



Machine Learning for Genomics

Introduction to Genomics and ML - Overview of ML in Biological Contexts

Sarwan Ali

Department of Computer Science
Georgia State University

 Understanding Machine Learning in Genomics 

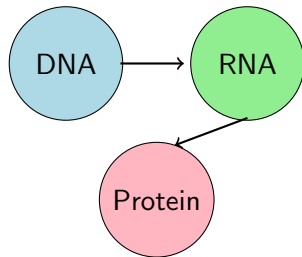
Today's Learning Journey

- 1 Introduction to Biological Data
- 2 Machine Learning Paradigms in Biology
- 3 Common ML Algorithms in Genomics
- 4 Specific Applications
- 5 Challenges and Considerations
- 6 Future Directions
- 7 Summary

What Makes Biological Data Unique?

Key Characteristics:

- High-dimensional data
- Noisy measurements
- Heterogeneous data types
- Complex dependencies
- Limited sample sizes



Central Dogma

Challenge

Traditional statistical methods often fail due to the **curse of dimensionality** and complex biological relationships.

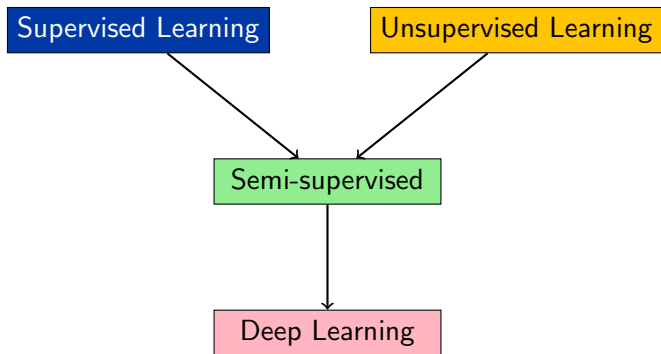
Types of Genomic Data

Data Type	Size	Description	ML Application
DNA Sequences	3.2B bp	Genetic code	Pattern recognition
RNA-seq	$10^4 - 10^5$ genes	Gene expression	Clustering, classification
ChIP-seq	Genome-wide	Protein-DNA binding	Peak calling
Hi-C	$10^6 - 10^9$ contacts	3D genome structure	Network analysis
Single-cell	$10^3 - 10^4$ cells	Cell heterogeneity	Dimensionality reduction

Key Insight

Each data type requires specialized ML approaches due to unique characteristics and challenges.

ML Paradigms: Overview



Biological Context:

- Labels often expensive/difficult to obtain
- High noise-to-signal ratio
- Need for interpretable models

Supervised Learning in Genomics

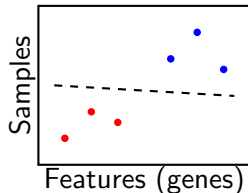
Classification Tasks:

- Disease prediction from genetic variants
- Tissue type classification from expression
- Functional annotation of genes
- Pathogenic variant identification

Regression Tasks:

- Gene expression prediction
- Drug response prediction
- Protein binding affinity

Classification Example



Example: Cancer Classification

Input: Gene expression profiles (20,000+ features)

Output: Cancer type (breast, lung, colon, etc.)

Challenge: $p \gg n$ problem (more features than samples)

Unsupervised Learning in Genomics

Why Unsupervised Learning?

- Discover hidden patterns in biological data
- Reduce dimensionality for visualization
- Identify cell types or disease subtypes
- Find co-expressed gene modules

Clustering Applications:

- Single-cell RNA-seq cell typing
- Patient stratification
- Gene co-expression networks
- Protein family classification

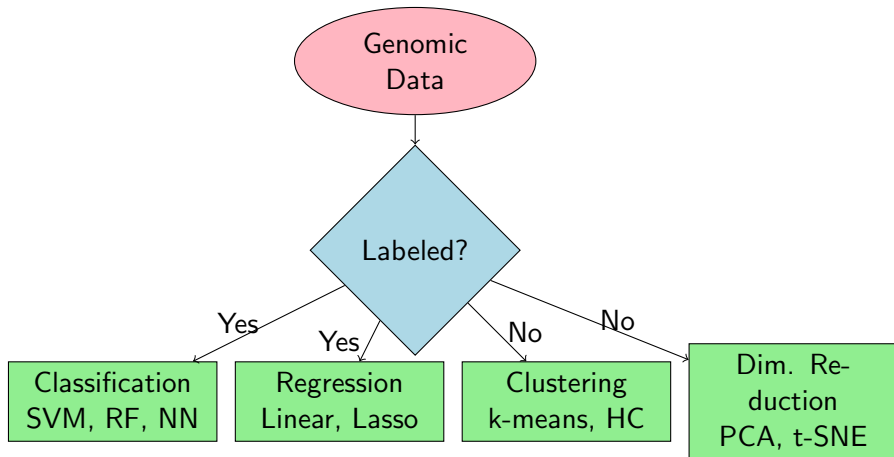
Dimensionality Reduction:

- PCA for expression data
- t-SNE for visualization
- UMAP for single-cell data
- Autoencoders for feature learning

Challenge

Biological interpretation of clusters can be difficult without domain knowledge.

Algorithm Selection Framework



Popular Algorithms: Strengths and Weaknesses

Algorithm	Strengths	Weaknesses	Genomics Use
Random Forest	Handles missing data, Feature importance	Poor with continuous numerical relationships	Gene selection, Disease prediction
SVM	High-dimensional data, Kernel flexibility	Parameter tuning, Not probabilistic	Sequence classification, Protein prediction
Neural Networks	Complex patterns, Non-linear	Black box, Overfitting risk	Deep genomics, Image analysis
Naive Bayes	Fast, interpretable Probabilistic	Independence assumption	Text mining, Functional annotation
k-means	Simple, fast	Assumes spherical clusters Need to specify k	Expression clustering, Cell type discovery

Selection Criteria

Consider: data size, interpretability needs, computational resources, and biological context.

Feature Selection in High-Dimensional Biology

The Curse of Dimensionality:

- Typical: 20,000+ genes, 100-1000 samples
- Many features irrelevant or redundant
- Risk of overfitting increases

Filter Methods:

- Statistical tests (t-test, ANOVA)
- Correlation-based selection
- Information gain
- Mutual information

Embedded Methods:

- LASSO regularization: $\min_{\beta} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$
- Elastic Net: $\min_{\beta} \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$

Wrapper Methods:

- Forward/backward selection
- Recursive feature elimination
- Genetic algorithms

Application 1: Gene Expression Analysis

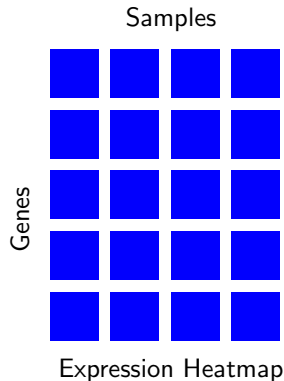
Problem: Identify differentially expressed genes between conditions

Traditional Approach:

- 1 Normalize expression data
- 2 Apply statistical tests (t-test, Wilcoxon)
- 3 Correct for multiple testing (FDR)
- 4 Set significance thresholds

ML Approach:

- 1 Feature selection (variance filtering)
- 2 Classification (disease vs. healthy)
- 3 Feature importance ranking
- 4 Pathway enrichment analysis



Tools

DESeq2, edgeR (traditional) vs. Machine Learning approaches using scikit-learn, TensorFlow

Application 2: Sequence Analysis

DNA/RNA/Protein Sequence Classification

Sequence Representation:

- One-hot encoding
- k-mer frequencies
- Position weight matrices
- Word embeddings

Common Tasks:

- Promoter prediction
- Splice site detection
- Protein family classification
- Regulatory element identification

One-hot Encoding Example:

	A	T	G	C
A	1	0	0	0
T	0	1	0	0
G	0	0	1	0
C	0	0	0	1

Sequence: **ATGC**

1,0,0,0	0,1,0,0	0,0,1,0	0,0,0,1
---------	---------	---------	---------

Challenge


Variable sequence lengths require padding or specialized architectures (RNNs, CNNs).

Application 3: Single-Cell Analysis

Single-cell RNA sequencing (scRNA-seq) Analysis Pipeline

Raw count matrix → Quality control & filtering → Normalization & scaling

Dimensionality reduction (PCA) → Clustering (Louvain) → Visualization (UMAP/t-SNE)



ML Challenges:

- Dropout (zeros) in expression data
- Batch effects between experiments
- Cell type identification
- Trajectory inference (pseudotime)

Major Challenges in Biological ML

Data Challenges:

- High noise levels
- Missing values
- Batch effects
- Small sample sizes
- Class imbalance

Computational Challenges:

- Memory limitations
- Processing time
- Scalability issues

Biological Challenges:

- Interpretability needs
- Biological validation
- Reproducibility
- Generalization across populations

Ethical Considerations:

- Privacy concerns
- Algorithmic bias
- Clinical translation

Key Insight

Success requires close collaboration between computational scientists and biologists.

Evaluation Metrics in Biological Context

Beyond Accuracy: Biologically Relevant Metrics

Classification Metrics:

- Sensitivity (Recall): $\frac{TP}{TP+FN}$
- Specificity: $\frac{TN}{TN+FP}$
- Precision: $\frac{TP}{TP+FP}$
- F1-score: $\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
- AUC-ROC

Clustering Metrics:

- Silhouette score
- Adjusted Rand Index
- Normalized Mutual Information
- Biological homogeneity

Important

In medical contexts, false negatives (missing disease) may be more costly than false positives. Choose metrics that reflect biological and clinical priorities.

Cross-validation Strategies:

- Stratified k-fold for balanced classes
- Leave-one-patient-out for personalized medicine
- Temporal splits for longitudinal data

Interpretability in Biological ML

Why Interpretability Matters:

- Scientific discovery requires understanding mechanisms
- Clinical applications need explainable decisions
- Regulatory approval often requires interpretable models
- Debugging and improving models

Model-Agnostic Methods:

- SHAP (SHapley Additive exPlanations)
- LIME (Local Interpretable Model-agnostic Explanations)
- Permutation importance
- Partial dependence plots

Model-Specific Methods:

- Linear model coefficients
- Decision tree paths
- Random forest feature importance
- Neural network attention mechanisms

Example

SHAP values can identify which genes contribute most to cancer classification, providing biological insights for further investigation.

Emerging Trends in Genomics ML

Deep Learning Applications:

- Convolutional Neural Networks for sequence motifs
- Recurrent Neural Networks for temporal data
- Graph Neural Networks for biological networks
- Transformers for sequence analysis
- Variational Autoencoders for data generation

Multi-omics Integration:

- Genomics + Transcriptomics + Proteomics
- Multi-modal learning approaches
- Tensor factorization methods

Federated Learning:

- Privacy-preserving collaboration
- Multi-institutional studies
- Addressing data sharing limitations

Foundation Models:

- Pre-trained on large biological datasets
- Transfer learning for specific tasks
- Examples: DNABERT, ProtBERT, scBERT

Causal Inference:

- Moving beyond correlation
- Understanding biological mechanisms
- Drug target discovery

Python Libraries:

- `scikit-learn`: General ML
- `pandas`: Data manipulation
- `numpy`: Numerical computing
- `scipy`: Scientific computing
- `matplotlib/seaborn`: Visualization
- `scanpy`: Single-cell analysis
- `BioPython`: Biological computing

Cloud Platforms:

- Google Cloud Platform (Life Sciences API)
- Amazon Web Services (AWS Batch)
- Microsoft Azure (Genomics)
- Terra (Broad Institute)

R Libraries:

- `Bioconductor`: Bioinformatics suite
- `DESeq2`: Differential expression
- `Seurat`: Single-cell analysis
- `randomForest`: Machine learning
- `caret`: Classification and regression

Deep Learning:

- `TensorFlow/Keras`
- `PyTorch`
- `JAX`

Key Takeaways


- 1 **Biological data is unique:** High-dimensional, noisy, and complex requiring specialized ML approaches
- 2 **Multiple paradigms apply:** Supervised, unsupervised, and semi-supervised learning all have important roles
- 3 **Feature selection is crucial:** Dealing with high-dimensional data requires careful feature engineering
- 4 **Evaluation must be biologically relevant:** Standard ML metrics may not capture biological significance
- 5 **Interpretability is essential:** Black-box models are often insufficient for biological applications
- 6 **Collaboration is key:** Success requires close partnership between computational and biological experts
- 7 **Field is rapidly evolving:** Deep learning and multi-omics approaches are opening new possibilities

Next Steps

Practice with real genomic datasets and explore domain-specific tools and libraries.

Thank You!

 Questions?

 Contact: sali85@student.gsu.edu