GCB 5370-001 202510 Advanced Computational Biology

Jump to Today ✎ Edit

GCB537 Advanced Computational Biology

Term: Spring 2025

Instructor: Yoseph Barash & Noam Auslander

Objectives:

1. Discuss current topics and related papers in genomics and computational biology
2. Review important concepts for computer science and statistics as they apply to computational biology.
3. Implement a subset of those in the context of tackling compbio research questions (shift from question answering to research oriented tasks/mindset).
4. Learn to evaluate, criticize, and summarize research papers in genomics and computational biology
5. Close the gap on basic genomic/bioinformatics tools you should know by now (TA hours)
6. Experiment/evaluate, and try to improve tools/algorithms from topics covered in the course (final project).

Requirement: Background in statistics, biology, genetics and genomics, and computer science.

NOTE:

1. **This is NOT a bioinformatics lab. Prior background is assumed.**
2. **None-GCB students need to be approved by the instructors.**

Time and Location: Tuesdays and Thursdays 1:45-3:15pm, Richards B205

TA hours:

Di Wu, **Monday 10am - 11am**, Zoom (https://upenn.zoom.us/j/2415154726)

Farica Zhuang, **Wednesday 10am - 11am**, Zoom (https://upenn.zoom.us/j/94615025853?pwd=tZxtIlvaJoFsoi1NvgbVmJmXakTo8z.1)

Course format:

The course consists of lectures, paper discussions, analysis/coding tasks, and final project presentation. Some lectures review current topics in computational biology, while others review material in computer science and statistical modeling relevant for these topics and compbio in general. In paper discussion classes, papers are selected to match topics discussed in the review lectures, with emphasis on algorithm design and best practice for data interpretation and presentation. In parallel, students will be given coding and analysis assignments that correspond to topics and/tools discussed in class. Final project will consist of a comp research task (e.g. as related to the student's research) that each student will need to define and describe how the plan to solve using the tools/methods learned in class.

Paper discussion.

1. The course is divided into "units" covering current topics in comp bio research

Each unit starts with a review lecture, followed by one or more paper discussion classes led by the instructors or students.

1. For paper discussion classes, 1-2 students will be designated to present the paper. Emphasis will be given to understanding the computational methods, model assumptions, evaluation process, overall significance and open issues/directions. To ensure the quality of the presentation, send PowerPoint files to the instructor or discuss with the instructor *at least two days before the scheduled presentation*.
2. The leading team will also submit questions on the paper to the instructors five days before the discussion. The question list will be circulated to the entire class before the class. Students will submit their anonymous responses and the presenters will grade those responses, together with the instructors/TA/guest lecturers.
3. After the class, all other students at the presentation will send a grade (between 1=unprepared and 5=excellent) or any constructive comments to the instructor by email *by end of the day*. Comments will be forwarded to the presenters anonymously to improve their future presentations.

Coding/analysis assignments:

As the course progresses students will be given analysis tasks that involve coding. The analysis tasks will relate to topics covered in class. Some of those tasks may be over predefined data/tools and some may be more open ended (“bring your data” and/or “come up with a scientific question and matching analysis for a given/chosen dataset”). Learning proper scientific coding will be implicitly included in these tasks while a major focus will be on the analysis of the data. Some of the tasks will be interconnected and build upon each other.

Final project:

Each student will need to submit a final project. The final project is a proposal for how to tackle a specific computational research task. The proposal (up to 2 pages long) should include a clear definition of the task (the "What"), along with a brief review of related work (was it done already? how? differences from what you propose?), and the details of how the students plans to go about addressing this task using the tools/methods learned in class (the "How"). Students are encouraged to select a task which fits within their own research but can also define a task that is generally within the topics/challenges in compBio covered in class. Students can NOT use existing work (e.g. describe the what and how from a paper). Emphasis should be given on clear/precise definition of the task and the evaluation of performance (how do you know it's doing well? How do you compare? etc.)

Course evaluation:

Evaluation will be a combination of class participation, paper presentation, answers to papers’ questions, coding/analysis assignments and the final project.

Attendance/Participation:

Given the nature of the course, **participation and attendance in class is mandatory** - you can't just show up whenever and read material later. If you are unable to attend a specific day/class you should let us know in advance, get it approved, and coordinate with us completing the assignments as needed (e.g. submit paper questions).

Note: Showing up late to class (beyond the 1-5min grace due to weather or conflicting class schedule) will not be allowed. If you need to be more than a few minutes late you need to seek approval in advance.

Also, **active class participation** is encouraged and will be part of the overall evaluation in the course. That includes answering paper questions, discussions/questions during class, feedback to presenters etc.

Using external resources/copying:

Students can and are encouraged to discuss problems/exercises at a high level (e.g., "How do you think you can validate this claim?"). However, no sharing of code is allowed, and students should exclusively write their own code, analysis, and final project for the course. Detection of suspected copying from external sources (including past submissions or public material) will be transferred directly to the University authorities who handle ethics, with potentially severe consequences for the students involved.

Tentative list of topics:

Efficient coding in genomics (TA reviews): Scripting, unix text editors, reproducible coding, numerical stability, efficient coding using matrix manipulations.

Optimization methods - (Stochastic) Gradient Descent, generalized EM.

Probabilistic view of regression

Clustering: supervised/unsupervised, probabilistic models (Naive Bayes, GMM), spectral

Dimensionality reduction methods

MCMC

Ensemble learning: Boosting vs Bagging, Decision trees, Random forest vs Tree boosting.

Current computational topics in human genetics: GWAS, QTL, Fine Mapping, Colocalization.

DL models for genomics: CNN, (V)AE, Transformers.

DL optimization and interpretation for genomics.

RNA Sequencing, transcriptomics, Gene expression and RNA splicing.

Peak Calling (ChiP/ATAC/CLIP etc.)

Sequence Motif Finding - From traditional ML to DL

Measures of performance, class imbalance and overfitting with biological data

Feature selection: filter, wrapper, and embedded methods. Utility and caveats

Multi omics integration

Molecular evolution: phylogenetics and selection

Assembly: naïve, string graph assembly, de Bruijn graph-based (briefly)

Single cell Genomics/Transcriptomics

Cancer genomics and ML/DL in cancer research

Missing data and data censoring. Data imputation, semi supervised and PU learning

Course Summary:

Date	Details	Due
Thu Jan 16, 2025	Course Intro + RNA Sequencing (https://canvas.upenn.edu/calendar?event_id=4704562&include_contexts=course_1841073)	1:45pm to 3:15pm
Tue Jan 21, 2025	RNA-Seq (cont.) (https://canvas.upenn.edu/calendar?event_id=4704567&include_contexts=course_1841073)	1:45pm to 3:15pm
Thu Jan 23, 2025	Paper Presentation DESeq1+2 (https://canvas.upenn.edu/calendar?event_id=4704566&include_contexts=course_1841073)	1:45pm to 3:15pm
	Questions for Paper Presentation #1 - DESeq2 (https://canvas.upenn.edu/courses/1841073/assignments/13062786)	due by 1:45pm
	Feedback for Paper Presentation #1 - DESeq2 (https://canvas.upenn.edu/courses/1841073/assignments/13062773)	due by 11:59pm
Tue Jan 28, 2025	Single Cell Guest Lecture - Dana Silverfish (https://canvas.upenn.edu/calendar?event_id=4706246&include_contexts=course_1841073)	12am
Thu Jan 30, 2025	Matrix Decomposition Review, Correcting Confounders (https://canvas.upenn.edu/calendar?	1:45pm to 3:15pm

	event_id=4704542&include_contexts=course_1841073)	
Tue Feb 4, 2025	Paper Presentation #2 - Single Cell (https://canvas.upenn.edu/calendar?event_id=4704553&include_contexts=course_1841073)	1:45pm to 3:15pm
	Questions for Paper Presentation #2 - Single cell (https://canvas.upenn.edu/courses/1841073/assignments/13062811)	due by 1:45pm
	Feedback for Paper Presentation #2 - Single Cell (https://canvas.upenn.edu/courses/1841073/assignments/13062778)	due by 11:59pm
	A0 Shell Scripting Exercise (https://canvas.upenn.edu/courses/1841073/assignments/13062764)	due by 11:59pm
Thu Feb 6, 2025	Peak Calling (ChIP/CLIP/ATAC-seq) + HMM Reminder (https://canvas.upenn.edu/calendar?event_id=4704561&include_contexts=course_1841073)	1:45pm to 3:15pm
Tue Feb 11, 2025	Paper Presentation #3 - Peak Calling (https://canvas.upenn.edu/calendar?event_id=4704564&include_contexts=course_1841073)	1:45pm to 3:15pm
	Questions for Paper Presentation # 3- PureCLIP (https://canvas.upenn.edu/courses/1841073/assignments/13062808)	due by 1:45pm
Thu Feb 13, 2025	Measures of Performance, Class Imbalance, Overfitting (https://canvas.upenn.edu/calendar?event_id=4704555&include_contexts=course_1841073)	1:45pm to 3:15pm
Tue Feb 18, 2025	Feature Selection, Filter Wrapper, Embedded (https://canvas.upenn.edu/calendar?event_id=4704556&include_contexts=course_1841073)	1:45pm to 3:15pm

Thu Feb 20, 2025	Motif Finding Review (https://canvas.upenn.edu/calendar?event_id=4704550&include_contexts=course_1841073)	1:45pm to 3:15pm
Tue Feb 25, 2025	DL for Motif Finding + interpratation (https://canvas.upenn.edu/calendar?event_id=4704565&include_contexts=course_1841073)	1:45pm to 3:15pm
	A1 - Splicing Regulation and RBP Binding (https://canvas.upenn.edu/courses/1841073/assignments/13062765)	due by 11:59pm
Thu Feb 27, 2025	Paper Presentation #4 - Motif Finding (Basset) (https://canvas.upenn.edu/calendar?event_id=4704552&include_contexts=course_1841073)	1:45pm to 3:15pm
	Questions for Paper #4 - Basset (https://canvas.upenn.edu/courses/1841073/assignments/13062805)	due by 1:45pm
	Feedback for Paper #4 Presentation - Basset (https://canvas.upenn.edu/courses/1841073/assignments/13062769)	due by 11:59pm
Tue Mar 4, 2025	Regression Classification (https://canvas.upenn.edu/calendar?event_id=4704557&include_contexts=course_1841073)	1:45pm to 3:15pm
Thu Mar 6, 2025	Clustering (https://canvas.upenn.edu/calendar?event_id=4704554&include_contexts=course_1841073)	1:45pm to 3:15pm
Tue Mar 18, 2025	Paper Presentation #5 (clustering) (https://canvas.upenn.edu/calendar?event_id=4704544&include_contexts=course_1841073)	1:45pm to 3:15pm
	Questions for paper #5 - SPICi (https://canvas.upenn.edu/courses/1841073/assignments/13062815)	due by 1:45pm

	A2 - Motif Finding (https://canvas.upenn.edu/courses/1841073/assignments/13062766)	due by 11:59pm
Thu Mar 20, 2025	Intro to Human Genetics (https://canvas.upenn.edu/calendar?event_id=4704560&include_contexts=course_1841073)	1:45pm to 3:15pm
Tue Mar 25, 2025	TBD? Human Genetics Paper? (https://canvas.upenn.edu/calendar?event_id=4704548&include_contexts=course_1841073)	1:45pm to 3:15pm
Thu Mar 27, 2025	Single Cell #2 Guest Lecture - Avi Srivastava (https://canvas.upenn.edu/calendar?event_id=4706268&include_contexts=course_1841073)	12am
Tue Apr 1, 2025	Paper Presentation #7 (Single Cell) (https://canvas.upenn.edu/calendar?event_id=4704545&include_contexts=course_1841073)	1:45pm to 3:15pm
	Questions for Paper Presentation #7 (Single Cell) (https://canvas.upenn.edu/courses/1841073/assignments/13062813)	due by 1:45pm
	Feedback for Paper Presentation #7 Single Cell (https://canvas.upenn.edu/courses/1841073/assignments/13062783)	due by 11:59pm
Thu Apr 3, 2025	Cancer genomics (https://canvas.upenn.edu/calendar?event_id=4704540&include_contexts=course_1841073)	1:45pm to 3:15pm
Tue Apr 8, 2025	Comparative genomics, phylogenetics and genome assembly (https://canvas.upenn.edu/calendar?event_id=4704547&include_contexts=course_1841073)	1:45pm to 3:15pm
	A3 - EM Algorithm (https://canvas.upenn.edu/courses/1841073/assignments/13062767)	due by 11:59pm

Thu Apr 10, 2025	Microbiome (https://canvas.upenn.edu/calendar?event_id=4704558&include_contexts=course_1841073)	1:45pm to 3:15pm
Tue Apr 15, 2025	DL for sequential input in the context of DNA/RNA/protein: from RNNs to attention (transformers) (https://canvas.upenn.edu/calendar?event_id=4704541&include_contexts=course_1841073)	1:45pm to 3:15pm
Thu Apr 17, 2025	Deep Learning Model Interpretation (https://canvas.upenn.edu/calendar?event_id=4706269&include_contexts=course_1841073)	12am
Tue Apr 22, 2025	Ensemble learning – Boosting vs Bagging (https://canvas.upenn.edu/calendar?event_id=4704559&include_contexts=course_1841073)	1:45pm to 3:15pm
Thu Apr 24, 2025	MCMC Review (https://canvas.upenn.edu/calendar?event_id=4706270&include_contexts=course_1841073)	12am
Tue Apr 29, 2025	Course Summary: Open Challenges in Computational Biology (https://canvas.upenn.edu/calendar?event_id=4704546&include_contexts=course_1841073)	1:45pm to 3:15pm
	A4 - Cancer Genomics (https://canvas.upenn.edu/courses/1841073/assignments/13062768)	due by 11:59pm
	Feedback for Paper #5 Presentation - SPICi (https://canvas.upenn.edu/courses/1841073/assignments/13062770)	
	Feedback for Paper Presentation #3 - PureCLIP (https://canvas.upenn.edu/courses/1841073/assignments/13062780)	
	Feedback for Paper Presentation #9 (ExPecto)	