25BIO801 Comparative Genomics with AI and Cloud Computing 2 2 0 4

Course Description: This interdisciplinary course familiarizes students with the principles and applications of comparative genomics, emphasizing the utilization of artificial intelligence (AI) and cloud computing to analyze and interpret complicated genetic data. Students will investigate how comparative genomics reveals evolutionary links, functional conservation, and gene divergence among species.

The course will encompass fundamental principles of comparative genomics, including gene orthology, synteny, genome alignment, and phylogenetics.

Practical implementations of machine learning (ML) and deep learning models for genome annotation, evolutionary analysis, and pattern identification.

Cloud platforms (AWS, Google Cloud, CyVerse) are utilized for extensive genomic data storage, processing, and collaboration.

Practical expertise with bioinformatics tools (BLAST, OrthoFinder, JBrowse2) and programming in Python for artificial intelligence model construction.

Students will undertake project-based assignments utilizing authentic genomic information from plants, animals, or microorganisms, and acquire skills in designing reproducible procedures on scalable cloud infrastructure.

Learning Objectives: This course provides students with a distinctive combination of genomics, artificial intelligence, and cloud computing skills that are progressively sought after in academia, industry, and research institutions. Students will acquire practical expertise in genomic data analysis, constructing machine learning models, and deploying bioinformatics processes on scalable cloud platforms. Graduates of this course will be adeptly equipped for professions in genome research, precision medicine, agri-genomics, biotechnology, and AI-driven life sciences through the amalgamation of biological knowledge and computational proficiency. The course cultivates critical thinking, problem-solving, and data-driven decision-making—crucial competencies for bioinformatics positions in pharmaceutical firms, research institutions, and genomics startups.

Pedagogy: The pedagogy will comprise a blend of lecture-based instruction and active learning approaches, such as participation in a journal club to present, discuss, and critique research papers, as well as participatory workshops.

Course Outline

Module 1: Introduction to Comparative Genomics

- Fundamentals of Comparative Genomics
  - Genome organization across species
  - Evolutionary relationships and phylogenomics
  - Genomic synteny and orthology
- Comparative Genomics Tools and Databases

    - BGDB.org, NCBI Genome, Ensembl, UCSC Genome Browser
    - Tools: CoGe, BLAST, OrthoMCL, COG/KOG
- Genome Evolution and Functional Genomics
    - Gene duplication and divergence
    - Non-coding DNA, regulatory elements, and epigenomics
    - Structural variations and genome rearrangements

Module 2: Applications of AI in Genomics and Bioinformatics

- Introduction to AI in Genomics
    - Overview of AI and machine learning (ML) in biological research
    - Standard ML algorithms used in genomics
    - Tools:
        - SciKit-Learn for bioinformatics
        - TensorFlow and PyTorch for deep learning applications
- AI in Genome Annotation
    - Automated gene prediction
    - Functional annotation using AI models
    - Protein structure prediction (AlphaFold, RoseTTAFold)
    - Case Study: AI-assisted genome annotation of Arabidopsis thaliana
- AI for Variant Calling and Genomic Data Analysis
    - AI-powered SNP and indel detection
    - Predicting pathogenicity of genetic variants
    - Tool: DeepVariant (Google's AI-based variant caller)
- Deep Learning in Genomics
    - CNNs and RNNs for sequence classification
    - Transformer models for DNA sequence analysis (e.g., DNABERT)
    - Tool: DNABERT for genomic sequence prediction
    - Case Study: Deep learning for enhancer prediction in mammalian genomes
- AI for Comparative Genomics and Evolutionary Biology
    - AI-driven phylogenetics
    - Machine learning for detecting evolutionary selection
- Ethical Considerations in AI-Driven Genomics
    - Bias in AI models
    - Privacy and security of genomic data

Module 3: Cloud Computing for Genomic Research (AWS Focus)

- Introduction to Cloud Computing in Bioinformatics
    - Benefits of cloud computing
    - Overview of AWS for life sciences
- AWS Services for Genomics
    - Amazon S3 for genomic data storage
    - AWS Lambda for serverless computing
    - AWS Batch for large-scale genomic workflows
    - Hands-on Example: Running BLAST on AWS Batch with NCBI datasets
- AI/ML on AWS for Bioinformatics

- Amazon SageMaker for genomic data analysis
- AWS Deep Learning AMI for AI-powered bioinformatics
- Using AI-powered AWS services for sequence analysis
- Use Case: Deploying a genome annotation model on AWS SageMaker

Workshops and Practical Sessions

Workshop 1: AI in Comparative Genomics

- Hands-on training with AI tools for genome annotation
- Using deep learning models for sequence classification
- Tools & Data:
  - DNABERT for sequence annotation
  - RefSeq genomes for annotation

Workshop 2: AWS for Genomics Data Storage and Processing

- Setting up AWS S3 for large-scale genomic datasets
- Running BLAST on AWS Batch
- Tools & Data:
  - Amazon S3 storage for whole-genome sequences
  - AWS Lambda for processing genomic metadata

Workshop 3: AI for Variant Calling and Disease Prediction

- Using AI-based tools for SNP annotation
- Machine learning for predicting disease-associated mutations
- Tools & Data:
  - DeepVariant for variant calling
  - 1000 Genomes dataset for variant prediction

Workshop 4: Cloud-Based AI Workflows for Comparative Genomics

- Deploying a genomic AI model on AWS SageMaker
- Running large-scale phylogenetic analyses on AWS
- Tools & Data:
  - Amazon SageMaker for ML model deployment
  - OrthoDB datasets for phylogenetic analysis

Participation Component (PhD-Level Addendum)

Title: Scholarly Engagement: Journal Club & Critical Discourse in Comparative Genomics

Overview:
PhD students will actively participate in a structured journal club format, engaging in the presentation, discussion, and critique of peer-reviewed scientific publications related to comparative genomics, AI-driven genomics, and cloud-based genomic computing. This component complements the technical and computational modules by fostering critical thinking, academic dialogue, and scientific communication skills.

Objectives:

- Develop expertise in reading and evaluating primary scientific literature.
- Strengthen oral communication through presentations and peer discussions.
- Encourage deeper understanding of emerging trends in comparative genomics.
- Foster interdisciplinary connections between genomics, AI, and cloud infrastructure.

Format:

- Weekly or biweekly sessions during the semester.
- Each student will:
  - Select a recent scientific paper relevant to course themes.
  - Present the study, methods, results, and implications to the class.
  - Lead a structured discussion, including critiques and future directions.
- Students are expected to participate actively in peer presentations.

Example Topics:

- Recent advances in phylogenomic tree reconstruction using AI.
- Novel tools for genome alignment and synteny analysis.
- Applications of transformer models in functional genomics.
- Case studies on genome annotation using cloud-deployed models.
- Ethical concerns in AI-driven variant prediction pipelines.

References

Comparative Genomics
Author: Melody S. Clark (Ed.)
Publisher: Springer
Description: A foundational reference for understanding comparative genomics methodologies and applications.

Bioinformatics and Functional Genomics
Author: Jonathan Pevsner
Publisher: Wiley-Blackwell
Description: Widely used text that introduces key bioinformatics tools, including those used in genome comparisons.

Machine Learning for Bioinformatics and Biostatistics *Editors*: Paolo Frasconi, Riccardo Martoglia, Andrea Passerini *Publisher*: Springer
Description: Introduces ML approaches applied in genomic data analysis.

Deep Learning for the Life Sciences
Authors: Bharath Ramsundar, Peter Eastman, Patrick Walters, Vijay Pande
Publisher: O'Reilly Media
Description: Practical guide to using deep learning on biological datasets, including genomics.

Cloud Computing for Science and Engineering
Authors: Ian Foster & Dennis Gannon
Publisher: MIT Press
Description: A straightforward guide on using cloud platforms (AWS, Google Cloud, etc.) for scalable data analysis.

Genomes 4
Author: T.A. Brown
Free online version available
Description: Offers a comprehensive overview of genome structure and analysis.

Online Tutorials & Docs:

- NCBI BLAST+ documentation
- OrthoFinder GitHub
- JBrowse2 User Guide
  CyVerse Learning Center

Cloud Computing Platforms:

- AWS Educate
- Google Cloud Genomics
- CyVerse Discovery Environment

Course Outcomes:

- Comprehend fundamental ideas of comparative genomics and their applications.
- Discover how artificial intelligence transforms genetic research, annotation, and data analysis.
- Acquire practical expertise with artificial intelligence techniques for genomic analysis.
- Examine the significance of cloud computing (AWS) in genomics and bioinformatics research.
- Facilitate practical workshops on AWS services for data storage, processing, and machine learning in genomics.