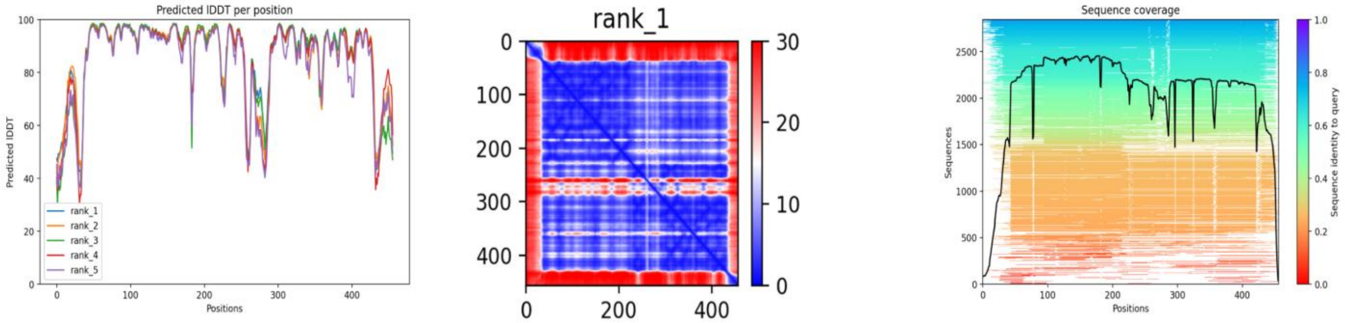
a. The Evoformer: This module refines the sequence alignment and pair representation. It uses a cross-attention mechanism to detect patterns of interaction between amino acids

b. The Structure Module: This part takes the refined data from the Evoformer and builds the 3D atomic coordinates of the protein. It uses physical simulation methods to ensure the final structure is realistic

A key feature is recycling, where the model iteratively refines its own predictions to achieve high stability and accuracy.

6.11.2. Output and Confidence Metrics

Beyond providing a 3D structure (in PDB format), AlphaFold2 generates crucial outputs that help researchers judge the model's reliability, as shown in **Fig. 44**.



(a) pLDDT (b) PAE (c) MSA Coverage

Figure 44. AlphaFold2 output visualization (Sibli *et al.* 2025).

a. pLDDT (per-residue confidence score): This score, from 0 to 100, shows the model's confidence in the position of each individual amino acid. A higher score means higher confidence

b. Predicted Aligned Error (PAE) plot: This plot shows the expected position error between any two residues

- Both axes represent residue positions in the protein sequence
- The color indicates the predicted error in angstroms (Å). Blue means low error (high confidence)**, and** red means high error (low confidence)
- The blue diagonal line represents each residue aligning with itself, where the error is zero.

c. Multiple Sequence Alignment (MSA) Coverage: This shows the evolutionary information available for different parts of the protein

- A color gradient identifies regions with high sequence conservation (blue/purple) and variable regions (red/orange)
- A black line shows the number of sequences aligned at each position. Dips in this line indicate regions with less evolutionary data (Sibli *et al.* 2025)

6.12. Molecular Docking

6.12.1. Definition

Molecular docking is a computational method used to predict how a small molecule (ligand) binds to a target protein (receptor). It models the preferred orientation, position, and 3D conformation of the ligand when interacting with the protein. This approach is widely applied in drug discovery to screen potential inhibitors, optimize lead compounds, and study molecular interactions.

6.12.2. Methodological Framework

The process involves simulating the binding between a ligand and a receptor to evaluate complementarity and stability. Key steps include:

a. Preparation of Structures: Optimizing the 3D structures of the ligand and receptor for docking simulations

b. Binding Site Identification: Defining the region on the receptor where the ligand is likely to bind

c. Docking Simulation: Using algorithms to generate and score possible binding poses based on energy and interaction compatibility

d. Pose Evaluation: Analyzing the generated poses to identify the most stable and biologically relevant binding mode

e. Interaction Analysis: Examining specific molecular forces (e.g., hydrogen bonds, hydrophobic contacts, electrostatic interactions) that stabilize the complex

6.12.3. Key Software Tools

Commonly used tools for molecular docking include:

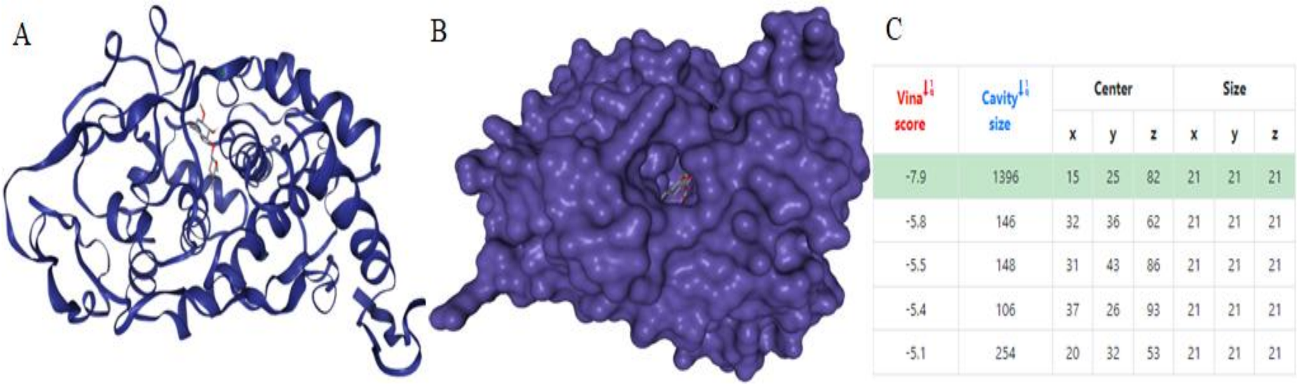
a. AutoDock/AutoDock Vina: Open-source programs offering flexible docking with a balance of accuracy and computational efficiency

b. Molecular Operating Environment (MOE): A commercial platform integrating docking, visualization, and simulation tools for drug design

c. HADDOCK: A docking software that incorporates experimental data to improve predictions of protein-ligand and protein-protein interactions

d. PLANTS: A tool designed for virtual screening and lead optimization, using algorithms inspired by ant colony optimization

An example of a molecular docking output is illustrated in **Figure 45**, showing the interaction between a guaiacyl dimer ligand and the *Bjerkandera adusta* lignin peroxidase receptor.



Vina score	Cavity size	Center			Size		
		x	y	z	x	y	z
-7.9	1396	15	25	82	21	21	21
-5.8	146	32	36	62	21	21	21
-5.5	148	31	43	86	21	21	21
-5.4	106	37	26	93	21	21	21
-5.1	254	20	32	53	21	21	21

Figure 45. Molecular docking of guaiacyl 4-O-5 guaiacyl dimer ligand with *Bjerkandera adusta* lignin peroxidase receptor. A. 3D bound configurations. B. binding cavitie. C Vina docking scores. The protein-ligand interactions were predicted using the CB-Dock server (http://clab.labshare.cn/cb-dock/php/index.php) (Bakli *et al.* 2025).

6.12.4. Applications and Advancements

Molecular docking is critical for:

a. Virtual screening of large compound libraries to identify drug candidates

b. Predicting binding affinity and specificity of ligands

c. Understanding structure-activity relationships (SAR) and mechanisms of drug action

d. Assessing potential adverse effects of compounds

Recent advances integrate artificial intelligence (AI), machine learning (ML), and deep learning (DL) to enhance the accuracy and efficiency of docking predictions. AI-driven models now improve pose prediction, enable rapid screening, and aid in discovering novel therapeutic agents.

6.12.5. Limitations

Challenges include accurately modeling flexibility in receptors and ligands, accounting for solvation effects, and predicting binding energies with high precision. Ongoing refinements aim to address these limitations through improved scoring functions and hybrid approaches combining docking with molecular dynamics simulations (Rout *et al.* 2025).

6.13. Molecular Dynamics Simulation

6.13.1. Definition

Molecular dynamics (MD) simulation is a computational technique that models the dynamic behavior of biomolecules over time. Unlike static structural methods (e.g., X-ray crystallography or cryo-EM), MD simulations capture atomic-level movements, enabling the study of conformational changes, molecular interactions, and flexibility in realistic environments.

6.13.2. Methodological Framework

MD simulations solve Newton's equations of motion for atoms within a defined system, typically following these steps:

a. System Preparation: Constructing the molecular system (e.g., protein-ligand complex) and embedding it in a solvated environment with ions

b. Energy Minimization: Reducing steric clashes and optimizing atomic positions to achieve a stable starting configuration

c. Equilibration: Gradually adjusting temperature and pressure to mimic experimental conditions (e.g., physiological settings)

d. Production Run: Simulating atomic trajectories over time to observe dynamics, interactions, and conformational changes

e. Analysis: Extracting thermodynamic, kinetic, and structural insights from the simulated trajectories

6.13.3. Key Software Tools

Widely used MD simulation packages include:

a. GROMACS: A high-performance, open-source package optimized for simulating large biomolecular systems efficiently

b. AMBER: A suite of tools featuring advanced force fields and methods for energy minimization and free-energy calculations

c. CHARMM: A versatile program with extensive parameterization options for modeling macromolecules and their interactions

d. NAMD: A parallelized tool designed for high-performance computing environments, enabling large-scale simulations

e. Advanced Techniques: Methods like metadynamics and adaptive biasing force (ABF) simulations enhance the sampling of free-energy landscapes and rare events

6.13.4. Applications and Advancements

MD simulations are indispensable for:

a. Studying protein folding, conformational changes, and allostery.

b. Investigating ligand-binding mechanisms, kinetics, and thermodynamics.

c. Bridging the gap between static structures and functional dynamics in biologics.

d. Guiding drug design by elucidating binding pathways and residence times. Recent advances include machine learning-augmented force fields, enhanced sampling algorithms, and integrations with experimental data to improve accuracy and predictive power.

6.13.5. Limitations

Challenges include high computational costs, force field inaccuracies, and difficulties in simulating large systems or long timescales. Ongoing developments focus on improving scalability, accuracy, and accessibility through hybrid methods and AI-driven approaches (Rout *et al.* 2025).

The following table summarizes essential bioinformatics tools and methods for analyzing biological sequences and structures, including both genes and proteins (**Table 13**).

Table 13. Overview of essential bioinformatics tools and software categorized by function, highlighting traditional, machine learning (ML) and Deep learning (DL) methods used in sequence analysis, structure prediction, docking, MD simulation, and functional annotation (Rout *et al.* 2025).

Category	Tool/software	Type	Description
Sequence retrieval	UniProt, NCBI GenBank, EMBL-EBI	Traditional	Databases for protein and nucleotide sequences
	iLearnPlus	ML	ML-based sequence feature encoding tool
Sequence search and alignment	BLAST, HMMER, FASTA	Traditional	Local/global alignment tools
	ClustalW, MUSCLE, T-coffee	Traditional	Multiple sequence alignment tools
	DeepBLAST	DL	Deep learning-based sequence comparison
	ML-based alignment	ML	Neural network-based alignment
Homology modeling	SWISS-MODEL, MODELLER, I-TASSER, Phyre^2	Traditional	Homology-based 3D structure prediction
	AlphaFold	ML	Deep learning-based structure prediction
Model refinement	ModRefiner, SwissSidechain	Traditional	Structure energy minimization tools
	DeepRefiner	DL	Deep learning for structure refinement
Molecular docking	AutoDock, HADDOCK, ClusPro, PLANTS	Traditional	Protein-ligand/protein-protein docking tools
	DeepDock	DL	DNN-based docking and affinity prediction
Molecular dynamics simulation	GROMACS, AMBER, NAMD, LAMMPS	Traditional	Classical MD simulation software
Structure visualization	PyMOL, Chimera, VMD	Traditional	3D visualization of molecular structures
Protein-protein interaction	STRING, RosettaDock, ClusPro, HADDOCK	Traditional	Tools for PPI prediction and modeling
	AlphaFold-Multimer	ML	AI-based protein complex prediction
Cryo-EM structure resolution	RELION, CryoSPARC	Traditional	Cryo-EM reconstruction and refinement
Functional genomics	MEGA, PhyML	Traditional	Phylogenetic and evolutionary analysis
	DeepFRI	DL	Functional residue prediction using GNN
Big data and multi-omics	Galaxy, Bioconductor	Traditional	Multi-omics data integration platforms
	DeepChem, TensorFlow, SchNet, GNNs	DL	ML frameworks for bioinformatics and cheminformatics

6.14. Virtual Screening

Virtual screening is a key computer-based method used to find new medicines. It allows scientists to quickly search massive digital libraries of molecules to identify the most promising drug candidates, saving significant time and money compared to traditional lab-only methods. This approach began in the 1990s and is now a standard tool in drug discovery.

The process, summarized in **Fig. 46**, typically relies on one of two main strategies. The first, called Structure-Based Virtual Screening (SBVS), uses the known 3D shape of a target protein to simulate how different molecules might bind to it, like finding the right key for a lock. The second, Ligand-Based Virtual Screening (LBVS), is used when a known drug exists; it finds new molecules that have a similar shape.

The overall workflow involves selecting a protein target, preparing a large and diverse digital compound library, and then using docking software to screen them. The top-ranked compounds then undergo further computer-based filtering to predict their absorption, distribution, metabolism, excretion, and Toxicity (ADME/T)—essential properties for a safe and effective drug. More advanced analysis, like Molecular Dynamics (MD) simulations, may also be used to study the stability of the drug-protein interaction. This entire *in-silico* process results in a shortlist of the best hits for experimental testing in the lab (Panwar *et al.* 2024).

Researchers use various software programs for this, from commercial packages like Schrödinger to free, open-source tools like AutoDock. While not a perfect replacement for experiments, virtual screening is a powerful first step that has successfully identified potential new drugs for cancer, infectious diseases, and other disorders (Panwar *et al.* 2024).

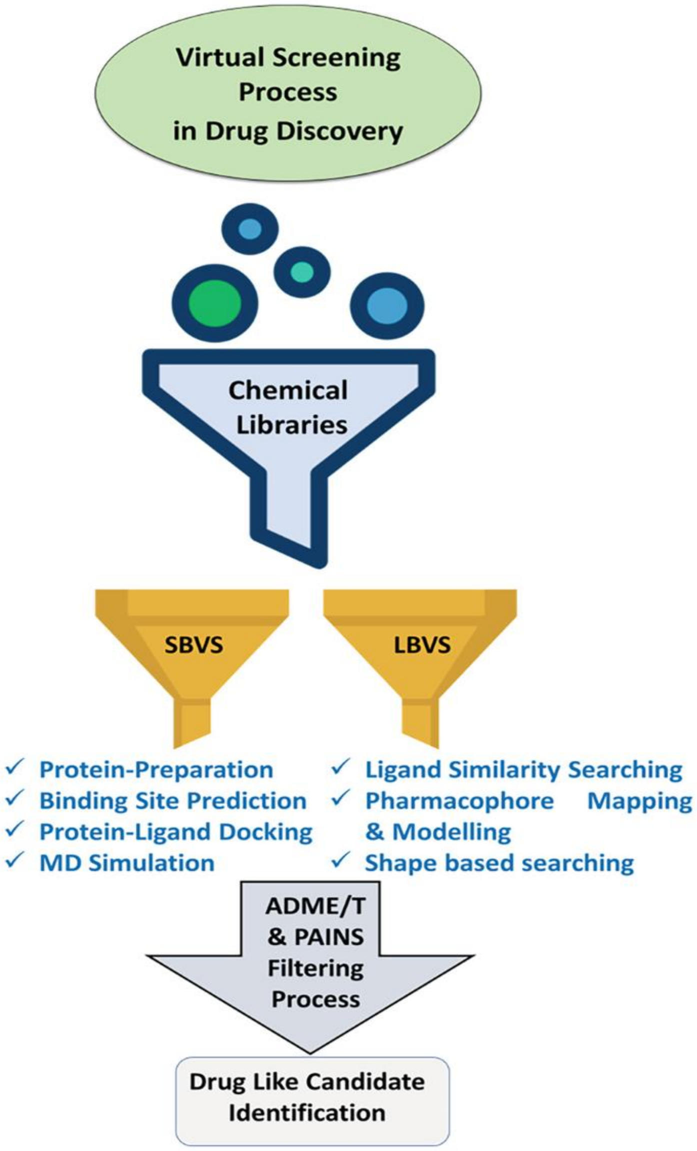

Virtual Screening Process in Drug Discovery
Chemical Libraries
SBVS
LBVS
✓ Protein-Preparation
✓ Binding Site Prediction
✓ Protein-Ligand Docking
✓ MD Simulation
✓ Ligand Similarity Searching
✓ Pharmacophore Mapping & Modelling
✓ Shape based searching
ADME/T & PAINS Filtering Process
Drug Like Candidate Identification

Figure 46. Schematic representation of virtual screening process in drug discovery. Abbreviations: SBVS; structure-based virtual screening, LBVS; ligand-based virtual screening, ADME/T; absorption, distribution, metabolism, excretion, and Toxicity, MD; molecular dynamics (Panwar *et al.* 2024).

6.15. Conclusion, Future Prospects, and Challenges

Structural bioinformatics uses computational methods to study the 3D shapes and functions of biological molecules like proteins, DNA, and RNA. This field has led to important discoveries and continues to grow rapidly. By combining computer modeling with experimental data, researchers have uncovered how these molecules work and interact, leading to advances in drug discovery, protein design, and understanding disease.

Recent advances in AI and machine learning have significantly improved protein structure prediction. These tools allow researchers to analyze complex biological data, which is vital for studying viruses and designing drugs and vaccines. By modeling viral interactions with human cells, scientists can identify potential inhibitors. AI has dramatically accelerated this design process, as shown in **Figure 47** (Rout *et al.* 2025).

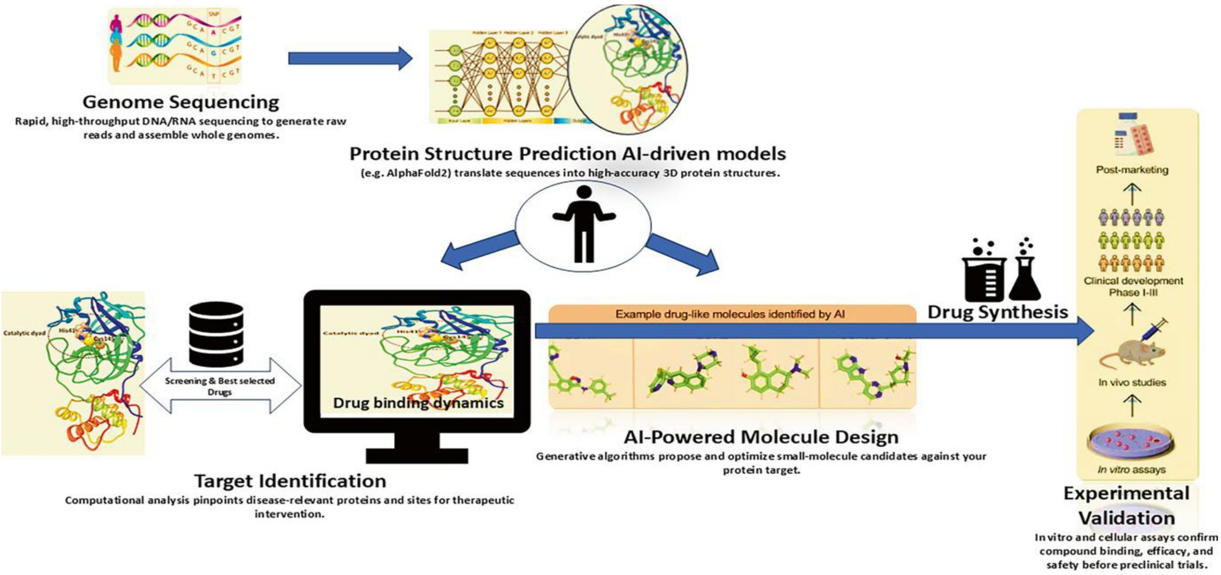

SNP
GCAACGT
GCAGCGT
GCATCGT
Genome Sequencing
Rapid, high-throughput DNA/RNA sequencing to generate raw reads and assemble whole genomes.
Protein Structure Prediction AI-driven models
(e.g. AlphaFold2) translate sequences into high-accuracy 3D protein structures.
Catalytic dyad
His41
Cys145
Screening & Best selected Drugs
Drug binding dynamics
Target Identification
Computational analysis pinpoints disease-relevant proteins and sites for therapeutic intervention.
Example drug-like molecules identified by AI
AI-Powered Molecule Design
Generative algorithms propose and optimize small-molecule candidates against your protein target.
Drug Synthesis
Post-marketing
Clinical development Phase I-III
In vivo studies
In vitro assays
Experimental Validation
In vitro and cellular assays confirm compound binding, efficacy, and safety before preclinical trials.

Figure 47. AI-driven structural analysis pipeline (Rout *et al.* 2025).

Despite these successes, structural bioinformatics still faces significant challenges. Predicting the structure of large or complex proteins remains difficult, and integrating data from multiple sources requires continued development. Molecular simulations need more accurate force fields to better represent real-world conditions, and new computational methods are needed to handle the growing volume of biological data (Wang 2024).

Looking forward, structural bioinformatics holds significant potential for growth and innovation. Integrating computational and experimental approaches will yield a fuller understanding of biological molecules and their functions in health and disease. These advancements will accelerate the development of new therapies and expand our knowledge of living systems. Despite existing challenges, the field's future is promising, with ongoing research paving the way for new discoveries and applications (Israr *et al.* 2024).

Chapter 7. Structure and Function Relationship

7.1. Fundamental Principles of Structure-Function Relationships

The structure-function relationship is a foundational concept in biology that demonstrates how the three-dimensional arrangement of molecules determines their biological roles. This principle applies to various biological macromolecules including proteins, nucleic acids, and membranes. For proteins, the specific arrangement of amino acids creates unique three-dimensional structures that define their functions. Hemoglobin's quaternary structure, for example, is essential for its ability to bind and release oxygen effectively, demonstrating a clear structure-function relationship in a critical biological process. Similarly, enzyme structure, including the shape and chemical properties of their active sites, directly influences their catalytic activity and substrate binding (Klebe 2024a).

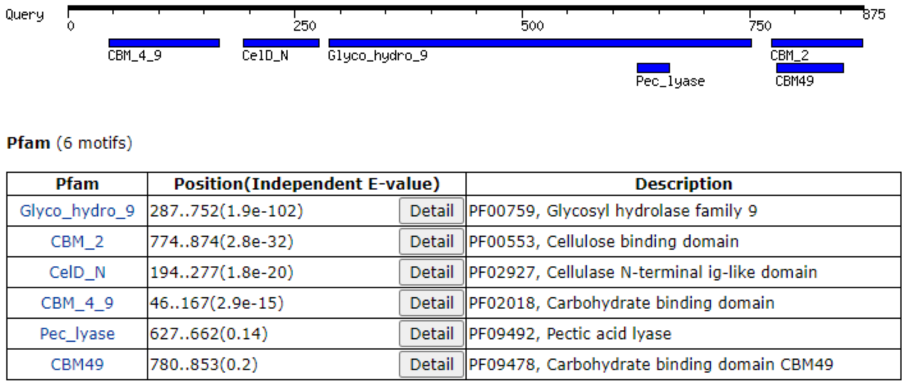
In nucleic acids, DNA's double helix structure is crucial for genetic information storage and transmission. The specific base pairing and helical configuration enable accurate replication and transcription, highlighting how structural integrity supports biological function. Cell membranes also exemplify this relationship, where the lipid bilayer arrangement regulates substance movement into and out of cells. Embedded proteins like ion channels and receptors show how structural features determine their specific roles in cellular signaling and transport (Sajid *et al.* 2024).

7.2. Computational Methodologies for Structure-Function Analysis

7.2.1. Structural Motif Identification and Domain Analysis

The relationship between protein structure and function represents a critical area in computational biology, particularly for drug discovery and protein engineering. Computational methods include motif finding, conserved domain databases (CDD), Pfam, STRING, binding site prediction, docking, and quantitative structure-activity relationship (QSAR) analyses (Ausiello *et al.* 2009).

Motif Finder identifies recurring structural motifs in proteins that are crucial for understanding protein function. Structural motifs often associate with specific ligand-binding capabilities, even across different protein folds, indicating conserved functional roles (Ausiello *et al.* 2009). As illustrated in the motif analysis of *Actinoalloteichus hoggarensis* endoglucanase (**Figure 48**), these patterns can reveal critical functional domains. Conserved Domain Databases (CDD) and Pfam provide information on conserved protein domains critical for predicting protein function. They help identify family-specific residue packing motifs for structure-based function prediction (Sayers *et al.* 2024).



Pfam (6 motifs)

Pfam	Position(Independent E-value)		Description
Glyco_hydro_9	287..752(1.9e-102)	Detail	PF00759, Glycosyl hydrolase family 9
CBM_2	774..874(2.8e-32)	Detail	PF00553, Cellulose binding domain
CelD_N	194..277(1.8e-20)	Detail	PF02927, Cellulase N-terminal ig-like domain
CBM_4_9	46..167(2.9e-15)	Detail	PF02018, Carbohydrate binding domain
Pec_lyase	627..662(0.14)	Detail	PF09492, Pectic acid lyase
CBM49	780..853(0.2)	Detail	PF09478, Carbohydrate binding domain CBM49

Figure 48. Motif analysis of *Actinoalloteichus hoggarensis* endoglucanase determined by MotifFinder (Bakli *et al.* 2023)**.**

7.2.2. Binding Site Prediction and Characterization

Accurate identification of ligand-binding sites is crucial for structure-based drug design. Methods like FTSite and GalaxySite use structural information to predict binding sites with high accuracy, essential for understanding protein-ligand interactions and guiding drug discovery (Liao *et al.* 2022).

COACH and COACH-D are consensus algorithms that combine multiple prediction methods to identify protein-ligand binding sites. COACH integrates TM-SITE and S-SITE, based on substructure comparison and sequence profile alignment respectively. COACH-D further refines predictions by incorporating molecular docking to improve ligand-binding poses, reducing steric clashes by 85% compared to COACH alone (Liao *et al.* 2022).

The predicted binding sites for *Bjerkandera adusta* lignin peroxidase demonstrate COACH's analytical capabilities (**Figure 49**) (Bakli *et al.* 2025).

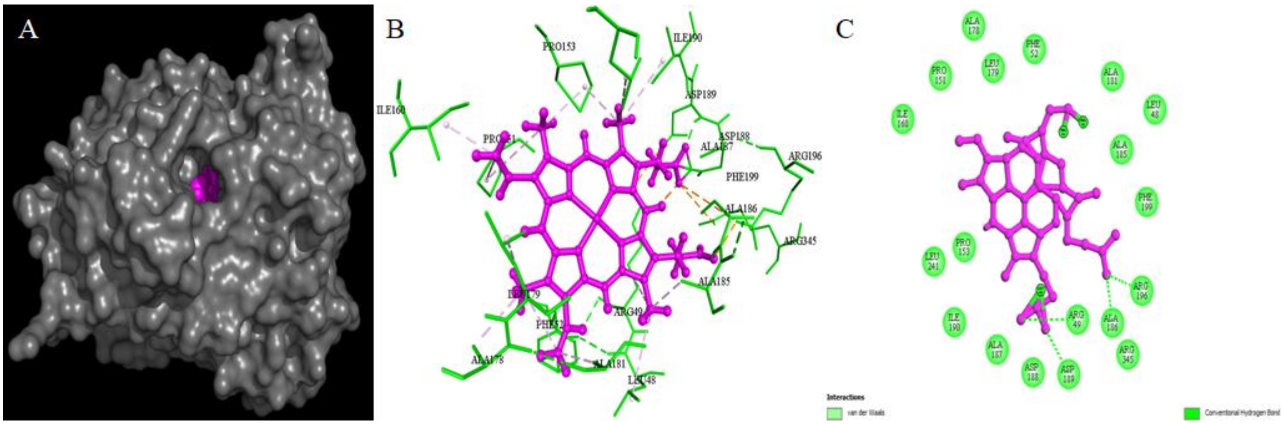


Figure 49. Predicted binding sites in complex with ligand of *Bjerkandera adusta* lignin peroxidase by COACH analysis. A. Protein surface representation (PyMOL); B. 3D ligand-receptor interactions with key residues; C. 2D interaction diagrams (BIOVIA Discovery Studio Visualizer). The ligand is shown in purple (Bakli *et al.* 2025).

COFACTOR uses structure-based approaches to predict protein functions, including ligand-binding sites, by threading target structures through template libraries. It outperforms traditional sequence-based methods and ranked highly in community-wide experiments (Liao *et al.* 2022). This approach successfully identified ligand binding sites in *Actinophytocola algeriensis* PHA synthase (**Figure 50**) (Bakli *et al.* 2024).

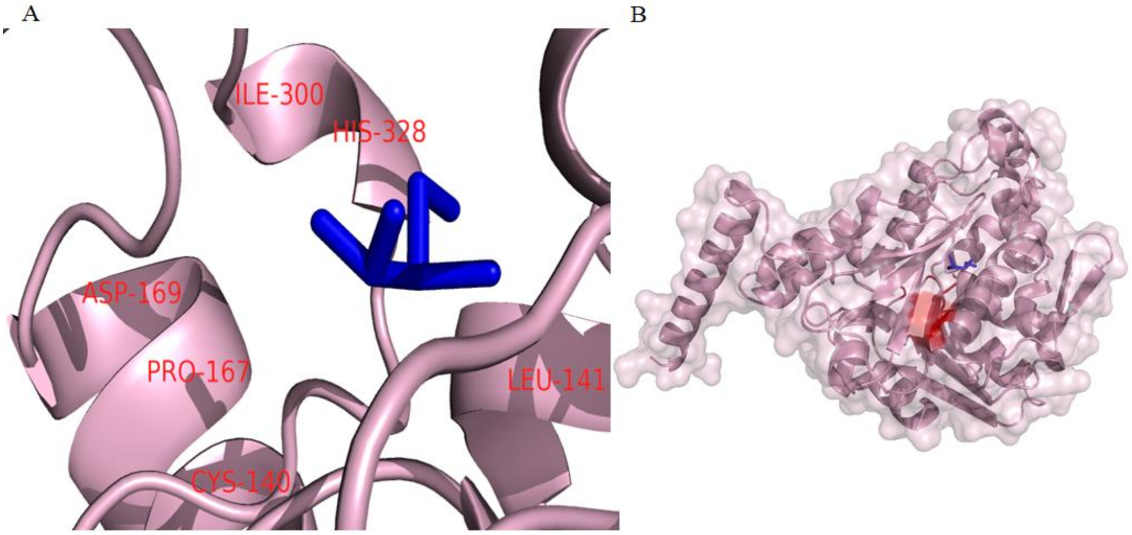


Figure 50. Ligand binding sites of the *Actinophytocola. algeriensis* PHA synthase predicted by COFACTOR. A. Ligand binding sites (Cys 140, Leu 141, Pro 167, Asp 169, Ile 300, and His 328). B. Surface view transparency of the protein with ligand. Sticks represent the ligand (blue). Ligand binding sites (red) within the protein structure (pink) (Bakli *et al.* 2024).

7.2.3. Protein Interaction Networks and Functional Context

STRING database predicts protein-protein interactions, providing insights into protein function and biological pathways. Understanding these interactions is essential for elucidating functional roles within cellular contexts.

Network-based approaches use binding site similarity networks to resolve complex structure-function relationships, providing deeper understanding of protein function and aiding accurate functional annotation (Szklarczyk *et al.* 2025).

7.3. Protein-Protein Interactions (PPIs)

7.3.1. Definition

Proteins are the primary functional molecules in biology, regulating cellular and molecular processes that define an organism’s health or disease state. Since proteins rarely act alone, their interactions with other proteins and biomolecules (e.g., DNA, RNA) are essential for their function. Studying these interactions—collectively termed protein-protein interactions (PPIs)—is critical for deciphering cellular mechanisms, signaling pathways, and identifying therapeutic targets (Rout *et al.* 2025).

7.3.2. Methodological Framework

Computational analysis of PPIs involves predicting and characterizing how proteins bind and interact. Key approaches include:

a. Data Integration: Combining experimental and computationally derived interaction evidence from diverse sources

b. Docking Simulations: Predicting the 3D structure of protein complexes using geometric and energetic compatibility

c. Interaction Network Construction: Mapping PPIs to visualize functional relationships and pathways within the cell

d. Energetic and Affinity Analysis: Evaluating binding stability, specificity, and thermodynamic properties of complexes

e. Validation: Cross-referencing predictions with experimental data to ensure biological relevance

7.3.3. Key Computational Tools

Widely used resources for PPI analysis include:

a. STRING: A comprehensive database that integrates experimental, textual, and computational data to predict PPIs and build interaction networks

As a practical example, the protein-protein interacting partners of a κ-carrageenase were identified using the STRING database (**Fig 51**), and are summarized in **Table 14** (Bakli *et al.* 2022).

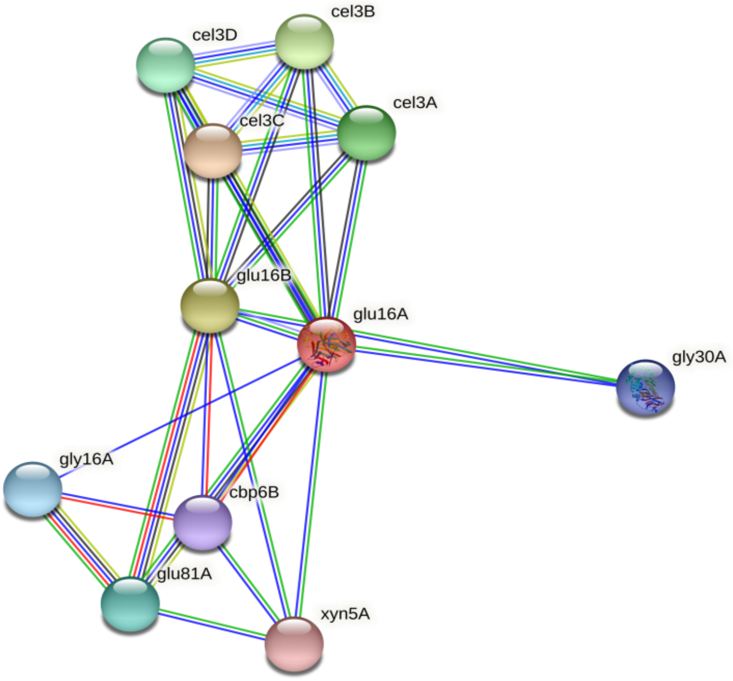


Figure 51. Protein-protein interacting partners of the κ-carrageenase using STRING database (Bakli *et al.* 2022)**.**

Table 14. Predicted functional protein partners of *Pseudomonas fluorescens* κ-carrageenase (Bakli *et al.* 2022).

Node	Annotation
cel3C	Glucan 1,4-beta-glucosidase
glu16B	Beta glucanase
cel3B	Glucan 1,4-beta-glucosidase
cel3A	Glucan 1,4-beta-glucosidase cel3A
cel3D	Putative 1,4-beta-D-glucan glucohydrolase cel3D
glu81A	Glucan endo-1,3-beta-glucanase
gly16A	Glycoside hydrolase
gly30A	Glycoside hydrolase family 30
cbp6B	Carbohydrate binding protein
xyn5A	Glucuronoarabinoxylan endo-1,4-beta-xylanase

b. ClusPro: A web-based docking tool that uses clustering algorithms to identify energetically favorable binding poses for protein pairs

c. RosettaDock: A flexible docking software that employs energy minimization and refinement to predict high-resolution protein complex structures

d. HADDOCK: A versatile docking platform that incorporates experimental constraints (e.g., NMR, mutagenesis) to improve predictions of protein-protein and protein-ligand complexes

7.3.4. Applications and Advancements

PPI analysis enables:

a. Elucidation of cellular pathways, complexes, and functional modules

b. Identification of novel drug targets by disrupting disease-relevant interactions

c. Understanding mechanisms of diseases caused by aberrant interactions (e.g., cancer, neurodegeneration)

Recent AI and deep learning models (e.g., AlphaFold-Multimer) significantly improve the accuracy and scale of PPI predictions, enabling high-throughput identification of interaction networks and cryptic binding sites.

7.3.5. Limitations

Challenges include false positives/negatives in prediction, difficulty modeling flexible interactions, and limited accuracy for transient or weak complexes. Integration of multi-omics data and better force fields are addressing these gaps (Rout *et al.* 2025).

7.4. Structure-Function Applications in Drug Discovery

7.4.1. Molecular Docking and Binding Affinity Prediction

Molecular docking predicts the preferred orientation of ligands when bound to proteins, crucial for understanding binding affinity and specificity. Docking often combines with QSAR analyses to predict potential drug candidate activity (Klebe 2024b).

7.4.2. Quantitative Structure-Activity Relationships (QSAR)

Quantitative Structure-Activity Relationships (QSAR) represent a computational approach that correlates molecular descriptors with biological activity to predict the properties of novel compounds. Since its inception in the 1960s, QSAR has evolved from simple regression models analyzing congeneric series to sophisticated AI and machine learning (ML) frameworks capable of processing massive, multi-dimensional datasets of structurally complex molecules (Pandey and Verma 2024).

In modern drug discovery, QSAR is an indispensable component, enabling high-accuracy, low-cost prediction of biological activities and ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity) properties. This allows for the efficient virtual screening of chemical libraries containing millions of candidate molecules. Classical QSAR relies on mathematical models that establish associations between biological activity and various molecular descriptors, such as 2D graphs and molecular fingerprints (Pandey and Verma 2024).

The integration of AI/ML has revolutionized QSAR by enabling the development of robust models that capture non-linear structure-function relationships. Deep neural networks and random forest algorithms are among the most powerful techniques for building these predictive models (Pandey and Verma 2024).

A critical recent advancement is the shift towards explainable AI in QSAR. By providing transparent, mathematically grounded explanations for their predictions, these models enhance interpretability, reduce bias, and foster collaboration between computational and scientific communities, thereby strengthening the entire drug discovery pipeline (Pandey and Verma 2024).

A foundational step for many structure-based approaches, including 3D-QSAR which relates 3D molecular fields to biological activity, is the computational placement of a ligand into a protein's binding site, a process known as molecular docking (Klebe 2024b).

Docking programs algorithmically fit ligand structures into a binding pocket, evaluating numerous possible orientations and conformations to predict the most stable and favorable binding mode. This process is conceptually illustrated by two key strategies for structure-based ligand design: either by incrementally "growing" a molecule from a central seed fragment or by "linking" separate, optimally placed fragments within the pocket **(Fig. 52)** (Klebe 2024b).

Case studies on targets such as the CCR5 receptor and the aryl hydrocarbon receptor (AhR) demonstrate the combined application of docking and QSAR. These studies are instrumental in predicting binding affinities, elucidating key ligand-receptor interactions, and identifying new therapeutic targets by highlighting the critical role of specific protein residues (Klebe 2024b).

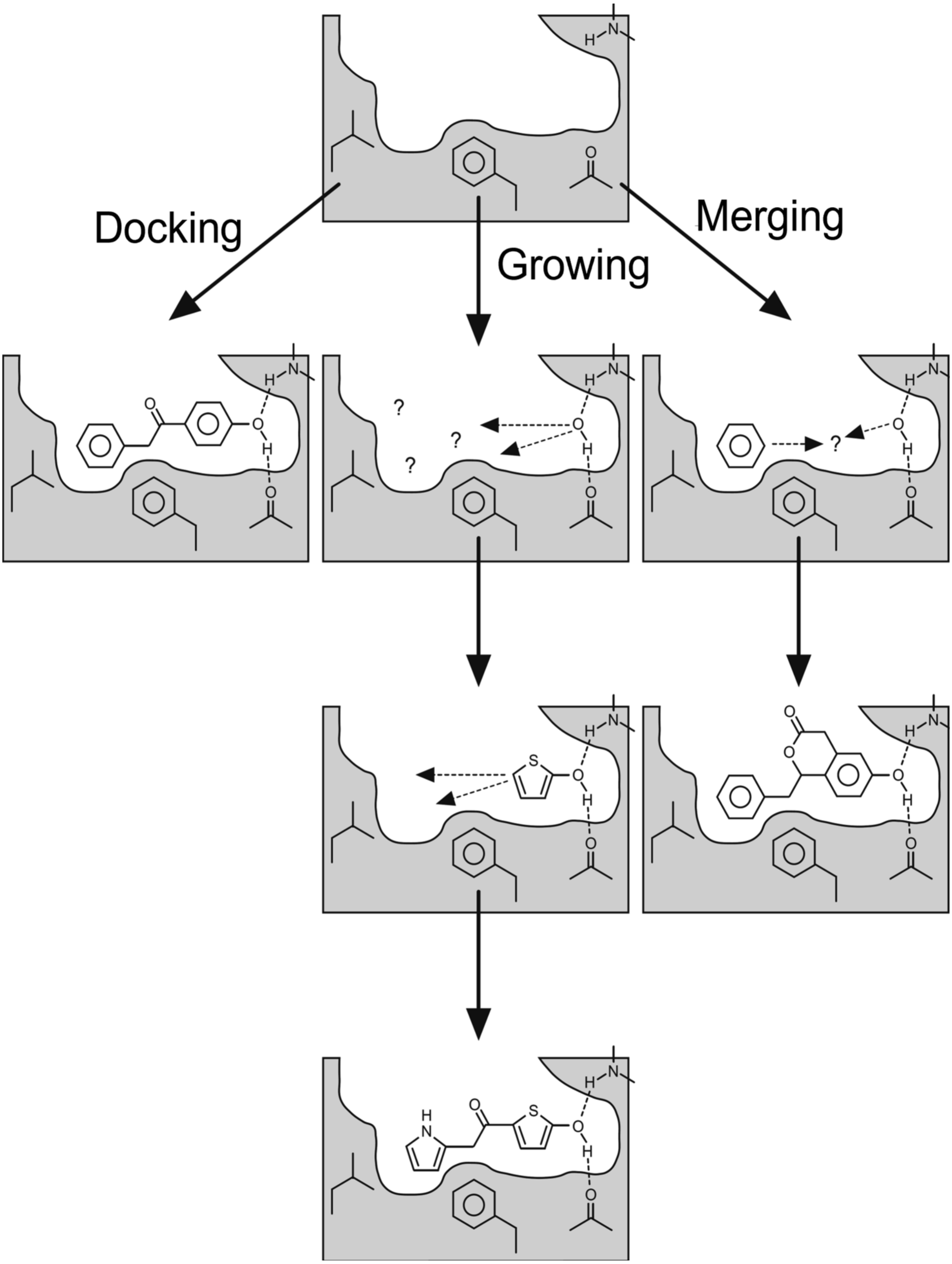


Figure 52. Strategies for structure-based ligand design. The process begins by docking complete ligand structures into a binding pocket (left). New molecules are then constructed using one of two approaches: (middle) fragment-based "growing," where functional groups are added to a central seed fragment, or (right) fragment "linking," where separate fragments placed independently within the pocket are connected (Klebe 2024b).

7.4.3. Automated Protocols and High-Throughput Applications

Tools like AADS provide automated workflows for active site identification, docking, and scoring, enhancing drug discovery efficiency by accurately predicting ligand binding sites and interactions (Singh *et al.* 2011). pocketZebra uses bioinformatics and geometry-based approaches to classify subfamily-specific binding sites, aiding in annotating proteins with unknown functions and engineering ligand-binding sites (Suplatov *et al.* 2014).

7.5. Specialized Structural Features and Functional Implications

7.5.1. Cofactor Binding Motifs

Alternating handedness motifs, characterized by alternating left- and right-handed conformations, are specific to cofactor binding sites. They facilitate tight backbone turns and direct interactions with cofactors, providing a structural basis for cofactor specificity and guiding engineered protein fold design (Cahn *et al.* 2015).

7.5.2. Cross-Fold Structural Motifs and Functional Evolution

A methodology identifying structural motifs recurring across different protein folds recognizes the same ligand fragments. This approach highlights the functional significance of shared features in interacting ligands, even among evolutionarily unrelated proteins (Ausiello *et al.* 2009). Network-based approaches integrating fold, function, and ligand similarity information help identify versatile protein folds and provide insights into protein function evolution (Sykes *et al.* 2023).

7.6. Current Challenges and Future Directions

Despite advancements, challenges remain in accurately predicting protein functions and binding sites, particularly for proteins without close homologs or those exhibiting functional divergence. Prediction complications arise from protein dynamic nature and multiple binding sites presence (Ibitoye and Soliman 2025).

Integrating diverse data sources and computational methods is necessary to improve prediction accuracy and reliability. Promising directions include developing hybrid methods combining structural, sequence, and interaction information. Understanding evolutionary transitions of protein functions through structural motifs and binding site networks can provide deeper functional insights and guide novel protein design with desired properties (Luo and Cai 2024).

References

Abdi G, Jain M, Barwant M, Tendulkar R, Tendulkar M, Tariq M, Amir A (2024) Unveiling the Dynamic Role of Bioinformatics in Automation for Efficient and Accurate Data Processing and Interpretation. In: Singh V, Kumar A (eds) Advances in Bioinformatics. Springer Nature Singapore, Singapore, pp 279-319.

Ali MA, Zahoor A, Niaz Z, Jabran M, Anas M, Shafique I, Ahmad HM, Usama M, Abbas A (2024) Bioinformatics and Computational Biology. In: Ijaz S, Ul Haq I, Mohamed Ali H (eds) Trends in Plant Biotechnology. Springer Nature Singapore, Singapore, pp 281-334.

Allmer J (2023) Noncoding RNA databases. Current Pharmaceutical Biotechnology 24 (7):825-831

Altman RB Bioinformatics in support of molecular medicine. In: Proceedings of the AMIA Symposium, 1998. p 53

Altman RB, Dugan JM (2003) Defining bioinformatics and structural bioinformatics. Structural Bioinformatics 44:1-14

Ausiello G, Gherardini PF, Gatti E, Incani O, Helmer-Citterich M (2009) Structural motifs recurring in different folds recognize the same ligand fragments. BMC Bioinformatics 10 (1):182.

Badar M (2023) A Guide to Applied Machine Learning for Biologists. Springer International Publishing

Bakli M, Bouacem K, Paşcalău R, Șmuleac L, Jaouadi B, Al-Madhagi H, Nassar H, Alessa AH, Alsaigh AA (2025) Bioinformatics analyses of lignin peroxidases from the smoky bracket fungi *Bjerkandera adusta* for endocrine disrupting chemical bioremediation. Journal of Biomolecular Structure and Dynamics:1-14

Bakli M, Bouras N, Paşcalău R, Șmuleac L (2022) Bioinformatic Characterization of a Kappa-Carrageenase from *Pseudomonas fluorescens*. Advanced Research in Life Sciences 6 (1):33-39

Bakli M, Bouras N, Paşcalău R, Șmuleac L (2023) *In silico* Structural and Functional Characterization of an Endoglucanase from *Actinoalloteichus hoggarensis*. Advanced Research in Life Sciences 7 (1):135-141

Bakli M, Paşcalău R, Șmuleac L (2024) AlphaFold Modeling and Computational Analysis of a PHA Synthase from *Actinophytocola algeriensis*. Advanced Research in Life Sciences 8 (1):39-44

Baxevanis AD, Bader GD, Wishart DS (2020) Bioinformatics. John Wiley & Sons, Bayat A (2002) Science, medicine, and the future: Bioinformatics. BMJ (Clinical research ed) 324 (7344):1018-1022

Beck T, Shorter T, Brookes AJ (2020) GWAS Central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies. Nucleic Acids Research 48 (D1):D933-D940

Berman HM, Burley SK (2025) Protein Data Bank (PDB): Fifty-three years young and having a transformative impact on science and society. Quarterly reviews of biophysics 58:e9

Blanco A, Blanco G (2022) Chapter 23 - Regulation of gene expression. In: Blanco A, Blanco G (eds) Medical Biochemistry (Second Edition). Academic Press, pp 569-581.

Bogaerts B, Van Braekel J, Van Uffelen A, D'aes J, Godfroid M, Delcourt T, Kelchtermans M, Milis K, Goeders N, De Keersmaecker SCJ, Roosens NHC, Winand R, Vanneste K (2025) Galaxy @Sciensano: a comprehensive bioinformatics portal for genomics-based microbial typing, characterization, and outbreak detection. BMC Genomics 26 (1):20

Boguski MS (1994) Bioinformatics. Current Opinion in Genetics & Development 4 (3):383-388

Brandies PA, Hogg CJ (2021) Ten simple rules for getting started with command-line bioinformatics. vol 17. Public Library of Science San Francisco, CA USA

Burley SK, Piehl DW, Vallat B, Zardecki C (2024) RCSB Protein Data Bank: supporting research and education worldwide through explorations of experimentally determined and computationally predicted atomic level 3D biostructures. IUCrJ 11 (3):279-286

Cahn JK, Brinkmann-Chen S, Spatzal T, Wiig JA, Buller AR, Einsle O, Hu Y, Ribbe MW, Arnold FH (2015) Cofactor specificity motifs and the induced fit mechanism in class I ketol-acid reductoisomerases. Biochemical Journal 468 (3):475-484

Calvino G, Farro J, Zampatti S, Peconi C, Megalizzi D, Trastulli G, Andreucci S, Cascella R, Strafella C, Caltagirone C, Grifalchi F, Giardina E (2025) From Genomics to AI: Revolutionizing Precision Medicine in Oncology. Applied Sciences 15 (12):6578

Chen Q (2024) Association Analysis Techniques and Applications in Bioinformatics. Springer Nature Singapore

Cheng Y, Ji C, Zhou H-Y, Zheng H, Wu A (2023) Web Resources for SARS-CoV-2 Genomic Database, Annotation, Analysis and Variant Tracking. Viruses 15 (5):1158

Clark AJ, Lillard Jr JW (2024) A comprehensive review of bioinformatics tools for genomic biomarker discovery driving precision oncology. Genes 15 (8):1036

Claverie JM, Notredame C (2011) Bioinformatics For Dummies. Wiley

Consortium TU (2024) UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Research 53 (D1):D609-D617

Crossa J, Montesinos-Lopez OA, Costa-Neto G, Vitale P, Martini JWR, Runcie D, Fritsche-Neto R, Montesinos-Lopez A, Pérez-Rodríguez P, Gerard G, Dreisigacker S, Crespo-Herrera L, Pierre CS, Lillemo M, Cuevas J, Bentley A, Ortiz R (2025) Machine learning algorithms translate big data into predictive breeding accuracy. Trends in Plant Science 30 (2):167-184

Danielewski M, Szalata M, Nowak JK, Walkowiak J, Słomski R, Wielgus K (2025) History of Biological Databases, Their Importance, and Existence in Modern Scientific and Policy Context. Genes 16 (1):100

Dapkūnas J, Margelevičius M (2022) The COMER web server for protein analysis by homology. Bioinformatics 39 (1)

Das R, Sharma S, Rakshit D (2024) Statistical and Biological Data Analysis Using Programming Languages. In: Genomics Data Analysis for Crop Improvement. Springer, pp 1-31

Deléage G, Gouy M, de Brevern A (2021) Bioinformatique - 3e éd.: De la séquence à la structure des protéines. Dunod

Demirbaga Ü, Aujla GS, Jindal A, Kalyon O (2024) Big Data Analytics in Bioinformatics. In: Big Data Analytics: Theory, Techniques, Platforms, and Applications. Springer, pp 265-284

DineshDarsi, R S, Krishna PJS, Sushma Pairwise Sequence Alignment in Biological Sequences using Machine Learning. In: 2023 Second International Conference on Advances in Computational Intelligence and Communication (ICACIC), 7-8 Dec. 2023 2023. pp 1-5

Edwards YJ, Cottage A (2003) Bioinformatics methods to predict protein structure and function: A practical approach. Molecular biotechnology 23 (2):139-166

Esposito S, Carputo D, Cardi T, Tripodi P (2020) Applications and Trends of Machine Learning in Genomics and Phenomics for Next-Generation Breeding. Plants 9 (1):34

Ferrari C, Guerra C (2003) Geometric methods for protein structure comparison. In: Mathematical Methods for Protein Structure Analysis and Design: CIME Summer School, Martina Franca, Italy, July 9-15, 2000. Advanced Lectures. Springer, pp 57-82

Ferreira FJN, Carneiro AS (2025) AI-Driven Drug Discovery: A Comprehensive Review. ACS Omega 10 (23):23889-23903

Gaudelet T, Day B, Jamasb AR, Soman J, Regep C, Liu G, Hayter JB, Vickers R, Roberts C, Tang J (2021) Utilizing graph machine learning within drug discovery and development. Briefings in bioinformatics 22 (6)

Goldfarb T, Kodali VK, Pujar S, Brover V, Robbertse B, Farrell CM, Oh D-H, Astashyn A, Ermolaeva O, Haddad D (2025) NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. Nucleic Acids Research 53 (D1):D243-D257

Greener JG, Kandathil SM, Moffat L, Jones DT (2022) A guide to machine learning for biologists. Nature reviews Molecular cell biology 23 (1):40-55

Gromiha MM, Ridha F, Selvaraj S (2025) Protein structural bioinformatics: an overview.

Harper R (1994) Access to DNA and protein databases on the internet. Current Opinion in Biotechnology 5 (1):4-18

Hasan ME, Osman MA, Hemeida AA, ElHefnawi M (2025) Bioinformatics and Artificial Intelligence Approaches in Metabolic Pathways. In: Metabolic Dynamics in Host-Microbe Interaction. Springer, pp 17-32

Hesper B, Hogeweg P (1970) Bioinformatica: een werkconcept. Kameleon 1 (6):28-29

Hollander M, Do T, Will T, Helms V (2021) Detecting rewiring events in protein-protein interaction networks based on transcriptomic data. Frontiers in bioinformatics 1:724297

Hu Y, Guo X, Yun Y, Lu L, Huang X, Jia S (2025) DisGeNet: a disease-centric interaction database among diseases and various associated genes. Database 2025:baae122

Ibitoye OE, Soliman ME (2025) Machine Learning in Enhancing Protein Binding Sites Predictions-What Has Changed Since Then? Combinatorial Chemistry & High Throughput Screening 28 (10):1640-1653

Iqbal N, Kumar P (2023) From Data Science to Bioscience: Emerging era of bioinformatics applications, tools and challenges. Procedia Computer Science 218:1516-1528

Ishengoma E, Rhode C (2022) Using SPAdes, AUGUSTUS, and BLAST in an automated pipeline for clustering homologous exome sequences. Current Protocols 2 (5):e449

Israr J, Alam S, Siddiqui S, Misra S, Singh I, Kumar A (2024) Advances in Structural Bioinformatics. In: Singh V, Kumar A (eds) Advances in Bioinformatics. Springer Nature Singapore, Singapore, pp 35-70

Jain K, Pandita P, Mathuria A, Mehak, Das D, Saini A, Mani I (2024) Emerging Tools for Generating Genomics Data. In: Advances in Genomics: Methods and Applications. Springer, pp 1-39

Jiang Y, Wang J, Sun A, Zhang H, Yu X, Qin W, Ying W, Li Y, Chang C, Wang X, Xie L, Liu W, Liu J, Zhang X, Yan Q, Zou Y, Zhao C, Sun H, Zhang J, Su S, Gao Q, He F (2025) The coming era of proteomics-driven precision medicine. National Science Review 12 (8).

Kaithal P, Kanchan S, Kesheri M (2024) Recent advances in biological omics databases and tools in human health. Microbial omics in environment and health:311-341

Kamble P, Nagar PR, Bhakhar KA, Garg P, Sobhia ME, Naidu S, Bharatam PV (2024) Cancer pharmacoinformatics: Databases and analytical tools. Functional & Integrative Genomics 24 (5):166

Kans J (2024) Entrez direct: E-utilities on the UNIX command line. In: Entrez programming utilities help. National Center for Biotechnology Information (US)

Karabayev D, Molkenov A, Yerulanuly K, Kabimoldayev I, Daniyarov A, Sharip A, Ashenova A, Zhumadilov Z, Kairov U (2021) re-Searcher: GUI-based bioinformatics tool for simplified genomics data mining of VCF files. PeerJ 9:e11333

Karim MR, Beyan O, Zappa A, Costa IG, Rebholz-Schuhmann D, Cochez M, Decker S (2021) Deep learning-based clustering approaches for bioinformatics. Briefings in bioinformatics 22 (1):393-415

Kellici TF, Hristozov D, Morao I (2024) AI‐Based Protein Structure Predictions and Their Implications in Drug Discovery. Computational Drug Discovery: Methods and Applications 1:227-253

Klebe G (2024a) Drug Research: Yesterday, Today, and Tomorrow. In: Drug Design: From Structure and Mode-of-Action to Rational Design Concepts. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 3-14

Klebe G (2024b) Protein Modeling and Structure-Based Drug Design. In: Drug Design: From Structure and Mode-of-Action to Rational Design Concepts. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 309-321

Koonin EV, Galperin MY (2003) Principles and Methods of Sequence Analysis. In: Sequence — Evolution — Function: Computational Approaches in Comparative Genomics. Springer US, Boston, MA, pp 111-192

Le NQK, Do DT, Hung TNK, Lam LHT, Huynh T-T, Nguyen NTK (2020) A computational framework based on ensemble deep neural networks for essential genes identification. International journal of molecular sciences 21 (23):9070

Liao J, Wang Q, Wu F, Huang Z (2022) In Silico Methods for Identification of Potential Active Sites of Therapeutic Targets. Molecules 27 (20):7103

Lipomi DJ, Ramji RS (2024) Forces Between Atoms, Ions, and Molecules. In: Introduction to Nanoengineering. Royal Society of Chemistry

Lu YY, Noble WS, Keich U (2024) A BLAST from the past: revisiting blastp's E-value. Bioinformatics 40 (12)

Luo Y, Cai J (2024) Deep learning for the prediction of protein sequence, structure, function, and interaction: applications, challenges, and future directions. Current Proteomics 21 (6):561-579

Ma X-K, Yu Y, Huang T, Zhang D, Tian C, Tang W, Luo M, Du P, Yu G, Yang L (2024) Bioinformatics software development: Principles and future directions. The Innovation Life 2 (3):100083-100081-100083-100011

Mahanandia NC, Biswal S, Nayak C, Farooqi MS, Srivastava S, Mishra DC, Chaturvedi KK, Sharma A (2025) Bioinformatics Tools: Insights from Structural Approaches. In: Rout AK, Singh RK, Shukla AK, Behera BK (eds) Advances in Omics Technologies: Exploring Genomics, Proteomics, and Metabolomics. Springer Nature Singapore, Singapore, pp 177-191

Martinez X, Chavent M, Baaden M (2020) Visualizing protein structures—tools and trends. Biochemical Society Transactions 48 (2):499-506

Martins PM, Mayrink VD, de A. Silveira S, da Silveira CH, de Lima LH, de Melo-Minardi RC How to compute protein residue contacts more accurately? In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, 2018. pp 60-67

Mughal AA (2021) Cybersecurity architecture for the cloud: protecting network in a virtual environment. International Journal of Intelligent Automation and Computing 4 (1):35-48

Nalina V, Prabhu D, Sahayarayan JJ, Vidhyavathi R (2025) Advancements in AI for Computational Biology and Bioinformatics: A Comprehensive Review. Artificial Intelligence (AI) in Cell and Genetic Engineering:87-105

Narlikar L, Ovcharenko I (2009) Identifying regulatory elements in eukaryotic genomes. Briefings in Functional Genomics & Proteomics 8 (4):215-230

Nei M (2019) Phylogenetic trees. In: Molecular evolutionary genetics. Columbia University Press, pp 287-326

Ogunjobi TT, Ohaeri PN, Akintola OT, Atanda DO, Orji FP, Adebayo JO, Abdul SO, Eji CA, Asebebe AB, Shodipe OO (2024) Bioinformatics applications in chronic diseases: A comprehensive review of genomic, transcriptomics, proteomic, metabolomics, and machine learning approaches. Medinformatics

Oyelade J, Isewon I, Ogunlana O, Aworunse O, Oyesola O, Aromolaran O, Dokumu T, Ademuwagun I, Iheagwam F, Babatunde E, Dania O, Obembe O (2020) Chapter 2 - Overview of the human genome. In: Forero DA, Patrinos GP (eds) Genome Plasticity in Health and Disease. Academic Press, pp 9-26

Paananen J, Fortino V (2020) An omics perspective on drug target discovery platforms. Briefings in bioinformatics 21 (6):1937-1953

Pandey AK, Verma S (2024) Computational Approaches for Structure-Assisted Drug Discovery and Repurposing. In: Chaudhary A, Sethi SK, Verma A (eds) Unraveling New Frontiers and Advances in Bioinformatics. Springer Nature Singapore, Singapore, pp 163-192

Panwar U, Murali A, Khan MA, Selvaraj C, Singh SK (2024) Virtual Screening Process: A Guide in Modern Drug Designing. In: Gore M, Jagtap UB (eds) Computational Drug Discovery and Design. Springer US, New York, NY, pp 21-31

Perez G, Barber GP, Benet-Pages A, Casper J, Clawson H, Diekhans M, Fischer C, Gonzalez JN, Hinrichs AS, Lee CM (2025) The UCSC genome browser database: 2025 update. Nucleic Acids Research 53 (D1):D1243-D1249

Pettini F, Visibelli A, Cicaloni V, Iovinelli D, Spiga O (2021) Multi-Omics Model Applied to Cancer Genetics. International journal of molecular sciences 22 (11):5751

Pevsner J (2015) Bioinformatics and Functional Genomics. Wiley

Pevsner J, Safari aORMC (2009) Bioinformatics and Functional Genomics, Second Edition. Wiley-Blackwell

Qiu X, Li H, Ver Steeg G, Godzik A (2024) Advances in AI for protein structure prediction: implications for cancer drug discovery and development. Biomolecules 14 (3):339

Robin X, Waterhouse AM, Bienert S, Studer G, Alexander LT, Tauriello G, Schwede T, Pereira J (2024) The SWISS‐model repository of 3D protein structures and models. Open Access Databases and Datasets for Drug Discovery:175-199

Rout AK, Singh RK, Shukla AK, Behera BK (2025) Advances in Omics Technologies: Exploring Genomics, Proteomics, and Metabolomics. Springer Nature Singapore

Saif R, Nadeem S, Khaliq A, Zia S, Iftekhar A (2023) Mathematical understanding of sequence alignment and phylogenetic algorithms: a comprehensive review of computation of different methods. Advancements in Life Sciences 9 (4):401-411

Sajid S, Rahman SU, Mahmood S, Bashir S, Habib M (2024) Fundamentals of Cellular and Molecular Biology. Bentham Science Publishers Pte. Limited

Sapundzhi F, Popstoilov M, Lazarova M RMSD Calculations for Comparing Protein Three-Dimensional Structures. In: International Conference on Numerical Methods and Applications, 2022. Springer, pp 279-288

Sayed M, Hwang JY, Park JW (2018) Sequence homology in circular RNA detection. Paper presented at the Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems, Honolulu, Hawaii

Sayers Eric W, Beck J, Bolton Evan E, Brister J R, Chan J, Connor R, Feldgarden M, Fine Anna M, Funk K, Hoffman J, Kannan S, Kelly C, Klimke W, Kim S, Lathrop S, Marchler-Bauer A, Murphy Terence D, O'Sullivan C, Schmieder E, Skripchenko Y, Stine A, Thibaud-Nissen F, Wang J, Ye J, Zellers E, Schneider Valerie A, Pruitt Kim D (2024) Database resources of the National Center for Biotechnology Information in 2025. Nucleic Acids Research 53 (D1):D20-D29

Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Sherry ST, Yankie L, Karsch-Mizrachi I (2023) GenBank 2023 update. Nucleic Acids Research 51 (D1):D141-D144

Schmidt B, Hildebrandt A (2024) From GPUs to AI and quantum: three waves of acceleration in bioinformatics. Drug Discovery Today 29 (6):103990

Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Research 18 (20):6097-6100

Schwalbe H, Audergon P, Haley N, Amaro CA, Agirre J, Baldus M, Banci L, Baumeister W, Blackledge M, Carazo JM, Carugo KD, Celie P, Felli I, Hart DJ, Hauß T, Lehtiö L, Lindorff-Larsen K, Márquez J, Matagne A, Pierattelli R, Rosato A, Sobott F, Sreeramulu S, Steyaert J, Sussman JL, Trantirek L, Weiss MS, Wilmanns M (2024) The future of integrated structural biology. Structure 32 (10):1563-1580

Schwarz RF, Tamuri AU, Kultys M, King J, Godwin J, Florescu AM, Schultz J, Goldman N (2016) ALVIS: interactive non-aggregative visualization and explorative analysis of multiple sequence alignments. Nucleic Acids Research 44 (8):e77-e77

Sibli SA, Panagiotou VP, Makris C (2025) Enhancing protein structure predictions: DeepSHAP as a tool for understanding AlphaFold2. Expert Systems with Applications 286:127853

Singh GB (2025a) BLAST. In: Fundamentals of Bioinformatics and Computational Biology: Methods and Exercises in MATLAB. Springer Nature Switzerland, Cham, pp 157-169.

Singh GB (2025b) Introduction to Bioinformatics and Computational Biology. In: Fundamentals of Bioinformatics and Computational Biology: Methods and Exercises in MATLAB. Springer Nature Switzerland, Cham, pp 3-11

Singh GB (2025c) Protein Sequence Alignment. In: Fundamentals of Bioinformatics and Computational Biology: Methods and Exercises in MATLAB. Springer Nature Switzerland, Cham, pp 141-156

Singh GB (2025d) Sequence Homology. In: Fundamentals of Bioinformatics and Computational Biology: Methods and Exercises in MATLAB. Springer Nature Switzerland, Cham, pp 113-140

Singh T, Biswas D, Jayaram B (2011) AADS-An automated active site identification, docking, and scoring protocol for protein targets based on physicochemical descriptors. Journal of chemical information and modeling 51 (10):2515-2527

Singh V, Kumar A (2024) Advances in Bioinformatics. Springer Nature Singapore

Sofi MY, Shafi A, Masoodi KZ (2022) Chapter 5 - Pairwise sequence alignment. In: Sofi MY, Shafi A, Masoodi KZ (eds) Bioinformatics for Everyone. Academic Press, pp 37-45.

Solano-Roman A, Cruz-Castillo C, Offenhuber D, Colubri A (2019) NX4: a web-based visualization of large multiple sequence alignments. Bioinformatics 35 (22):4800-4802

Suplatov D, Kirilin E, Arbatsky M, Takhaveev V, Švedas V (2014) pocketZebra: a web-server for automated selection and classification of subfamily-specific binding sites by bioinformatic analysis of diverse protein families. Nucleic Acids Research 42 (W1):W344-W349

Swargam S, Kumari I (2023) An Introduction to the Integration of Systems Biology and OMICS data for Animal Scientists. In: Systems Biology, Bioinformatics and Livestock Science. Bentham Science Publishers, pp 1-16

Sykes J, Holland BR, Charleston MA (2023) A review of visualisations of protein fold networks and their relationship with sequence and function. Biological reviews of the Cambridge Philosophical Society 98 (1):243-262

Szklarczyk D, Nastou K, Koutrouli M, Kirsch R, Mehryary F, Hachilif R, Hu D, Peluso ME, Huang Q, Fang T (2025) The STRING database in 2025: protein networks with directionality of regulation. Nucleic Acids Research 53 (D1):D730-D737

Tiwary BK (2022) Biological Databases. In: Bioinformatics and Computational Biology: A Primer for Biologists. Springer Singapore, Singapore, pp 11-31

Trivedi R, Nagarajaram HA (2020) Substitution scoring matrices for proteins ‐ An overview. Protein Science 29 (11):2150-2163

Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M (2019) Applications of machine learning in drug discovery and development. Nature reviews Drug discovery 18 (6):463-477

Villalba GC, Matte U (2021) Fantastic databases and where to find them: Web applications for researchers in a rush. Genetics and molecular biology 44 (2):e20200203

Vitorino R (2024) Transforming Clinical Research: The Power of High-Throughput Omics Integration. Proteomes 12 (3):25

Wang M, Nie Z, He Y, Vasilakos AV, Cheng Q, Ren Z (2025) Deep learning methods for protein representation and function prediction: A comprehensive overview. Engineering Applications of Artificial Intelligence 155:110977

Wang Y (2024) Advancements in Protein Structure Prediction: Novel Bioinformatics Algorithms and Applications. Theoretical and Natural Science 7 (1):32-37

Wei H, Shao J, Liu B (2024) Biological sequence analysis: Advances, medical applications, and challenges. Fundamental Research

Weltz J, Volfovsky A, Laber EB (2022) Reinforcement learning methods in public health. Clinical therapeutics 44 (1):139-154

Wolde T, Bhardwaj V, Pandey V (2025) Current Bioinformatics Tools in Precision Oncology. MedComm 6 (7):e70243

Wünschiers R (2025) Computational Biology: A Practical Introduction to Bio Data Juggling with Worked Examples. Springer Nature Switzerland

Yan J, Wang X (2022) Unsupervised and semi‐supervised learning: The next frontier in machine learning for plant systems biology. The Plant Journal 111 (6):1527-1538

Yingngam B (2024) Introduction to Bioinformatics. Artificial Intelligence and Machine Learning in Drug Design and Development:23-66

Zhang C, Wang Q, Li Y, Teng A, Hu G, Wuyun Q, Zheng W (2024) The historical evolution and significance of multiple sequence alignment in molecular structure and function prediction. Biomolecules 14 (12):1531

Zhang S, Liu K, Liu Y, Hu X, Gu X (2025) The role and application of bioinformatics techniques and tools in drug discovery. Frontiers in Pharmacology 16:1547131

Zhang Z (2024) Expanding bioinformatics: Toward a paradigm shift from data to theory. Fundamental Research

Zhou L, Feng T, Xu S, Gao F, Lam TT, Wang Q, Wu T, Huang H, Zhan L, Li L, Guan Y, Dai Z, Yu G (2022) ggmsa: a visual exploration tool for multiple sequence alignment and associated data. Briefings in bioinformatics 23 (4)

Ziemann M, Poulain P, Bora A (2023) The five pillars of computational reproducibility: bioinformatics and beyond. Briefings in bioinformatics 24 (6):bbad375

Zou D, Ma L, Yu J, Zhang Z (2015) Biological Databases for Human Research. Genomics, Proteomics & Bioinformatics 13 (1):55-63

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

Zvelebil M, Baum JO (2007) Understanding bioinformatics. Garland Science.