BIOINFORMATICS INTERNSHIP PROGRAM SYLLABUS

A guide for interns and supervisors

Advanced Bioinformatics Internship Program: AI, Cloud & Next-Gen Omics

Module 1: Introduction to Bioinformatics and Advanced Biological Databases

- Overview of bioinformatics, omics technologies, and role in precision medicine
- Advanced biological databases: NCBI, Ensembl, UniProt, AlphaFold Database (AFDB), STRING 12.5
- Emerging databases: MetaGraph for petabase-scale sequence search, Database Commons catalog
- Multi-omics data integration and federated database systems
- API-based data retrieval using NCBI E-utilities and Ensembl REST API
- Cloud-based genomic data repositories: Registry of Open Data on AWS, ENA Cloud

Hands-on Training:

- Query advanced databases using programmatic access (Python/R APIs)
- Navigate AlphaFold Protein Structure Database for predicted structures
- Use STRING 12.5 for regulatory network analysis with directionality
- Retrieve bulk genomic data from cloud repositories
- Integrate multi-source data using federated queries

Assignments:

1.

2.

3.

Module 2: Advanced Sequence Alignment and Deep Learning Approaches

- Advanced alignment algorithms: HAlign4 for millions of sequences, BetaAlign (AI-based MSA)
- Deep learning-based sequence alignment using transformer models
- Profile HMM-based searches with HMMER3 and HH-suite
- Ultra-fast alignment tools: DIAMOND for BLASTX/BLASTP, CS-BLAST for context-specific search
- GPU-accelerated alignment: CUDASW++ for Smith-Waterman on GPUs
- Cloud-based alignment services: AWS Batch for parallel BLAST workflows
- Phylogenomic analysis with IQ-TREE2 and RAxML-NG

Hands-on Training:

- Align millions of sequences using HAlign4 with parallel computing
- Apply BetaAlign transformer models for deep learning-based MSA
- Perform sensitive homology searches with HH-suite profile HMMs
- Run GPU-accelerated alignments for large-scale datasets
- Deploy BLAST workflows on AWS cloud infrastructure
- Construct maximum-likelihood phylogenetic trees with bootstrap support

Assignments:

1. illion+ qu ce u ng li n4 ar c pare per rma v lusta
2.
3.

Module 3: AlphaFold3-Based Structure Prediction and Validation

- AlphaFold3 architecture: Evoformer blocks, diffusion networks, and confidence scoring
- Protein structure prediction with AlphaFold3 for monomers and multimers
- Protein-protein complex prediction and interface analysis
- Protein-nucleic acid interactions (DNA/RNA binding prediction)
- Protein-ligand binding pose prediction and drug design applications
- Structure quality assessment: pLDDT, PAE, ipTM metrics

- Limitations of AlphaFold3 for intrinsically disordered regions and flexible domains
- Integration with experimental data: cryo-EM, X-ray crystallography validation
- RoseTTAFold2 and ESMFold as alternative prediction tools

Hands-on Training:

- Predict protein structures using AlphaFold3 web server and local installation
- Model protein-protein complexes and analyze binding interfaces
- Predict protein-DNA/RNA interactions for transcription factor studies
- Generate protein-ligand binding poses for drug discovery
- Validate predictions against experimental structures (PDB)
- Calculate confidence metrics and identify high/low confidence regions
- Compare AlphaFold3 with RoseTTAFold2 and ESMFold predictions

Assignments:

1. multi-domain protein using Alp
2. protein c and
3. Predict transcription factor-DN ing and identi .

Module 4: Advanced Systems Biology and Multi-Omics Integration

- Advanced network biology: regulatory networks, signaling pathways, metabolic networks
- STRING 12.5 for directed regulatory interactions and co-complex networks
- Multi-omics data integration: genomics, transcriptomics, proteomics, metabolomics
- Network topology analysis: centrality measures, community detection, network motifs
- Dynamic network modeling and pathway flux analysis
- Machine learning for network inference and gene regulatory network reconstruction
- Single-cell multi-omics network analysis
- Cloud-based systems biology platforms

Hands-on Training:

- Build directed regulatory networks using STRING 12.5
- Integrate multi-omics datasets for systems-level analysis

- Perform network topology analysis and identify hub genes/proteins
- Apply machine learning for gene regulatory network inference
- Analyze single-cell multi-omics networks
- Use Cytoscape with advanced plugins for network visualization

Assignments:

1.
2.
3.

Module 5: Advanced Computer-Aided Drug Design (CADD) and AI-Driven Approaches

- Modern drug discovery pipeline: AI-driven target identification and validation
- Advanced molecular docking: SwissDock 2024 (Attracting Cavities 2.0), OEDocking suite
- Deep learning-based docking: DiffDock, EquiBind, TANKBind
- Covalent docking and induced-fit docking protocols
- Virtual screening on cloud platforms: AWS-based high-throughput docking
- Fragment-based drug design and structure-based optimization
- ADMET prediction using AI models
- Binding free energy calculations: MM/GBSA, MM/PBSA, FEP/TI methods

Hands-on Training:

- Perform advanced docking with SwissDock 2024 Attracting Cavities engine
- Apply deep learning docking tools (DiffDock) for pose prediction
- Run covalent docking for targeted covalent inhibitor design
- Deploy virtual screening workflows on AWS cloud
- Calculate binding free energies using MM/GBSA
- Predict ADMET properties using AI-based tools
- Optimize lead compounds using structure-activity relationships

Assignments:

1.

2.

3. .

Module 6: Advanced Linux, HPC, and Container Technologies

- Advanced Linux system administration for bioinformatics servers
- High-Performance Computing (HPC) cluster architecture and job scheduling
- SLURM workload manager for HPC job submission and management
- Containerization with Docker and Singularity for reproducible workflows
- Kubernetes for orchestrating containerized bioinformatics pipelines
- Version control with Git and collaborative development with GitHub/GitLab
- Advanced shell scripting and workflow automation
- Environment management: Conda, Mamba, and virtual environments
- Nextflow and Snakemake workflow management systems

Hands-on Training:

- Configure and manage bioinformatics software on Ubuntu/CentOS servers
- Submit and manage jobs on HPC clusters using SLURM
- Create Docker containers for bioinformatics tools and workflows
- Build Singularity containers for HPC environments
- Deploy Nextflow/Snakemake pipelines with containerization
- Set up Git repositories and collaborative workflows
- Automate complex analyses with advanced bash scripting

Assignments:

1.

2.

3.

Module 7: Advanced Molecular Dynamics Simulations and Machine Learning Potentials

- GROMACS 2025 new features: PLUMED interface, neural network potentials, AMD GPU support
- Advanced force fields: AMBER ff19SB with CMAP support, CHARMM36m, Martini3 coarse-grained
- Neural Network Potentials (NNPs) for ab initio accuracy simulations
- Enhanced sampling methods: umbrella sampling, metadynamics, replica exchange
- Free energy perturbation (FEP) and thermodynamic integration (TI)
- GPU-accelerated MD simulations: CUDA and HIP backends
- Cloud-based MD simulations on AWS and Google Cloud
- Advanced trajectory analysis: RMSD, RMSF, principal component analysis, dynamic cross-correlation
- Markov State Models (MSM) for conformational dynamics
- Integration with AlphaFold3 structures for MD simulation setup

Hands-on Training:

- Install and configure GROMACS 2025 with GPU and PLUMED support
- Set up MD simulations using neural network potentials
- Run enhanced sampling simulations with metadynamics
- Perform free energy calculations using FEP/TI methods
- Analyze complex trajectories using principal component analysis
- Build Markov State Models from MD trajectories
- Deploy MD workflows on AWS cloud infrastructure
- Prepare AlphaFold3 structures for MD simulations

Assignments:

1.
2.
3.

Module 8: Cloud Computing and Big Data Infrastructure for Bioinformatics

- Cloud platforms for bioinformatics: AWS HealthOmics, Google Cloud Life Sciences, Azure Genomics
- AWS services: EC2, S3, Batch, Lambda, and HealthOmics for genomics workflows
- Serverless bioinformatics: AWS Lambda functions for event-driven analysis
- Scalable storage solutions: S3, EFS, FSx Lustre for genomic data
- Big data processing: Apache Spark for genomics, Dask for parallel computing
- Infrastructure as Code (IaC): Terraform and CloudFormation for reproducible deployments
- Cost optimization strategies for cloud-based bioinformatics
- Data security and compliance: HIPAA, GDPR in cloud environments
- Registry of Open Data on AWS for public genomic datasets

Hands-on Training:

- Deploy end-to-end genomics pipeline on AWS HealthOmics
- Create serverless bioinformatics functions with AWS Lambda
- Set up scalable storage with S3 and FSx Lustre
- Process large-scale genomic data with Apache Spark
- Deploy infrastructure using Terraform/CloudFormation
- Implement cost monitoring and optimization strategies
- Access and analyze public datasets from AWS Registry of Open Data

Assignments:

1.
2.
3.

Module 9: Advanced Programming and AI Integration in Bioinformatics

- Advanced Python: Object-oriented programming, decorators, generators, async programming
- Bioconductor ecosystem: Advanced genomic analysis in R
- Deep learning frameworks: TensorFlow, PyTorch, JAX for biological applications

- Large language models for biology: ESM-2, ProtGPT, BioBERT
- Graph neural networks for molecular property prediction
- Attention mechanisms and transformers for sequence analysis
- AI-driven protein design: ProteinMPNN, RFdiffusion
- AutoML for bioinformatics: automated feature engineering and model selection
- Real-time data streaming and analysis with Apache Kafka
- API development with FastAPI for bioinformatics services

Hands-on Training:

- Build deep learning models with PyTorch for genomic variant calling
- Apply protein language models (ESM-2) for structure prediction
- Implement graph neural networks for drug-target interaction
- Use transformers for DNA sequence motif discovery
- Design novel proteins with AI-based tools
- Create REST APIs for bioinformatics tools using FastAPI
- Deploy ML models with model serving platforms (TensorFlow Serving)

Assignments:

1. [illegible]
2. [illegible]
3. [illegible]

Module 10: Advanced Capstone Project with Industry Standards

- End-to-end bioinformatics project design and implementation
- Integration of multiple omics data types and analysis methods
- Application of AI/ML techniques to biological problems
- Cloud-based scalable analysis infrastructure
- Reproducible research practices: containers, version control, documentation
- Scientific communication: manuscript writing, figure creation, presentations
- Open science principles: data sharing, code publication, preprints
- Peer review process and collaborative research

Hands-on Training:

- Design comprehensive research project addressing real biological question
- Implement analysis pipeline using advanced tools from all modules
- Deploy project on cloud infrastructure with full reproducibility
- Create publication-quality figures and comprehensive documentation
- Present findings in scientific format with peer review
- Publish code on GitHub and data on public repositories

Assignments:

1.
2.
3.

****