

VU LEARNING HUB

vulearninghub.xyz

HIGHLIGHTED HANDOUT

BIF602

Bioinformatics Computing II

BIF602 by VU Learning HUB

Prepared for VU students • Free study resources

<https://vulearninghub.xyz>

VU STUDY HANDOUT

BIF602

Bioinformatics Computing II

This handout covers the advanced computational methods of bioinformatics: sequence alignment algorithms, hidden Markov models, phylogenetic inference, genome assembly, gene prediction, protein structure analysis, and high-throughput data analysis. Each lecture builds the theory first and then shows how it is applied to real biological data.

21 lecture-wise topics | key terms | exam pointers



Table of Contents

UNIT 1: ADVANCED SEQUENCE ALIGNMENT

- Lecture 1: From Pairwise to Multiple Alignment
- Lecture 2: Substitution Matrices - PAM and BLOSUM
- Lecture 3: Heuristic Database Search - BLAST

UNIT 2: MULTIPLE ALIGNMENT AND PHYLOGENETICS

- Lecture 4: Multiple Sequence Alignment
- Lecture 5: Phylogenetic Trees - Distance Methods
- Lecture 6: Phylogenetics - Parsimony and Likelihood

UNIT 3: PATTERN MATCHING AND HIDDEN MARKOV MODELS

- Lecture 7: Motifs, Profiles and PSSMs
- Lecture 8: Hidden Markov Models
- Lecture 9: Gene Prediction

UNIT 4: GENOME ANALYSIS

- Lecture 10: Genome Assembly
- Lecture 11: Genome Annotation and Comparative Genomics
- Lecture 12: Next-Generation Sequencing Data Analysis

UNIT 5: PROTEIN STRUCTURE AND SYSTEMS BIOLOGY

- Lecture 13: Protein Structure Prediction
- Lecture 14: Protein-Protein Interactions and Networks
- Lecture 15: Systems Biology and Pathway Analysis

UNIT 6: STRUCTURAL BIOINFORMATICS AND EMERGING AREAS

- Lecture 16: Molecular Docking and Drug Design
- Lecture 17: Molecular Dynamics Simulations
- Lecture 18: Metagenomics
- Lecture 19: Single-Cell and Spatial Omics
- Lecture 20: Machine Learning in Bioinformatics
- Lecture 21: Biological Data Management and Ethics

UNIT 1: ADVANCED SEQUENCE ALIGNMENT

Lecture 1: From Pairwise to Multiple Alignment

Learning Objectives

- Distinguish global, local and semi-global alignment
- Explain why optimal alignment is computationally expensive
- Define scoring matrices and gap penalties

Key Terms & Definitions

Global alignment

An alignment that spans the entire length of both sequences, e.g. Needleman-Wunsch.

Local alignment

An alignment of the most similar subsequences, e.g. Smith-Waterman.

Substitution matrix

A table of scores for aligning each residue pair; e.g. PAM and BLOSUM matrices.

Bioinformatics Computing II begins where introductory bioinformatics ends: not with using BLAST, but with understanding the **algorithms underneath** it. Sequence alignment is the foundational operation of computational biology: it arranges two or more biological sequences so that evolutionarily or functionally equivalent positions line up.

An alignment is judged by a **scoring function**. Matches and mismatches are scored with a substitution matrix (PAM or BLOSUM for proteins), while gaps are penalized with an affine penalty: a gap-open cost plus a smaller gap-extension cost per residue. Affine gaps model biology better than linear gaps because a single insertion/deletion event usually spans several residues.

The two classic dynamic programming algorithms are **Needleman-Wunsch** (global, 1970) and **Smith-Waterman** (local, 1981). Both fill an $(m+1) \times (n+1)$ matrix in $O(mn)$ time and use traceback to recover the optimal path. Local alignment differs only in that negative scores are reset to zero and traceback starts from the maximum cell.

- **Global alignment** is appropriate for sequences of similar length that are homologous over their full length.
- **Local alignment** finds conserved domains inside otherwise unrelated sequences.
- **Semi-global (overlap) alignment** penalizes end gaps; useful for assembling overlapping reads.
- BLOSUM62 is the default for BLAST protein searches; PAM matrices suit closely related sequences.

Example

Aligning two globin sequences globally reveals the conserved heme-binding histidines; aligning a short motif locally against a genome finds all occurrences of a transcription-factor binding site.

VERY IMPORTANT — Exam Pointer

Very important for exams: state the time and space complexity of Needleman-Wunsch ($O(mn)$ both) and explain exactly how Smith-Waterman differs (zero-flooring and max-cell traceback). This is asked in almost every paper.

Section Summary

Alignment algorithms are dynamic programming methods whose behavior is controlled by substitution matrices and gap penalties; the choice of global vs local depends on the biological question.

Lecture 2: Substitution Matrices - PAM and BLOSUM

Learning Objectives

- Explain how PAM matrices are constructed
- Explain how BLOSUM matrices are constructed
- Choose the right matrix for a search

Key Terms & Definitions

PAM (Point Accepted Mutation)

A matrix family built from observed mutations in closely related proteins; PAM250 is a common distant-homology choice.

BLOSUM (BLOcks SUBstitution Matrix)

Matrices built from conserved blocks of protein families; BLOSUM62 is the standard for database searches.

Log-odds score

$\log(\text{observed frequency} / \text{expected frequency})$; positive means the substitution occurs more often than chance.

Substitution matrices quantify the **evolutionary plausibility** of each residue replacement. Margaret Dayhoff built PAM1 from 1,575 accepted mutations in 71 closely related protein families, then extrapolated higher PAM numbers by matrix multiplication - PAM250 corresponds to 250 mutations per 100 residues.

BLOSUM matrices take a different route: Steven Henikoff extracted **ungapped conserved blocks** from the BLOCKS database and counted substitutions directly, clustering sequences at a given identity threshold. BLOSUM62 clusters at 62% identity; lower numbers (BLOSUM45) suit distant relatives, higher numbers (BLOSUM80) suit close relatives.

Both families use **log-odds scores**; positive entries mean the pair is observed more often than expected by chance. The choice of matrix is the single most important parameter of a database search after the E-value threshold.

- **BLOSUM62**: default for general protein database searches.
- **PAM250 / BLOSUM45**: distant homologs (twilight zone, <30% identity).
- **BLOSUM80 / PAM30**: very close homologs, short exact matches.
- DNA alignments use simpler matrices (e.g. +5/-4 match/mismatch) because only 4 letters exist.

VERY IMPORTANT — Exam Pointer

Exam pointer: be able to explain why BLOSUM62 is preferred over PAM250 for BLAST (built from real conserved blocks rather than extrapolated mutations).

Section Summary

PAM extrapolates from close relatives; BLOSUM counts directly in conserved blocks. Matrix choice should match the evolutionary distance you expect.

Lecture 3: Heuristic Database Search - BLAST

Learning Objectives

- Describe the BLAST algorithm steps
- Interpret E-value and bit score
- Explain gapped vs ungapped BLAST

Key Terms & Definitions

E-value

The number of hits expected by chance with a score $\geq S$; $E < 0.001$ is a common significance cutoff.

Bit score

A normalized alignment score independent of the scoring system; higher is better.

HSP (High-scoring Segment Pair)

A local alignment segment without gaps that scores above the threshold.

BLAST (Basic Local Alignment Search Tool) sacrifices guaranteed optimality for **speed**: it finds most significant matches in seconds where Smith-Waterman would take hours. The algorithm works in three phases: (1) break the query into words (default 3 residues for proteins, 11 for nucleotides), (2) find database words scoring above threshold T against query words, (3) extend each seed hit in both directions to form HSPs.

Gapped BLAST (BLAST 2.0) then performs a **banded Smith-Waterman** extension around promising HSPs, dramatically improving sensitivity. PSI-BLAST iterates: it builds a position-specific scoring matrix (PSSM) from significant hits and re-searches, detecting remote homologs invisible to a single pass.

The **E-value** is the key statistic: $E = Kmn(e^{-\lambda S})$, where m and n are query and database sizes. An E-value of 0.01 means you expect 0.01 chance hits at that score - i.e., the hit is almost certainly real. Bit scores let you compare hits from searches with different scoring systems.

- **Word size (W)**: smaller W = more sensitive, slower (e.g. $W=2$ for short motifs).
- **Threshold (T)**: lower T = more seeds = more sensitive, slower.
- **PSI-BLAST**: iterate to find remote homologs via profiles.
- **BLAST variants**: blastn (DNA-DNA), blastp (protein-protein), blastx (translated DNA vs protein), tblastn (protein vs translated DNA), tblastx (translated vs translated).

Example

Searching a novel bacterial protein with blastp finds homologs; a second PSI-BLAST iteration using the PSSM then detects a distant archaeal homolog with only 22% identity.

VERY IMPORTANT — Exam Pointer

Very important: define E-value precisely and explain why it depends on database size (searching a bigger database raises E-values for the same score).

Section Summary

BLAST uses seed-and-extend heuristics to approximate Smith-Waterman at a fraction of the cost; E-values and bit scores make hits comparable and interpretable.

UNIT 2: MULTIPLE ALIGNMENT AND PHYLOGENETICS

Lecture 4: Multiple Sequence Alignment

Learning Objectives

- Explain the progressive alignment strategy
- Describe the guide-tree approach of Clustal
- List applications of MSA

Key Terms & Definitions

Progressive alignment

Build the MSA by aligning sequences pairwise in order of a guide tree (Feng-Doolittle).

Sum-of-pairs score

A common MSA objective: sum of all pairwise alignment scores in each column.

Guide tree

A rough phylogenetic tree used to decide the order of progressive alignment.

Multiple sequence alignment (MSA) aligns three or more sequences simultaneously to reveal **conserved columns** - the functionally critical residues. The exact dynamic programming solution is $O(L^N)$ for N sequences and is infeasible, so practical tools use heuristics.

The **progressive** strategy (ClustalW/Clustal Omega) first computes all pairwise alignments, builds a guide tree by neighbor-joining, then aligns profiles progressively from the most similar pair outward. Its weakness is the **greedy** nature: early mistakes are frozen into the alignment ("once a gap, always a gap").

Modern refinements include **consistency-based** methods (T-Coffee: combine multiple pairwise alignments into a library, then align consistently with it) and **iterative refinement** (MUSCLE, MAFFT: realign and improve until convergence). MAFFT adds FFT-based speed for thousands of sequences.

- **Clustal Omega**: fast, good default for protein families.
- **MUSCLE**: iterative refinement, accurate for moderate sizes.

- **MAFFT**: very fast, handles large datasets; L-INS-i mode is most accurate.
- **T-Coffee**: combines heterogeneous information (structures, profiles).
- MSA applications: phylogenetic trees, motif/domain detection, primer design, structure prediction.

VERY IMPORTANT — Exam Pointer

Exam pointer: explain the "once a gap, always a gap" problem of progressive alignment and how iterative refinement addresses it.

Section Summary

MSA is NP-hard in the exact form; progressive, consistency-based and iterative heuristics each trade speed against accuracy.

Lecture 5: Phylogenetic Trees - Distance Methods

Learning Objectives

- Define molecular phylogeny terms
- Work through UPGMA and neighbor-joining logic
- Explain bootstrapping

Key Terms & Definitions

Orthologs

Genes in different species derived from a common ancestor by speciation.

Paralogs

Genes related by duplication within a genome.

Bootstrap value

Percentage of resampled trees supporting a clade; >70% is usually considered reliable.

A phylogenetic tree represents **evolutionary relationships**; branch lengths estimate evolutionary distance. Distance methods first convert the MSA into a pairwise distance matrix (correcting for multiple substitutions with models like Jukes-Cantor or Kimura 2-parameter), then build the tree.

UPGMA (Unweighted Pair Group Method with Arithmetic Mean) clusters the closest pair iteratively, assuming a **molecular clock** (constant rate) - fast but often wrong when rates vary. **Neighbor-joining** relaxes the clock: it joins the pair that minimizes total tree length, producing an unrooted tree that tolerates rate variation.

Tree reliability is assessed by **bootstrapping**: resample alignment columns with replacement, rebuild the tree 100-1000 times, and report how often each clade appears. Bootstrap is a measure of sampling robustness, not of absolute truth.

- **UPGMA**: assumes molecular clock; produces rooted trees; sensitive to rate variation.
- **Neighbor-joining**: no clock assumption; unrooted; the standard distance method.
- **Jukes-Cantor**: simplest correction, assumes equal base frequencies and equal rates.
- **Kimura 2-parameter**: distinguishes transitions from transversions.

Example

A neighbor-joining tree of globin sequences recovers the alpha/beta duplication and the mammalian radiation; bootstrap values >90% support the major clades.

VERY IMPORTANT — Exam Pointer

Very important: compare UPGMA vs neighbor-joining assumptions, and explain what a bootstrap value of 95% actually means (and does not mean).

Section Summary

Distance methods trade model realism for speed; neighbor-joining is the workhorse, and bootstrapping quantifies support.

Lecture 6: Phylogenetics - Parsimony and Likelihood

Learning Objectives

- Explain maximum parsimony
- Explain maximum likelihood phylogeny
- Discuss long-branch attraction

Key Terms & Definitions

Maximum parsimony

Choose the tree requiring the fewest evolutionary changes.

Maximum likelihood

Choose the tree maximizing the probability of the observed data under an evolutionary model.

Long-branch attraction

Artifact where rapidly evolving lineages cluster together incorrectly.

Maximum parsimony selects the tree with the minimum number of character-state changes. It is intuitive and model-free but can be misled by **long-branch attraction**: two fast-evolving lineages independently acquire the same changes (homoplasy) and parsimony wrongly groups them.

Maximum likelihood (ML) evaluates $P(\text{data} \mid \text{tree, model})$ using an explicit substitution model (e.g. GTR+Gamma). It sums over all possible ancestral states at internal nodes via Felsenstein's pruning algorithm. ML is statistically consistent - with enough data and the right model it converges on the true tree - but is computationally heavy; heuristics (RAxML, IQ-TREE) make it practical.

Bayesian phylogenetics (MrBayes) samples trees from the posterior distribution with MCMC, reporting posterior probabilities for clades. Model selection (jModelTest/ModelFinder) matters: an under-parameterized model causes systematic error, an over-parameterized one wastes power.

- **Parsimony**: fast, no model, vulnerable to long-branch attraction.
- **Maximum likelihood**: model-based, consistent, slower; the current gold standard.

- **Bayesian:** posterior probabilities; needs convergence diagnostics (ESS, split frequencies).
- Always align carefully first - phylogeny is only as good as the MSA.

VERY IMPORTANT — Exam Pointer

Exam pointer: define long-branch attraction and name which methods resist it (likelihood/Bayesian with adequate models) and which do not (parsimony, UPGMA).

Section Summary

Character-based methods use every alignment column; likelihood and Bayesian approaches dominate modern phylogenetics because they model evolution explicitly.

UNIT 3: PATTERN MATCHING AND HIDDEN MARKOV MODELS

Lecture 7: Motifs, Profiles and PSSMs

Learning Objectives

- Define sequence motifs
- Build a position-specific scoring matrix
- Explain profile searching

Key Terms & Definitions

Motif

A short conserved sequence pattern with biological function, e.g. a binding site.

PSSM

Position-Specific Scoring Matrix: per-position scores for each residue, built from an MSA.

Profile

A generalization of a PSSM that also models insertions/deletions.

A **motif** is a recurring pattern - a transcription-factor binding site, a catalytic triad, a signal peptide. Motifs are described by consensus sequences, regular expressions (e.g. PROSITE patterns), or, most powerfully, **position-specific scoring matrices**.

A PSSM is built from an MSA of motif instances: for each column, compute the log-odds score of each residue. Scoring a new sequence means summing the per-position scores - positions with strong conservation dominate. PSI-BLAST's iterative PSSM is exactly this idea applied to whole proteins.

Profiles (Gribskov) extend PSSMs with position-specific gap penalties, and **profile HMMs** (HMMER) add a full probabilistic model of match, insert and delete states. Searching Pfam with HMMER annotates protein domains far more sensitively than BLAST.

- **Consensus:** simplest but loses quantitative information.
- **Regular expressions (PROSITE):** precise, but binary match/mismatch.

- **PSSM**: quantitative per-position scoring; used by PSI-BLAST.
- **Profile HMM (HMMER/Pfam)**: most sensitive; models indels probabilistically.

Example

The zinc-finger motif C-x(2,4)-C-x(12)-H-x(3,5)-H is found by a profile HMM across thousands of proteins, including very diverged members BLAST misses.

VERY IMPORTANT — Exam Pointer

Very important: explain how a PSSM is constructed from an MSA and why it is more sensitive than a single-sequence query.

Section Summary

Profiles turn a family alignment into a sensitive search model; profile HMMs are the state of the art for domain annotation.

Lecture 8: Hidden Markov Models

Learning Objectives

- Define the HMM components
- Explain Viterbi and forward algorithms
- List bioinformatics applications of HMMs

Key Terms & Definitions

Hidden states

The unobserved state sequence (e.g. exon/intron); emissions are the observed symbols.

Viterbi algorithm

Dynamic programming that finds the most probable hidden state path.

Forward-backward

Computes posterior probabilities of states; used in Baum-Welch training.

A **Hidden Markov Model** has hidden states (e.g. exon, intron, intergenic), each emitting observable symbols (A/C/G/T) with state-specific probabilities, plus transition probabilities between states. The states are "hidden" because we observe only the emissions.

Three classic problems: (1) **Evaluation** - probability of the observed sequence (forward algorithm); (2) **Decoding** - most likely state path (**Viterbi**, a dynamic programming traceback analogous to sequence alignment); (3) **Learning** - estimate parameters from data (**Baum-Welch**, an EM algorithm).

In bioinformatics, HMMs power **gene finders** (Genscan: states for exons, introns, splice signals), **profile HMMs** for protein families (HMMER/Pfam), and **pair-HMMs** for alignment. Training requires enough labeled data; decoding gives the annotation.

- **Forward algorithm**: total probability of observations; $O(TN^2)$.

- **Viterbi**: single best path; the decoding workhorse.
- **Baum-Welch (EM)**: unsupervised parameter training; can stick in local optima.
- Applications: gene prediction, profile search, CpG island detection, secondary structure.

VERY IMPORTANT — Exam Pointer

Exam pointer: be able to name the three HMM problems and the algorithm that solves each (forward / Viterbi / Baum-Welch).

Section Summary

HMMs model sequences as emissions from hidden biological states; Viterbi decoding turns a DNA sequence into a gene annotation.

Lecture 9: Gene Prediction

Learning Objectives

- List signals used by gene finders
- Explain ab initio vs homology-based prediction
- Discuss prediction accuracy limits

Key Terms & Definitions

Ab initio prediction

Predicting genes from sequence signals alone, e.g. Genscan, Augustus.

Splice site

GT...AG consensus at intron boundaries recognized by the spliceosome.

ORF (Open Reading Frame)

A stretch of codons without a stop; a necessary but not sufficient sign of a gene.

Ab initio gene finders scan DNA for statistical signals: start codons (ATG), stop codons, donor/acceptor **splice sites** (GT/AG), codon usage bias (hexamer frequencies differ between coding and non-coding), and promoter motifs. A generalized HMM integrates all signals into one parse of the sequence.

Homology-based methods align the genome to known proteins/ESTs: a region aligning to a known protein is almost certainly coding. Modern pipelines (Ensembl, NCBI) **combine** ab initio, homology, and RNA-seq evidence - this consensus approach is far more accurate than any single method.

Accuracy is measured at **nucleotide, exon and gene levels** (sensitivity vs specificity). Even the best pipelines miss short genes, alternative splice forms, and genes in repetitive regions; **manual curation** remains essential for reference genomes.

- **Signals**: start/stop codons, splice sites, codon bias, CpG islands, poly-A signals.
- **Genscan / Augustus**: classic ab initio HMM gene finders.
- **GenomeThreader / Exonerate**: spliced alignment of proteins to genome.
- **Consensus**: Ensembl/NCBI pipelines combine all evidence types.

VERY IMPORTANT — Exam Pointer

Very important: distinguish ab initio from homology-based prediction and explain why combining evidence improves accuracy.

Section Summary

Gene prediction fuses signal detection with statistical models; homology and transcript evidence rescue the weaknesses of pure ab initio methods.

UNIT 4: GENOME ANALYSIS

Lecture 10: Genome Assembly

Learning Objectives

- Contrast OLC and de Bruijn assembly
- Define N50 and coverage
- Explain paired-end scaffolding

Key Terms & Definitions

Contig

A contiguous assembled sequence without gaps.

Scaffold

Ordered/oriented contigs linked by paired reads, with gaps.

N50

Length of the shortest contig such that longer contigs cover 50% of the assembly.

Genome assembly reconstructs a genome from short **reads**. The classic **overlap-layout-consensus (OLC)** approach (used for Sanger/long reads) finds all overlaps, lays them out, and derives the consensus; it is accurate but $O(n^2)$ in overlaps.

De Bruijn graph assemblers (Velvet, SPAdes, for Illumina short reads) chop reads into k-mers; nodes are (k-1)-mers, edges are k-mers, and the genome is an Eulerian path. This is memory-efficient and handles huge data, but **repeats** longer than k collapse into tangled graph structures - the fundamental limit of short-read assembly.

Coverage (read depth) of 30-50x is typical for Illumina; **paired-end** reads (known insert size) order contigs into scaffolds and resolve small repeats. Assembly quality is reported with **N50** (higher is better) and validated by mapping reads back and checking gene completeness (BUSCO).

- **OLC**: long reads (PacBio/ONT); overlap graph; Canu, Flye.
- **De Bruijn**: short reads; k-mer graph; SPAdes, Velvet, MEGAHIT.
- **Repeats** are the main enemy; long reads resolve them.
- **N50, L50, BUSCO**: standard assembly quality metrics.

Example

A bacterial genome assembles into one circular contig with PacBio reads, but into 200 contigs with Illumina alone because of rRNA operon repeats.

VERY IMPORTANT — Exam Pointer

Exam pointer: define N50 precisely and explain why repeats break de Bruijn assemblies.

Section Summary

Assembly is graph traversal over reads; the read length vs repeat length race decides contiguity.

Lecture 11: Genome Annotation and Comparative Genomics

Learning Objectives

- Describe structural and functional annotation
- Explain synteny and whole-genome alignment
- Define ortholog detection by reciprocal best hit

Key Terms & Definitions

Structural annotation

Finding genes, exons, repeats on the assembled sequence.

Functional annotation

Assigning function via homology, domains (Pfam/InterPro), GO terms.

Synten

Conserved gene order between species, evidence of common ancestry.

After assembly comes **annotation**: structural (gene models, tRNAs with tRNAscan-SE, repeats with RepeatMasker) then functional. Functional annotation transfers knowledge by homology (BLAST vs UniProt), domain profiles (InterPro/Pfam), and controlled vocabularies (**Gene Ontology**: molecular function, biological process, cellular component).

Comparative genomics aligns whole genomes (MUMmer, LASTZ, Mauve) to find **syntenic blocks** - regions with conserved gene order. Breakpoints reveal inversions and translocations; conserved non-coding regions often mark enhancers.

Orthologs are found by **reciprocal best BLAST hits** (RBH): gene A in species 1 whose best hit is gene B in species 2, and vice versa. Ortholog sets (OrthoFinder/OrthoMCL) underpin phylogenomics and function transfer - with the caveat that function can diverge after duplication (paralogs).

- **RepeatMasker**: mask repeats before gene prediction.
- **InterPro**: integrates Pfam, PROSITE, PRINTS domain signatures.
- **GO annotation**: standardized function vocabulary enabling enrichment analysis.
- **RBH**: simple, widely used ortholog criterion; refined by phylogeny-aware methods.

VERY IMPORTANT — Exam Pointer

Very important: explain reciprocal best hit and why paralogs complicate function transfer.

Section Summary

Annotation turns raw sequence into biology; comparison across species reveals evolution and function.

Lecture 12: Next-Generation Sequencing Data Analysis

Learning Objectives

- Outline the NGS analysis workflow
- Explain read mapping and variant calling
- Describe RNA-seq quantification

Key Terms & Definitions

Read mapping

Aligning short reads to a reference genome (BWA, Bowtie2).

Variant calling

Detecting SNPs/indels from aligned reads (GATK).

FPKM/TPM

Normalized RNA-seq expression measures accounting for length and depth.

A standard **DNA-seq** workflow: QC (FastQC) -> trimming (Trimmomatic) -> mapping (**BWA-MEM/Bowtie2**, using Burrows-Wheeler indexing for speed) -> duplicate marking -> **variant calling** (**GATK HaplotypeCaller**: local re-assembly around candidate variants) -> annotation (SnEff/VEP).

Variant calls are filtered by quality metrics (QD, FS, MQ) and often **VQSR** (variant quality score recalibration). Population-scale projects then phase and impute genotypes. False positives concentrate in repetitive/low-complexity regions.

RNA-seq maps reads to the transcriptome (STAR, HISAT2) or pseudo-aligns (Salmon, Kallisto), counts reads per gene (featureCounts), normalizes (**TPM** preferred over FPKM for cross-sample comparison), and tests differential expression (**DESeq2/edgeR**: negative-binomial models with shrinkage).

- **BWA-MEM**: standard short-read mapper; handles split alignments.
- **GATK best practices**: the reference variant-calling pipeline.
- **TPM**: transcripts per million; sums to 1M per sample, comparable across samples.
- **DESeq2/edgeR**: differential expression with proper count statistics.

Example

RNA-seq of treated vs control cells: Salmon quantifies transcripts, DESeq2 finds 500 differentially expressed genes, GO enrichment highlights the stress-response pathway.

Section Summary

NGS analysis is a pipeline of specialized tools; statistics (not just software) decide which variants and genes are real.

UNIT 5: PROTEIN STRUCTURE AND SYSTEMS BIOLOGY

Lecture 13: Protein Structure Prediction

Learning Objectives

- Describe the levels of protein structure
- Contrast homology modeling, threading and ab initio
- Explain the CASP experiment

Key Terms & Definitions

Homology modeling

Build a 3D model from a template structure of a homologous protein.

Threading

Fit the sequence onto known folds without clear homology.

CASP

Critical Assessment of Structure Prediction: blind biennial benchmarking.

Protein structure has four levels: **primary** (sequence), **secondary** (alpha-helices, beta-sheets, predicted by e.g. PSIPRED from profiles), **tertiary** (3D fold), **quaternary** (complexes). Structure determines function, and structure is more conserved than sequence.

Homology (comparative) modeling works when a template with >30% identity exists: align target to template, copy coordinates, model loops, refine. **Threading** (fold recognition) scores the sequence against a library of folds using environment-dependent potentials - it detects remote relationships. **Ab initio** (Rosetta) folds from physics/statistics alone; historically limited to small proteins.

AlphaFold2 (CASP14, 2020) revolutionized the field: deep learning on MSAs and pairwise distances predicts structures near experimental accuracy for most single chains. The **AlphaFold Protein Structure Database** now covers 200M+ proteins, but disordered regions and complexes still need care (AlphaFold-Multimer addresses complexes).

- **>30% identity**: homology modeling reliable; **twilight zone 20-30%**: risky.
- **Threading**: for remote homologs with the same fold.
- **AlphaFold2**: MSA + attention network; transformed structural biology.
- **Validation**: Ramachandran plot, QMEAN, MolProbity scores.

VERY IMPORTANT — Exam Pointer

Very important: explain the 30% identity rule of thumb for homology modeling and what changed with AlphaFold2.

Section Summary

Structure prediction progressed from templates to deep learning; validation remains essential before biological interpretation.

Lecture 14: Protein-Protein Interactions and Networks

Learning Objectives

- List experimental PPI detection methods
- Explain network topology concepts
- Describe STRING database usage

Key Terms & Definitions

Y2H (Yeast two-hybrid)

Genetic assay detecting binary protein interactions.

Hub

A highly connected node in a biological network.

STRING

Database of known and predicted protein-protein interactions.

Protein interactions are detected experimentally by **yeast two-hybrid** (binary, high-throughput), **affinity purification-mass spectrometry** (complexes, co-complex), and computationally by **interologs** (transfer by homology), gene neighborhood/fusion, and co-expression.

Interaction **networks** are analyzed with graph theory: **degree** (number of partners), **hubs** (high-degree nodes, often essential), **clustering coefficient** (modularity), and **betweenness centrality** (bottlenecks controlling information flow). Biological networks are typically **scale-free** (few hubs, many low-degree nodes) and modular.

STRING integrates experimental, computational and text-mined interactions with confidence scores; **Cytoscape** visualizes and analyzes networks. Network propagation ("guilt by association") predicts gene function and disease genes.

- **Y2H**: binary interactions, prone to false positives; validate independently.
- **AP-MS**: captures complexes; quantitative variants (SILAC) find dynamics.
- **Hubs tend to be essential** - a key systems-biology finding.
- **Cytoscape + STRING**: standard network analysis workflow.

Example

A STRING network of DNA-repair proteins shows BRCA1 as a hub; its high betweenness explains why its mutation disrupts multiple pathways.

VERY IMPORTANT — Exam Pointer

Exam pointer: define hub and explain the observed correlation between hub status and gene essentiality.

Section Summary

Interactomics turns protein lists into networks; topology reveals function and vulnerability.

Lecture 15: Systems Biology and Pathway Analysis

Learning Objectives

- Define systems biology
- Explain GO and pathway enrichment analysis
- Interpret a KEGG pathway result

Key Terms & Definitions

Systems biology

Study of biological systems as integrated networks rather than isolated parts.

Enrichment analysis

Statistical test for over-representation of a gene set among experimental results.

KEGG

Kyoto Encyclopedia of Genes and Genomes: pathway maps and orthology.

Systems biology studies how components interact to produce emergent behavior: metabolic networks (flux balance analysis predicts growth phenotypes), signaling networks (ODE models of cascades), and gene regulatory networks (inferred from expression data).

After any omics experiment, **enrichment analysis** asks which functions are over-represented: the **hypergeometric test** (Fisher's exact) on GO terms or **KEGG pathways**, corrected for multiple testing (Benjamini-Hochberg FDR). **GSEA** avoids arbitrary cutoffs by scoring whole ranked gene lists.

Reactome and **KEGG** provide curated pathway diagrams; mapping expression or variant data onto them turns gene lists into mechanistic hypotheses. Always check the background gene set - enrichment against the whole genome vs the measured genes gives different answers.

- **Hypergeometric/Fisher test**: classic over-representation analysis.
- **FDR correction**: essential - thousands of terms are tested.
- **GSEA**: uses full ranked list, no cutoff needed.
- **FBA**: constraint-based metabolic modeling; predicts knockouts.

VERY IMPORTANT — Exam Pointer

Very important: explain multiple-testing correction in enrichment analysis and why the background set choice matters.

Section Summary

Systems biology integrates omics data with networks and pathways; enrichment statistics connect gene lists to biology.

UNIT 6: STRUCTURAL BIOINFORMATICS AND EMERGING AREAS

Lecture 16: Molecular Docking and Drug Design

Learning Objectives

- Explain the docking problem
- Distinguish scoring functions
- Describe virtual screening

Key Terms & Definitions

Docking

Predicting the bound pose and affinity of a small molecule in a protein binding site.

Virtual screening

Docking large compound libraries to find drug candidates computationally.

Pharmacophore

The 3D arrangement of features needed for biological activity.

Molecular docking predicts how a ligand binds a protein: the search algorithm samples ligand conformations and orientations, while the **scoring function** estimates binding free energy. Docking is used to propose binding modes for known actives and to screen libraries for new ones.

Scoring functions come in three flavors: **force-field based** (AMBER/CHARMM terms), **empirical** (weighted interaction terms fitted to data), and **knowledge-based** (statistical potentials from PDB structures). All are approximate - docking reliably finds poses but **ranks affinities poorly**, so top hits need experimental validation.

Virtual screening docks millions of compounds (e.g. ZINC database) against a target; **pharmacophore screening** is faster, matching 3D feature patterns instead. Structure-based drug design iterates: dock, synthesize, assay, co-crystallize, refine.

- **AutoDock Vina**: widely used open-source docking program.
- **Search**: genetic algorithms, Monte Carlo sample the pose space.
- **Scoring is the weak link**: expect enrichment, not exact K_d prediction.
- **ADMET filtering**: drug-likeness (Lipinski rules) before docking saves effort.

VERY IMPORTANT — Exam Pointer

Exam pointer: explain why docking is better at pose prediction than affinity ranking.

Section Summary

Docking accelerates drug discovery by prioritizing compounds, but computation proposes and experiment disposes.

Lecture 17: Molecular Dynamics Simulations

Learning Objectives

- Describe the MD simulation setup
- Explain force fields
- List what MD can and cannot do

Key Terms & Definitions

Force field

A parameterized energy function describing bonded and non-bonded interactions.

RMSD

Root-mean-square deviation: measures structural drift from a reference.

Enhanced sampling

Techniques (replica exchange, metadynamics) that accelerate rare-event sampling.

Molecular dynamics integrates Newton's equations for every atom: forces come from a **force field** (AMBER, CHARMM, GROMOS) with bonded terms (bonds, angles, dihedrals) and non-bonded terms (Lennard-Jones, Coulomb). A typical setup solvates the protein in explicit water, adds ions, minimizes, equilibrates, then runs production.

The integration **timestep** (~2 fs with constrained bonds) limits accessible timescales: routine simulations reach microseconds on GPUs, while folding can take milliseconds. **Enhanced sampling** methods (replica exchange, metadynamics, umbrella sampling) compute free energies along chosen coordinates.

MD reveals **conformational dynamics** invisible to static structures: allosteric transitions, induced fit, cryptic pockets. Analysis uses **RMSD/RMSF** (stability/flexibility), principal component analysis of motions, and Markov state models to extract kinetics.

- **Explicit solvent + PME electrostatics**: the standard for biomolecules.
- **Timescale gap**: simulation ns-us vs biological ms-s; enhanced sampling bridges it.
- **Validation**: compare with NMR order parameters, crystallographic B-factors.
- **Coarse-graining** (MARTINI): trade atomic detail for longer timescales.

VERY IMPORTANT — Exam Pointer

Very important: state the typical MD timestep and explain what limits the accessible simulation timescale.

Section Summary

MD turns static structures into movies; force-field accuracy and sampling limits bound what the movies can tell us.

Lecture 18: Metagenomics

Learning Objectives

- Contrast amplicon vs shotgun metagenomics
- Explain taxonomic binning
- Describe functional profiling

Key Terms & Definitions

16S rRNA sequencing

Amplicon sequencing of a marker gene for bacterial community profiling.

Binning

Grouping assembled contigs into draft genomes (MAGs).

MAG

Metagenome-Assembled Genome: a draft genome binned from environmental reads.

Amplicon (16S/ITS) sequencing profiles "who is there" cheaply: PCR amplifies a marker gene, reads are clustered into **OTUs/ASVs** (DADA2 denoises to exact sequence variants), and taxonomy is assigned against SILVA/Greengenes. It is biased by primers and gives no functional information.

Shotgun metagenomics sequences all DNA: assembly (MEGAHIT, metaSPAdes) then **binning** by coverage and tetranucleotide composition (MetaBAT2, CONCOCT) yields **MAGs** - draft genomes of uncultured microbes. Quality is assessed by single-copy marker completeness/contamination (CheckM).

Functional profiling (HUMANn) maps reads to gene families and pathways, revealing what the community can do. Metagenomics has uncovered the human gut microbiome's role in health and vast viral diversity in oceans.

- **DADA2**: denoising to amplicon sequence variants (ASVs), replacing OTU clustering.
- **Binning signals**: differential coverage across samples + k-mer composition.
- **CheckM**: completeness/contamination estimates for MAGs.
- **Always include controls**: contamination ("kitome") plagues low-biomass samples.

VERY IMPORTANT — Exam Pointer

Exam pointer: compare what amplicon vs shotgun metagenomics can and cannot tell you.

Section Summary

Metagenomics studies microbes without culturing; assembly plus binning recovers genomes from mixtures.

Lecture 19: Single-Cell and Spatial Omics

Learning Objectives

- Explain scRNA-seq workflow
- Describe clustering and marker detection
- Introduce spatial transcriptomics

Key Terms & Definitions

scRNA-seq

RNA sequencing of individual cells, revealing cell-type heterogeneity.

UMI

Unique Molecular Identifier: tags molecules to remove PCR duplicates.

Spatial transcriptomics

Gene expression mapped to tissue coordinates.

scRNA-seq (10x Chromium, Smart-seq) barcodes each cell's mRNA: droplets encapsulate cells with barcoded beads, **UMIs** tag individual molecules so PCR duplicates can be collapsed. After sequencing, the count matrix (cells x genes) is normalized, dimensionality-reduced (PCA), and **clustered** (Louvain/Leiden) to find cell types.

Analysis challenges: **dropouts** (zero inflation), batch effects (corrected by Harmony/integration), and doublet cells. **Marker genes** per cluster are found by differential expression; annotation uses reference atlases (Azimuth). **Pseudotime** (Monocle) orders cells along differentiation trajectories.

Spatial transcriptomics (Visium, MERFISH) adds tissue coordinates, linking expression to anatomy. **CITE-seq** measures surface proteins alongside RNA. Together these resolve development, tumor heterogeneity, and brain cell atlases at unprecedented resolution.

- **UMIs**: essential for accurate molecule counting. 
- **QC filters**: remove high-mitochondrial (dying) cells and doublets.
- **Integration**: align batches/conditions before comparing.
- **CellTypist/Azimuth**: automated annotation against references.

VERY IMPORTANT — Exam Pointer

Very important: explain what UMIs solve and why mitochondrial-percentage filtering is standard QC.

Section Summary

Single-cell methods decompose tissues into cell types; spatial methods put them back on the map.

Lecture 20: Machine Learning in Bioinformatics

Learning Objectives

- List ML applications in bioinformatics
- Explain cross-validation for biological data
- Discuss interpretability vs accuracy

Key Terms & Definitions

Cross-validation

Splitting data into folds to estimate generalization performance.

Data leakage

Test information leaking into training, inflating performance estimates.

Feature importance

Which inputs drive a model's predictions (e.g. SHAP values).

Machine learning pervades bioinformatics: **variant effect prediction** (CADD, AlphaMissense), **protein function** (DeepGO), **gene expression** prediction from sequence (Enformer), and **drug-target interaction** models. Deep learning excels where large labeled datasets exist.

Biological data demands careful **validation**: homology between train and test sets causes **data leakage** - split by sequence identity (e.g. <30%) or by protein family, not randomly. Report **precision-recall** for imbalanced problems (rare variants, binding sites) rather than accuracy alone.

Interpretability matters for biology: **feature attribution** (SHAP, attention maps) suggests mechanisms, e.g. which motif positions drive a splicing model. But correlation is not causation - model highlights are hypotheses requiring experiments.

- **Split by homology**, not randomly, to avoid leakage.
- **AUPRC** over accuracy for imbalanced biological problems.
- **External test sets** (new proteins, new species) are the real exam. ☰
- **Interpretability tools** generate hypotheses, not conclusions.

Example

A deep model predicts splice sites from sequence; splitting chromosomes (not random reads) between train/test prevents leakage from overlapping windows.

VERY IMPORTANT — Exam Pointer

Exam pointer: explain data leakage via homology and the correct splitting strategy.

Section Summary

ML is powerful but fragile on biological data; rigorous splitting and honest metrics separate real advances from artifacts.

Lecture 21: Biological Data Management and Ethics

Learning Objectives

- Name major biological databases
- Explain FAIR data principles
- Discuss ethics of genomic data

Key Terms & Definitions

FAIR

Findable, Accessible, Interoperable, Reusable - data stewardship principles.

INSDC

International Nucleotide Sequence Database Collaboration (GenBank/ENA/DDBJ).

dbGaP

Database of Genotypes and Phenotypes: controlled-access human data.

Primary databases archive raw data: **INSDC** (GenBank/ENA/DDBJ) for sequences, **SRA** for reads, **PDB** for structures, **GEO/ArrayExpress** for expression. Secondary databases curate: **UniProt** (protein function), **Ensembl** (genomes), **KEGG/Reactome** (pathways). Use accession numbers - they are the permanent identifiers of biology.

FAIR principles demand data be findable (DOIs, metadata), accessible, interoperable (standard formats: FASTQ, BAM, VCF), and reusable (clear licenses). Reproducibility needs versioned code, containers, and workflow managers (Snakemake/Nextflow).

Ethics: human genomic data is identifiable - access is controlled (**dbGaP**, EGA) with consent and data-use agreements. Indigenous data sovereignty (CARE principles), dual-use concerns (pathogen engineering), and equitable benefit-sharing are active debates every bioinformatician should know.

- **Accession numbers** (e.g. NM_000059) uniquely identify records - cite them.
- **FAIR + containers**: the reproducibility stack.
- **Controlled access** for identifiable human genomic data.
- **CARE principles**: Collective benefit, Authority, Responsibility, Ethics for Indigenous data.

VERY IMPORTANT — Exam Pointer

Very important: define FAIR and explain why human genomic data needs controlled access.

Section Summary

Databases, standards and ethics are the infrastructure that makes computational biology trustworthy and reusable.

Glossary of Key Terms

E-value — Expected number of chance hits at a given score; lower is more significant.

Ortholog — Genes separated by speciation; usually keep the same function.

PSSM — Position-specific scoring matrix built from a motif alignment.

Viterbi — Dynamic programming algorithm for the most probable HMM state path.

N50 — Assembly contiguity metric: shortest contig covering 50% of assembly length.

Bootstrap — Resampling procedure quantifying clade support in phylogenies.

Homology modeling — 3D structure prediction from a homologous template.

Enrichment analysis — Testing gene lists for over-represented functions/pathways.