HPC-Enabled Curriculum Enhancement of a General Bioinformatics Course at Albany State University.

SGX3 award
# 2231406

Revised Course Description

The revised General Bioinformatics Course offers a comprehensive and hands-on approach to modern bioinformatics, with a strong emphasis on high-performance computing (HPC) applications. This course integrates fundamental bioinformatics concepts with practical HPC skills, providing students with a solid foundation in both areas. The curriculum begins with an introduction to basic programming, coding, and HPC concepts, including access to scientific gateways and bash scripting. As the course progresses, it delves into advanced topics such as large-scale genomic data analysis, structural bioinformatics, HPC-based visualization of biological datasets, and HPC-enabled comparative genomics. Students will gain practical experience through project-based modules, utilizing a local supercomputer setup for hands-on learning. The course culminates in a final exam structured as an HPC-based project, allowing students to apply their acquired knowledge in a real-world scenario. By combining theoretical knowledge with practical skills in HPC and bioinformatics, this course prepares students for the computational challenges of modern biological research and equips them with valuable tools for data-intensive scientific discovery

Implementation Schedule

The course will be offered during the Spring and Fall Semesters, from 2025.

Weeks 1-9: Foundations
Weeks 1-2: Course introduction, scientific gateway access, bash basics
Weeks 3-4: Programming fundamentals for bioinformatics
Weeks 5-6: HPC fundamentals and introduction to bioinformatics
Weeks 7-9: Biological databases, sequence alignment, BLAST

Weeks 10-15: Advanced HPC Projects
Weeks 10-11: Large-scale genomic data analysis with HPC
Weeks 12-13: HPC for structural bioinformatics and data visualization
Weeks 14-15: HPC-enabled comparative genomics
Week 16: Review and Project Preparation

Week 17: Final Exam (HPC-based project)

Throughout each semester, students will utilize a local 4GPU supercomputer for hands-on HPC experience.

Sample HPC/Gateways Exercise

1. Parallel BLAST Search
Use an HPC cluster to perform a parallel BLAST search on a large sequence set. Split a FASTA file, submit parallel jobs, and combine results.

2. Large-Scale Multiple Sequence Alignment
Utilize HPC to align hundreds or thousands of sequences using tools like MAFFT or Clustal Omega optimized for parallel execution.

3. Genome Assembly Pipeline
Develop a genome assembly workflow on a scientific gateway. Upload raw data, configure and run assembly tools, and visualize results.

4. Machine Learning for Protein Structure Prediction
Use HPC to train and run a machine learning model (e.g. simplified AlphaFold-inspired) for protein structure prediction.

5. Large-Scale Phylogenetic Analysis
Conduct phylogenetic analysis on a large dataset using HPC. Perform multiple sequence alignment and build trees using parallel tools like RAxML or IQ-TREE.

6. Virtual Screening for Drug Discovery
Conduct a virtual screening experiment using molecular docking on an HPC system, utilizing tools like AutoDock Vina to identify potential drug candidates

Resource Needs/List

1. Textbook: "Bioinformatics and Functional Genomics" by Jonathan Pevsner, 3rd edition

2. Computer lab or personal computers with internet access

3. Python programming environment (e.g., Anaconda distribution)

4. Biopython library

5. Access to biological databases (e.g., NCBI, Ensembl, UniProt)

6. BLAST software or web interface
Multiple sequence alignment tools (e.g., Clustal Omega, MUSCLE)

7. Phylogenetic analysis software (e.g., MEGA, PhyML)
Machine learning libraries (e.g., scikit-learn, TensorFlow)

8. Access to AlphaFold database and visualization tools (e.g., PyMOL)

9. High-performance computing resources for large-scale analyses (e.g. local 4GP)

10. Version control system (e.g., Git)
Access to scientific literature databases

Gateway Community Mentor Syllabus Suggestions

Environment Setup
- Install Python 3.8+ and Visual Studio Code
- Install Python packages: biopython, pandas, numpy, matplotlib
- Learn basic Bash commands for file operations

Version Control
- Install Git and create a GitHub account
- Configure Git with your name and email
- Create a repository for your bioinformatics projects

Bioinformatics Tools
- Familiarize yourself with NCBI databases and BLAST
- Install Clustal Omega and HMMER
- Learn to use MEGA and RAxML for phylogenetic analysis

Programming
- Focus on Python basics, file I/O, Biopython, Pandas, and Matplotlib
- Practice writing Bash scripts for automation
- Learn to use Scikit-learn for machine learning tasks

Additional Resources
- Explore Biopython documentation
- Review NCBI tutorials
- Check relevant GitHub repositories for example code
- Remember to commit to your work regularly and practice good version control habits..

Datasets

- Public Bioinformatics Dataset from JEFworks GitHub repository: https://github.com/JEFworks/public-bioinformatics-datasets

- NCBI Datasets: https://github.com/ncbi/datasets

- Metagenomic Dataset from the CURE: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7511545/

- Genomics Data Lake from Azure Open Datasets: https://learn.microsoft.com/da-dk/azure/open-datasets/dataset-genomics-data-lake

Resources / Science Gateways

- ACCESS-CI
- EDGE Bioinformatics Gateway
- KEGG (Kyoto Encyclopedia of Genes and Genomes)
- NERSC (National Energy Research Scientific Computing Center)
- Science Gateways Community Institute (SGCI)
- ORNL Portal
- AnVIL

Use Cases

- Genomic Data Analysis
- Protein Structure Prediction
- Metagenomics Analysis
- Machine Learning for Genomic Classification
- Comparative Genomics

Possible Expansions

Team Teaching / Co-Teaching:
This course is taught in the Department of Natural Sciences. Students who enrolled in this course are Biology majors with little or no background in Computer Science. A 'sister' course *(CSCI 2300 Computational Informatics)* is taught in the Computer Sciences program Both courses can be co-taught to actuate a cross-disciplinary perspective for students.

Future consideration:
(1) Either consolidate the two courses and carefully arrange the contents; or
(2) Develop interdisciplinary collaboration (set up a Computational Biology Lab to expand opportunities in both directions to our students.

Authors

Olabisi Ojo, PhD
Department of Natural Sciences
(Biology Program)
Albany State University
olabisi.ojo@asurams.edu

HPC/Gateways Co-Mentor
Sheryl Bradford, PhD
School of Science, Aviation, Health & Technology.
Elizabeth City State University
sbradford@ecsu.edu

HPC/Gateways Mentor
Hector Corzo, PhD.
National Center for Computational Sciences,
Oak Ridge National Laboratory
hernandezchf@ornl.gov

MORE INFORMATION →
https://hackhpc.github.io/facultyhack-gateways23