



Machine Learning for Genomics

Introduction to Genomics and ML - Central Dogma Review

Sarwan Ali

Department of Computer Science
Georgia State University

 Understanding the Central Dogma 

Today's Learning Journey

- 1 Course Introduction and Motivation
- 2 Fundamentals of Molecular Biology
- 3 The Central Dogma of Molecular Biology
- 4 Transcription: DNA to RNA
- 5 Translation: RNA to Protein
- 6 From Genotype to Phenotype
- 7 Data Types and Computational Challenges
- 8 Applications and Future Directions
- 9 Summary and Next Steps

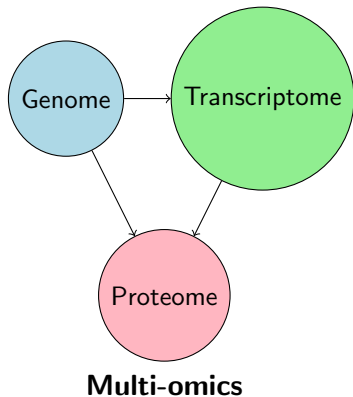
Why Machine Learning + Genomics?

The Data Explosion in Genomics:

- Human genome: 3.2 billion base pairs
- Single cell RNA-seq: 20,000+ genes per cell
- Population studies: millions of genomes
- Multi-omics integration challenges

Traditional approaches are insufficient for:

- Pattern recognition in high-dimensional data
- Predictive modeling of biological processes
- Integration of heterogeneous data types



Learning Objectives

By the end of this lecture, you will be able to:

- 1 **Recall** the central dogma of molecular biology and its key components
- 2 **Explain** the molecular processes of transcription and translation
- 3 **Identify** the information flow from genotype to phenotype
- 4 **Analyze** how variations in DNA sequence affect downstream processes
- 5 **Connect** molecular biology concepts to computational approaches
- 6 **Evaluate** the complexity and data types generated at each step

Foundation Knowledge

This review establishes the biological foundation essential for understanding how ML algorithms can be applied to genomic data analysis.

What is Genomics?

Definition

Genomics is the comprehensive study of an organism's complete set of genetic material, including genes and non-coding sequences.

Key Components:

- **Structural genomics:** Physical nature of genomes
- **Functional genomics:** Gene and genome function
- **Comparative genomics:** Genome structure/function across species
- **Pharmacogenomics:** Drug response variation

Scale and Complexity:

- Human genome: $\sim 3.2 \times 10^9$ bp
- $\sim 20,000$ - $25,000$ protein-coding genes
- $\sim 98\%$ non-coding regions
- Individual variation: $\sim 0.1\%$ between humans

Computational Challenge

How do we extract meaningful patterns and predictions from this massive, complex dataset?

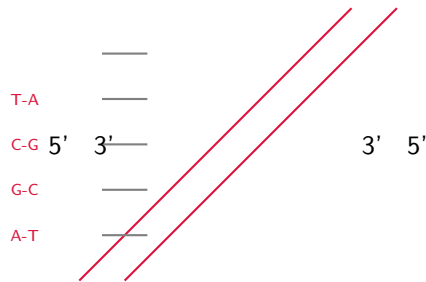
DNA Structure and Organization

DNA Double Helix Structure:

- Antiparallel strands (5' to 3' direction)
- Complementary base pairing: A-T, G-C
- Major and minor grooves
- Right-handed helix (B-form)

Hierarchical Organization:

- Nucleotides → DNA → Chromatin → Chromosomes
- Euchromatin vs. Heterochromatin
- Regulatory elements: Promoters, enhancers, silencers



ML Relevance

DNA sequence patterns, structural motifs, and regulatory elements are key features for machine learning algorithms in genomics.

Central Dogma: Historical Perspective

Francis Crick (1958, 1970)

"The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information."

Original Formulation:



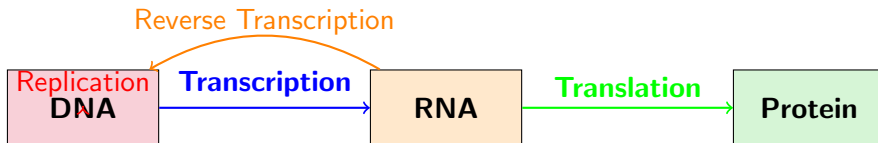
Key Principles:

- Unidirectional information flow
- Sequence-specific transfer
- Universal genetic code
- Information preservation

Modern Understanding

The central dogma has been refined to include reverse transcription, RNA editing, epigenetic modifications, and non-coding RNA functions.

Extended Central Dogma



- Pre-mRNA processing
 - Alternative splicing
 - RNA editing
 - microRNA regulation

- Post-translational modifications
 - Protein folding
 - Protein complexes

Implications for ML

Each step introduces complexity and variability that can be modeled computationally. Understanding these processes is crucial for feature engineering and model interpretation.

Transcription Overview

Definition

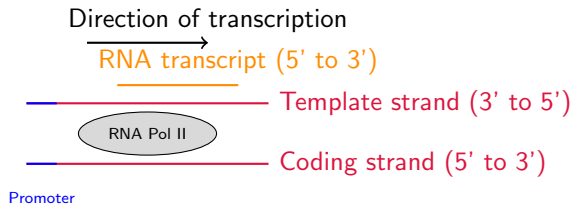
Transcription is the process by which genetic information in DNA is copied into RNA by RNA polymerase.

Key Players:

- **RNA Polymerase II:** Primary enzyme for mRNA
- **Transcription factors:** Sequence-specific DNA binding
- **Promoters:** Transcription initiation sites
- **Enhancers/Silencers:** Regulatory elements

Three Main Phases:

- (1) **Initiation**, (2) **Elongation**, (3) **Termination**



Computational Perspective

Transcription creates the first layer of information processing where DNA sequence features determine RNA expression levels.

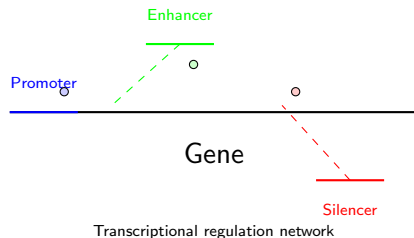
Transcriptional Regulation

Promoter Elements:

- **Core promoter:** TATA box, Initiator (Inr), DPE
- **Proximal promoter:** -200 to +200 bp from TSS
- **Distal promoter:** Beyond proximal region

Regulatory Mechanisms:

- **Positive regulation:** Activators, enhancers
- **Negative regulation:** Repressors, silencers
- **Chromatin structure:** Histone modifications
- **DNA methylation:** Epigenetic silencing



ML Applications

- Promoter prediction from sequence features
- Transcription factor binding site identification
- Gene expression prediction from regulatory elements
- Chromatin accessibility modeling

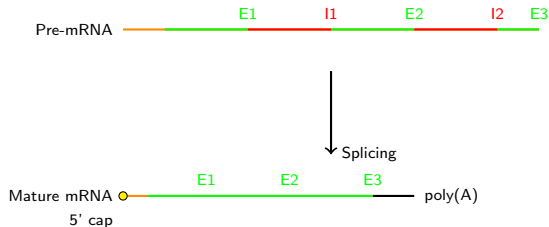
RNA Processing and Maturation

Pre-mRNA Processing Steps:

- 1 **5' Capping:** 7-methylguanosine cap
- 2 **3' Polyadenylation:** Poly(A) tail addition
- 3 **Splicing:** Intron removal, exon joining
- 4 **RNA editing:** Base modifications (A-to-I, C-to-U)

Alternative Splicing:

- Exon skipping
- Intron retention
- Alternative 5'/3' splice sites
- Mutually exclusive exons



Computational Complexity

Alternative splicing increases proteomic diversity: one gene can produce multiple protein isoforms. This creates a many-to-many mapping that ML algorithms must account for.

Translation Overview

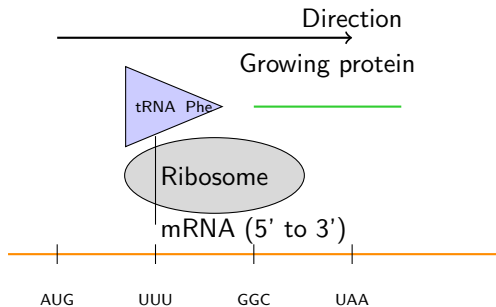
Definition

Translation is the process by which the sequence of nucleotides in mRNA is decoded to produce a specific sequence of amino acids in a protein.

Key Components:

- **Ribosomes:** Protein synthesis machinery
- **tRNA:** Amino acid carriers
- **Aminoacyl-tRNA synthetases:** tRNA charging
- **Translation factors:** Process regulation

Three Phases: (1) **Initiation:** Ribosome assembly,
(2) **Elongation:** Peptide chain growth,
(3) **Termination:** Release of protein



Information Transfer

Translation converts the 4-letter nucleotide alphabet into the 20-letter amino acid alphabet using the genetic code.

The Genetic Code

Key Properties:

- **Triplet code:** 3 nucleotides = 1 codon
- **Degenerate:** Multiple codons per amino acid
- **Universal:** Nearly identical across species
- **Non-overlapping:** Codons read sequentially
- **Comma-free:** No punctuation between codons

Special Codons:

- **Start codon:** AUG (Methionine)
- **Stop codons:** UAA, UAG, UGA
- **Wobble position:** 3rd position tolerance

Genetic Code Table

UUU → Phe	UUC → Phe
UUA → Leu	UUG → Leu
UCU → Ser	UCC → Ser
UAU → Tyr	UAC → Tyr
UGU → Cys	UGC → Cys

AUG → Met (Start)

UAA, UAG, UGA → Stop

64 codons → 20 amino acids

Computational Significance

The genetic code redundancy provides error tolerance and is crucial for understanding synonymous vs. non-synonymous mutations in evolutionary and disease studies.

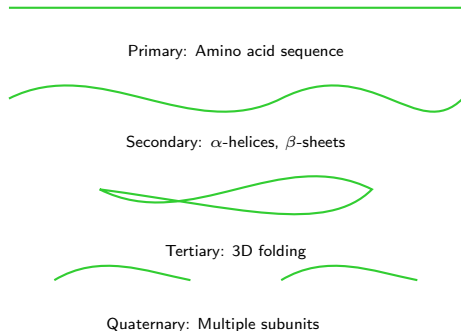
Protein Structure and Function

Protein Structure Hierarchy:

- 1 **Primary:** Amino acid sequence
- 2 **Secondary:** Local folding (α -helices, β -sheets)
- 3 **Tertiary:** 3D structure of single chain
- 4 **Quaternary:** Multiple subunit assembly

Structure-Function Relationship:

- Active sites and binding domains
- Allosteric regulation
- Post-translational modifications
- Protein-protein interactions



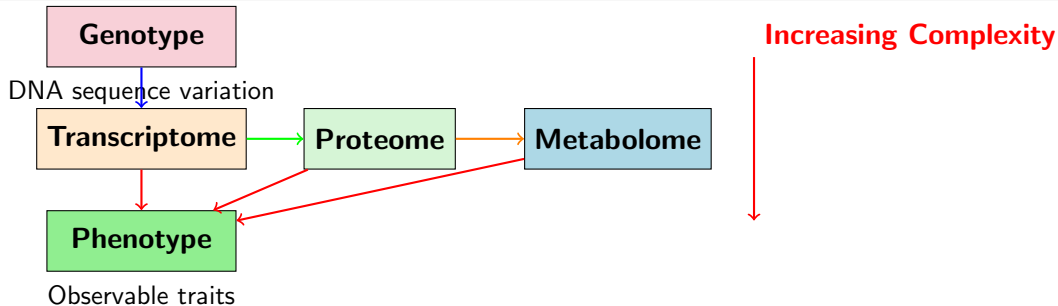
ML Applications in Protein Biology

Protein structure prediction, function annotation, interaction networks, and drug target identification all rely on computational approaches.

The Genotype-Phenotype Map

Definition

The **genotype-phenotype map** describes how genetic variation (genotype) influences observable characteristics (phenotype) through molecular and developmental processes.



Computational Challenge

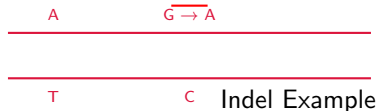
Understanding this multi-level, non-linear mapping is a central goal of computational genomics and systems biology.

Types of Genetic Variation

Single Nucleotide Variants (SNVs):

- Point mutations: $A \rightarrow G$, $C \rightarrow T$, etc.
- Most common type of variation
- ~4-5 million SNVs per individual genome
- Synonymous vs. non-synonymous effects

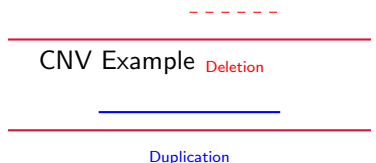
SNV Example



Structural Variants:

- **Insertions/Deletions (Indels):** 1-50 bp
- **Copy Number Variants (CNVs):** >50 bp
- **Chromosomal rearrangements:** Inversions, translocations
- **Repeat expansions:** Microsatellites, transposons

CNV Example



ML Feature Engineering

Different types of genetic variants require different computational representations and modeling approaches in machine learning pipelines.

Functional Consequences of Variation

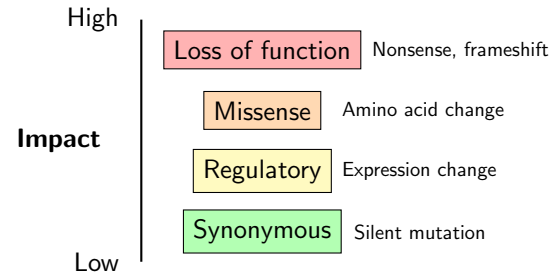
Protein-Coding Effects:

- **Synonymous:** No amino acid change
- **Missense:** Amino acid substitution
- **Nonsense:** Premature stop codon
- **Frameshift:** Reading frame alteration

Regulatory Effects:

- Promoter region variants
- Enhancer/silencer disruption
- Splice site mutations
- miRNA binding site changes

Consequence Prediction:



Computational Prediction

ML algorithms can predict the functional impact of genetic variants using sequence features, evolutionary conservation, and structural information.

Genomic Data Types

Sequence Data:

- **Whole Genome Sequencing (WGS):** Complete genome
- **Whole Exome Sequencing (WES):** Protein-coding regions
- **Targeted sequencing:** Specific genes/regions
- **RNA-seq:** Transcriptome profiling

Functional Genomics Data:

- **ChIP-seq:** Protein-DNA interactions
- **ATAC-seq:** Chromatin accessibility
- **Hi-C:** 3D chromosome organization
- **Single-cell omics:** Cell-type resolution

Data Scale:

WGS: 100 GB Per sample

WES: 5 GB Per sample

RNA-seq: 2 GB Per sample

ChIP-seq: 1 GB Per sample

Population studies:

Petabyte scale

Computational Requirements

Modern genomics datasets require specialized algorithms, distributed computing, and efficient data structures for analysis.

Machine Learning Challenges in Genomics

Data Challenges:

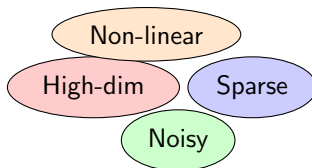
- **High dimensionality:** $p \gg n$ problem
- **Sparsity:** Many zeros in count data
- **Noise:** Technical and biological variation
- **Batch effects:** Platform/protocol differences
- **Missing data:** Dropout, technical failures

Biological Complexity:

- Non-linear relationships
- Epistatic interactions
- Temporal dynamics
- Tissue/cell-type specificity

Methodological Considerations:

- **Feature selection:** Curse of dimensionality
- **Normalization:** Cross-sample comparability
- **Validation:** Population stratification
- **Interpretability:** Biological relevance
- **Reproducibility:** Computational workflows



Genomic Data

Solution Approaches

Regularization, dimensionality reduction, ensemble methods, deep learning, and domain-specific preprocessing are key strategies.

Current ML Applications in Genomics

Variant Analysis:

- Variant calling algorithms
- Pathogenicity prediction
- Population genetics
- Pharmacogenomics

Gene Expression:

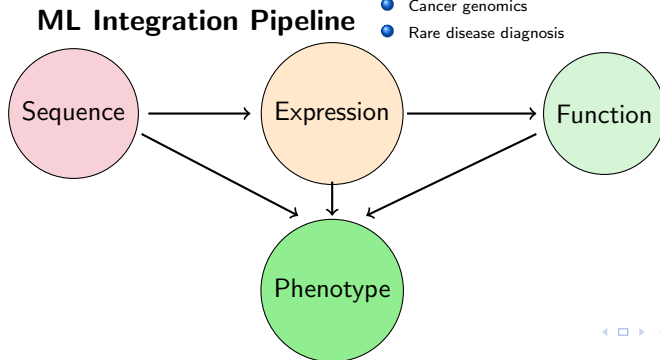
- Differential expression
- Co-expression networks
- Single-cell analysis
- Spatial transcriptomics

Regulatory Genomics:

- Motif discovery
- Enhancer prediction
- Chromatin state modeling
- 3D genome organization

Clinical Applications:

- Disease risk prediction
- Drug response modeling
- Cancer genomics
- Rare disease diagnosis



Emerging Trends and Technologies

Technical Advances:

- **Long-read sequencing:** PacBio, Oxford Nanopore
- **Single-cell multi-omics:** Simultaneous measurements
- **Spatial omics:** Tissue-level resolution
- **Real-time analysis:** Edge computing

AI/ML Developments:

- Deep learning architectures
- Graph neural networks
- Transformer models for sequences
- Federated learning approaches

Integration Challenges:

- Multi-modal data fusion
- Cross-species translation
- Population diversity
- Ethical considerations

Precision Medicine

ML + Genomics ← Personalized Therapy

Disease Prevention

Future Outlook

The integration of advanced ML techniques with comprehensive genomic data promises to revolutionize our understanding of biology and medicine.

Key Takeaways

Central Dogma Foundations:

- DNA \rightarrow RNA \rightarrow Protein information flow provides the basis for computational modeling
- Each step introduces complexity and variation that can be analyzed computationally
- Understanding molecular processes is essential for effective ML feature engineering

Computational Perspectives:

- Genomic data presents unique challenges: high dimensionality, sparsity, noise
- Multiple data types require integrated analytical approaches
- ML applications span from basic research to clinical translation

Looking Forward:

- Emerging technologies continue to generate new data types and scales
- Advanced ML methods offer unprecedented analytical capabilities
- Integration of biological knowledge with computational methods is crucial

Next: Types of genomic data: DNA sequences, RNA-seq, ChIP-seq, ATAC-seq

Thank You!

Questions?



Email: sali85@student.gsu.edu