Machine Learning for Genomics

Key Databases and Resources: NCBI, Ensembl, UCSC Genome Browser

Sarwan Ali

Department of Computer Science
Georgia State University

Today's Learning Journey

1. Introduction to Genomic Databases
2. NCBI: National Center for Biotechnology Information
3. Ensembl: Comparative Genomics Hub
4. UCSC Genome Browser: Visualization and Tracks
5. Database Comparison and Selection
6. Practical Applications and Case Studies
7. Emerging Trends and Future Directions

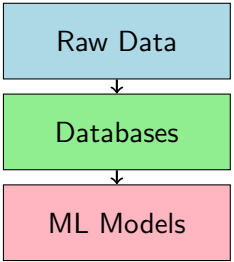
Why Genomic Databases Matter for ML

The Data Challenge:

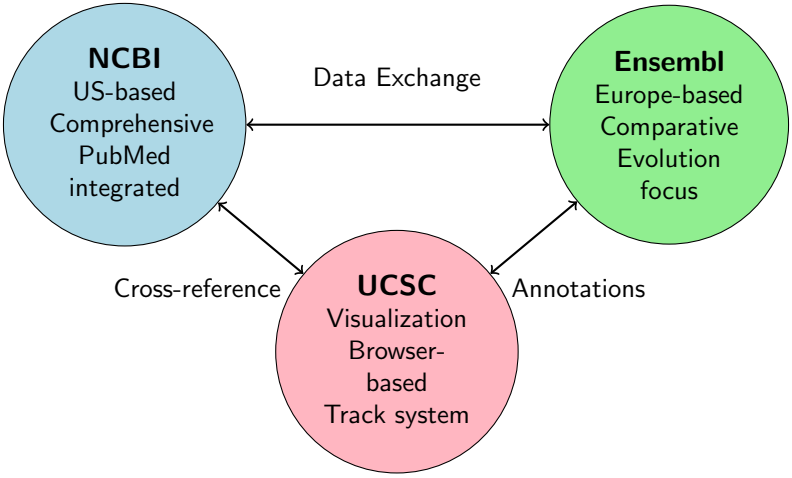
- Human genome: 3.2 billion base pairs
- 20,000+ protein-coding genes
- Millions of genetic variants
- Multi-omics data integration

ML Applications:

- Variant effect prediction
- Gene expression analysis
- Disease risk assessment
- Drug discovery



The Big Three: Overview



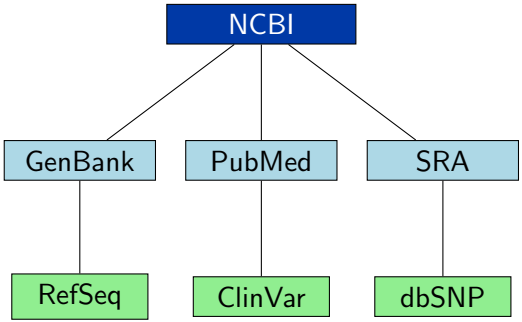
NCBI: The Comprehensive Repository

Key Features:

- Established: 1988
- Location: Bethesda, MD, USA
- Part of National Library of Medicine
- 40+ interconnected databases

URL: `https://www.ncbi.nlm.nih.gov/`

Mission: Advance science and health by providing access to biomedical and genomic information



NCBI Major Databases for ML

Database	Content	ML Use Cases	Access Methods
GenBank	DNA sequences	Sequence classification, motif discovery	E-utilities, FTP
RefSeq	Curated sequences	Reference datasets, annotation training	E-utilities, BLAST
SRA	Raw sequencing data	Feature extraction, quality control	SRA Toolkit, Cloud
dbSNP	Genetic variants	GWAS, variant effect prediction	E-utilities, VCF
ClinVar	Clinical variants	Pathogenicity prediction	E-utilities, XML
PubMed	Literature	Text mining, knowledge graphs	E-utilities, PMC OA

Data Formats: FASTA, GenBank, VCF, SAM/BAM, SRA, XML, JSON

NCBI E-utilities: Programmatic Access

E-utilities Suite: RESTful web services for database queries

Main Tools:

- `esearch` - Search databases
- `efetch` - Retrieve records
- `elink` - Find related records
- `einfo` - Database information
- `esummary` - Document summaries

Example Query:

```
# Search for BRCA1 sequences
https://eutils.ncbi.nlm.nih.gov/
entrez/eutils/esearch.fcgi?
db=nucleotide&term=BRCA1&
retmode=json

# Fetch sequence data
https://eutils.ncbi.nlm.nih.gov/
entrez/eutils/efetch.fcgi?
db=nucleotide&id=ACCESSION&
rettype=fasta
```

Rate Limits: 3 requests/second (10/second with API key)

NCBI for Machine Learning: Practical Examples

1. **Variant Effect Prediction:**
   - ClinVar: Training labels
   - dbSNP: Population frequencies
   - RefSeq: Reference sequences
2. **Gene Expression Analysis:**
   - GEO: Expression datasets
   - SRA: Raw RNA-seq data
   - PubMed: Literature context
3. **Sequence Classification:**
   - GenBank: Diverse sequences
   - RefSeq: High-quality references
   - Taxonomy: Phylogenetic labels
4. **Literature Mining:**
   - PubMed: 34M+ abstracts
   - PMC: Full-text articles
   - MeSH: Controlled vocabulary

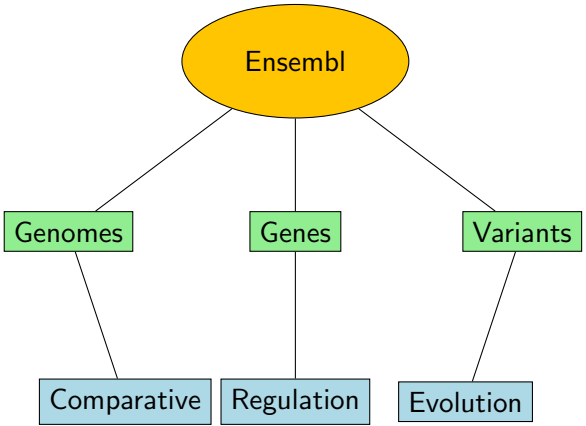
Ensembl: European Perspective on Genomics

Key Features:

- Established: 1999
- Location: Hinxton, UK (EMBL-EBI)
- Focus: Comparative genomics
- 270+ species genomes

URL: `https://www.ensembl.org/`

Mission: Provide comprehensive annotations for vertebrate genomes with emphasis on evolutionary relationships



Ensembl Data Types and ML Applications

Data Type	Description	ML Applications	Format
Gene Annotations	Protein-coding, non-coding genes	Gene prediction, classification	GTF, GFF3
Comparative Genomics	Orthology, paralogy	Phylogenetic ML, evolution	TSV, XML
Regulatory Features	Promoters, enhancers	Regulatory element prediction	BED, GFF
Variation Data	SNPs, indels, CNVs	Population genetics, GWAS	VCF, JSON
Expression Data	Tissue-specific expression	Expression prediction	TSV, JSON
Protein Domains	Pfam, InterPro domains	Protein function prediction	JSON, TSV

Ensembl REST API: Modern Access Pattern

RESTful Architecture: JSON-based, language-agnostic access

Key Endpoints:

- `/lookup/` - Gene/transcript info
- `/sequence/` - Genomic sequences
- `/homology/` - Comparative data
- `/variation/` - Variant information
- `/vep/` - Variant Effect Predictor

Example Queries:

```
# Get gene information
GET /lookup/symbol/homo_sapiens/BRCA1
Content-Type: application/json

# Get sequence data
GET /sequence/region/human/
17:43044294..43125364:1
Content-Type: text/x-fasta

# Variant Effect Prediction
POST /vep/human/region
{"variants": ["17:43044295 G A"]}
```

Rate Limits: 15 requests/second (55,000/hour)

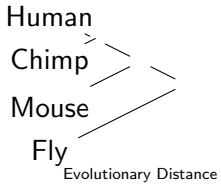
Ensembl Comparative Genomics: Evolutionary ML

Phylogenetic Trees:

- 270+ species relationships
- Gene trees vs species trees
- Duplication/speciation events

Orthology Predictions:

- One-to-one orthologs
- One-to-many relationships
- Confidence scores



ML Applications:

- Function transfer prediction
- Evolutionary rate estimation
- Conservation scoring

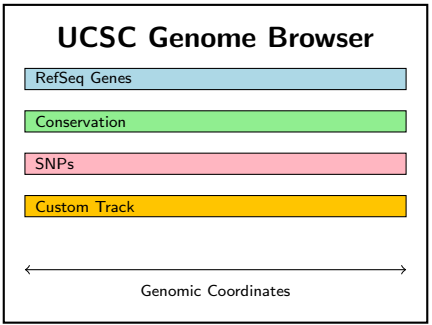
UCSC Genome Browser: The Visual Pioneer

Key Features:

- Established: 2000
- Location: Santa Cruz, CA, USA
- Focus: Genome visualization
- Track-based architecture

URL: `https://genome.ucsc.edu/`

Mission: Provide interactive visualization of genomic data with custom track integration



UCSC Track Types and Data Integration

Track Type	Data Source	ML Relevance	File Formats
Gene Predictions	RefSeq, GenCode	Feature annotation training	GTF, genePred
Conservation	PhyloP, PhastCons	Evolutionary constraints	WIG, bigWig
Variation	dbSNP, 1000 Genomes	Population genomics	VCF, BED
Regulation	ENCODE, Roadmap	Regulatory ML models	narrowPeak, BED
Expression	GTEx, ENCODE	Tissue-specific prediction	bigWig, BED
Repeats	RepeatMasker	Sequence masking, bias	BED, RMOut
Custom Tracks	User-uploaded	Experimental validation	Multiple formats

Key Advantage: Visual integration of multiple data types for hypothesis generation

UCSC Table Browser: Bulk Data Access

Programmatic Interface: Query and download genomic annotations

Query Parameters:

- **Clade/Genome:** Species selection
- **Group/Track:** Data category
- **Region:** Genomic coordinates
- **Output:** Format selection

Output Formats:

- BED, GTF, GFF
- Custom fields
- Statistical summaries
- Sequence extraction

URL Structure:

```
# Table Browser Query
https://genome.ucsc.edu/cgi-bin/
hgTables?hgsid=SESSION&
clade=mammal&org=Human&
db=hg38&hgta_group=genes&
hgta_track=refSeqComposite&
hgta_table=refGene&
hgta_regionType=range&
position=chr17:43000000-44000000&
hgta_outputType=bed&
hgta_outFileName=BRCA1_region.bed
```

Tip: Use sessions to save and share browser configurations

UCSC Data Hubs: Scalable Track Integration

Track Hubs Architecture:

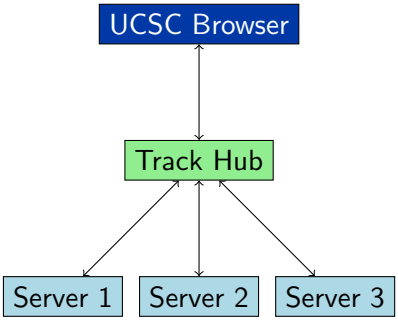
- Remote data hosting
- JSON/text configuration
- Scalable to thousands of tracks
- Real-time data updates

ML Applications:

- Model prediction visualization
- Training data validation
- Results comparison
- Interactive exploration

File Formats:

- bigWig: Signal data
- bigBed: Interval annotations
- BAM: Sequence alignments
- VCF: Variant calls



Distributed Data

Comparative Analysis: When to Use Which Database

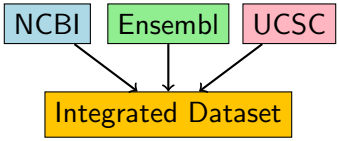
Criterion	NCBI	Ensembl	UCSC
Scope	Comprehensive, all domains	Vertebrate-focused	Genome visualization
Strengths	Literature integration, clinical data	Comparative genomics, clean annotations	Visual exploration, custom tracks
API Quality	E-utilities (XML/JSON)	Modern REST (JSON)	Table Browser (TSV/BED)
Update Frequency	Daily/weekly	Monthly releases	Continuous
ML Friendliness	Good programmatic access	Excellent structure	Visual validation
Data Volume	Largest repository	Curated subset	Annotation-focused
Best For	Literature mining, clinical ML	Evolutionary ML, clean datasets	Hypothesis generation, visualization

Integration Strategies: Combining Multiple Resources

Common Integration Patterns:

1. **Reference-Based:**
   - Use RefSeq IDs across platforms
   - Ensembl stable IDs for tracking
   - UCSC for coordinate mapping
2. **Coordinate-Based:**
   - Genomic coordinates (chr:start-end)
   - Assembly consistency (hg38/GRCh38)
   - LiftOver for version conversion
3. **Cross-Reference Tables:**
   - Gene symbol mapping
   - UniProt protein IDs
   - GO term annotations
4. **Metadata Integration:**
   - Sample information (SRA/GEO)
   - Experimental conditions
   - Quality metrics

Tool: BioMart for cross-database queries



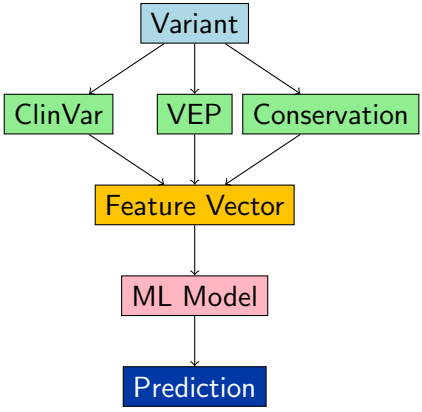
Case Study 1: Variant Effect Prediction Pipeline

Data Sources:

1. **NCBI ClinVar:** Pathogenic/benign labels
2. **Ensembl VEP:** Functional annotations
3. **UCSC Conservation:** Evolutionary scores
4. **NCBI dbSNP:** Population frequencies

Feature Engineering:

- Conservation scores (PhyloP, PhastCons)
- Protein domain disruption
- Splice site predictions
- Population allele frequencies



Case Study 2: Cross-Species Gene Expression Analysis

Objective: Predict tissue-specific expression across species

Data Integration:

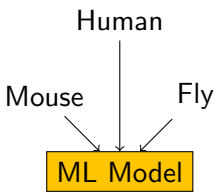
- **Ensembl:** Orthology relationships
- **NCBI GEO:** Expression datasets
- **UCSC:** Regulatory annotations

ML Approach:

- Multi-task learning
- Species as auxiliary tasks
- Phylogenetic regularization
- Transfer learning from model organisms

Results Validation:

- Cross-species correlation
- Functional enrichment
- Literature verification (PubMed)
- Experimental validation



Multi-species Learning

Best Practices for Database Integration

Data Quality Assurance:

- Version control and timestamps
- Data provenance tracking
- Quality score integration
- Batch effect correction

Reproducibility:

- Document database versions
- Archive query parameters
- Use persistent identifiers
- Container-based workflows

Performance Optimization:

- Batch API requests
- Local caching strategies
- Parallel processing
- Rate limit compliance

Error Handling:

```
try:
    response = api_call()
    if response.status_code == 429:
        time.sleep(retry_delay)
    elif response.status_code == 200:
        process_data(response.json())
except RequestException as e:
    log_error(e)
    use_cached_data()
```

Cloud-Based Genomics: The New Paradigm

Cloud Platforms:

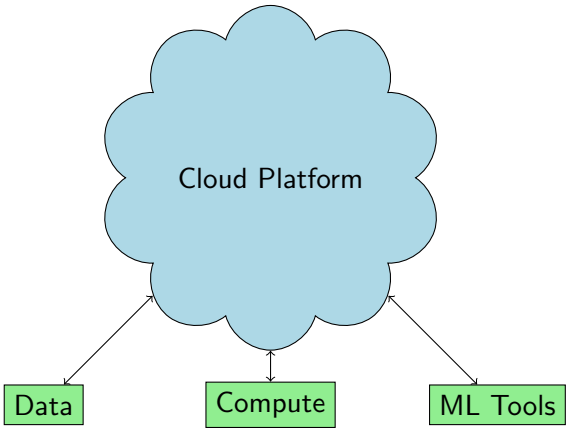
- **NCBI Cloud:** SRA on AWS/GCP
- **Ensembl Cloud:** Public datasets
- **Terra/AnVIL:** Analysis workspaces
- **Galaxy Project:** Workflow platforms

Advantages:

- Compute-near-data
- Scalable resources
- Reproducible workflows
- Cost optimization

ML Implications:

- Large-scale training
- Distributed computing
- Real-time inference
- Collaborative research



AI-Enhanced Database Features

Current Developments:

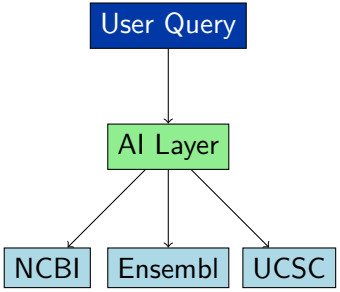
- **Semantic Search:** Natural language queries
- **Auto-annotation:** ML-based curation
- **Quality Control:** Automated error detection
- **Recommendation:** Relevant dataset suggestions

Future Possibilities:

- Conversational interfaces
- Predictive data modeling
- Automated hypothesis generation
- Real-time literature integration

Example Applications:

- "Find genes similar to BRCA1"
- Auto-generated gene summaries
- Variant pathogenicity prediction
- Cross-database entity linking



Summary and Key Takeaways

Database Selection Criteria:

- **NCBI:** Comprehensive, clinical focus
- **Ensembl:** Comparative, clean annotations
- **UCSC:** Visualization, hypothesis generation

Integration Best Practices:

- Use multiple sources for validation
- Maintain data provenance
- Version control everything
- Plan for scalability

ML Success Factors:

- Quality over quantity
- Domain knowledge integration
- Cross-validation strategies
- Biological interpretation

Future Trends:

- Cloud-native approaches
- AI-enhanced interfaces
- Real-time data streams
- Multi-modal integration

Remember: The database is your foundation - choose wisely, integrate carefully!

Resources and Further Reading

Official Documentation:

- NCBI E-utilities Guide
- Ensembl REST API Tutorial
- UCSC Table Browser Manual
- Galaxy Training Materials

Programming Libraries:

- Biopython (NCBI access)
- pyensembl (Ensembl data)
- pybedtools (UCSC formats)
- pandas (data manipulation)

Key Papers:

- NCBI Resource Coordinators (2018)
- Ensembl 2023 update (Nucleic Acids Res)
- UCSC Genome Browser (Nature Biotechnol)
- Genomic data science workflows

Online Courses:

- Coursera: Genomic Data Science
- edX: Introduction to Bioinformatics
- EMBL-EBI Training Portal
- Galaxy Project Tutorials

Questions & Discussion

 Thank You! 

Email: sali85@student.gsu.edu

Next: Linear algebra for genomics applications