CSI 5180. Topics in Artificial Intelligence: Machine Learning for Bioinformatics

Marcel Turcotte

2024-12-19

> "In the not so distant past, data generation was the bottleneck, now it is data mining, or extracting useful biological insights from large, complicated datasets."
>
> Xu, C. & Jackson, S. A.
> *Machine learning and complex biological data.*
> **Genome Biology 20**, (2019).

> **Warning**
>
> This website is currently being developed. Specifically, the syllabus is in draft form and will be subject to modifications. Feedback and suggestions are most welcomed.

Course info

	Day	Time	Location
Lecture 1	Tuesday	11:30-12:50	VNR 2095
Lecture 2	Friday	13:00–14:20	VNR 2095
Office hours	Tuesday	13:00-14:20	STE 5106

Description

Machine learning theories and methods with applications to biological sequence data, gene expression, genomics and proteomics.

Learning outcomes

Upon completion of the course, you will be able to:

- **Encode** and **clean** biological data for machine learning applications
- **Apply** modern machine learning methods to solve bioinformatics problems
- **Find** optimal values for the hyperparameters a given machine learning algorithm and data set • Use a sound methodology for your machine learning projects
- **Critically review** scientific publications in this field
- **Locate** and **critically evaluate** scientific information
- **Present** scientific content to a small technical audience

Outline

Here is a tentative and ambitious course outline.

- Overview
- Essential Cell Biology (Part 1)
- Essential Cell Biology (Part 2)
- Essential Bioinformatics Skills (Databases, APIs, Frameworks)
- Fundamentals of Machine Learning
- Feature Engineering
- Data Imputation
- Dimensionality Reduction
- Unsupervised Learning
- Linear and Logistic Regression
- Decision Trees, Random Forests and eXtreme Gradient Boosting
- Extreme Learning Machines
- Hidden Markov Models
- Kernel Methods
- Support Vector Machines
- Deep Learning: Fundamentals
- Deep Learning: Embeddings
- Deep Learning: Architectures
- Concept and Rule-based
- Learning Graphs
- Ensemble
- Semi-supervised Learning
- Data Integration
- Automated Scientific Discovery

Grading

The final course grade will be calculated as follows:

Category	Percentage
Assignments	30% (10% x 3)
Project	20%
Presentation	10%
Examinations	40% (20% x 2)

Except in programs and courses for which language is a requirement, all students have the right to produce their written work and to answer examination questions in the official language of their choice, regardless of the course's language of instruction.

Assignments

The assignments are done individually and there will be three of them. They are programming assignments with specific learning objectives in mind. For example, the learning objectives for the first assignment include: encode biological data for a specific machine learning task, implement two metrics to compare sequence data, apply an unsupervised learning algorithm to summarize some data.

Python will be used for our assignments, along with popular machine learning libraries, including Scikit-Learn, Keras, and TensorFlow.

Presentation

Papers in (refereed) journals and conference proceedings are the main vehicle for communicating scientific information. You must select a publication that presents either a specialized application or a more efficient algorithm on a topic that has been presented in class.

There will be one student presentation per lecture. Each presentation will be related to the topic of the lecture. Students are randomly assigned a date.

Learning objectives

- Thoroughly study of a specific topic in bioinformatics
- Familiarity with the modes of communicating research
- Develop your presentation skills

Deliverable

- 15 – 20 minutes presentation

Project

Learning objectives

- Thorough study of a specific bioinformatics problem using two machine learning approaches
- Learning to study autonomously

Deliverable

You will replicate the results from a recent scientific publication. You must create a suitable data set and apply at least two distinct algorithms. You will apply the methodology proposed in class to select the values of the hyperparameters and evaluate the result. You must hand in your data, your source code, as well as a short report (5-10 pages).

Examinations

There will be a midterm and final examination. Students that have 90% or above for their combined grade for the assignments, midterm, and presentation can be exempted from writing the final examination.

Material and resources

- Lecture notes (slides) and complementary resources will be posted on the course Web site: turcotte.xyz/teaching/csi-5180/lectures

Below, you will find a number of references to **Springer** Link, which provides our community with access to journals, books, series, protocols and reference documents, access is restricted to the University of Ottawa, based on your IP address.

Monographs

- Hiroshi Mamitsuka, *Textbook of Machine Learning and Data Mining: with Bioinformatics Applications*. Global Data Science Publishing, 2018. (On Amazon.ca)
- Hiroshi Mamitsuka, Ed., *Data Mining for Systems Biology: Methods and Protocols*. Humana Press, 2018. (link.springer.com/book/10.1007/978-1-4939-8561-6)
- Ravi Bhramaramba and Akula Chandra Sekhar, *Application of Computational Intelligence to Biology*. Springer, 2016.
- Zengyou He, *Data Mining for Bioinformatics Applications*. Woodhead Publishing, 2015.
- Pradipta Maji Sushmita Paul, *Scalable Pattern Recognition Algorithms: Applications in Computational Biology and Bioinformatics*. Elsevier, 2014.
- Hiroshi Mamitsuka, Charles DeLisi, Minoru Kanehisa, Eds., *Data Mining for Systems Biology: Methods and Protocols*. Humana Press, 2013. (link.springer.com/book/10.1007/978-1-62703-107-3)
- Sumeet Dua and Pradeep Chowriappa, *Data Mining for Bioinformatics*. CRC Press, 2012. (On Amazon.ca)
- Zheng Rong Yang, *Machine learning approaches to bioinformatics*. World Scientific, ISBN 978-981-4287-30-2, 2010.
- Yan-Qing Zhang Jagath C. Rajapakse, *Machine Learning in Bioinformatics*. Wiley, 2008.
- Pierre Baldi and Søren Brunak, *Bioinformatics: The Machine Learning Approach*. MIT Press, 2001. (On Amazon.ca)

Essential cell biology

- Wiesława Widłak, *Molecular Biology: Not Only for Bioinformaticians*. Vol. 8248, Springer, 2013. link.springer.com/book/10.1007/978-3-642-45361-8
- Terence A. Brown, *Genomes*. Oxford: BIOS Scientific Publishing, 2007. (QH 447 .B76 2007) (NCBI Bookshelf)
- Bruce Alberts, Karen Hopkin, Alexander Johnson, David Morgan, Keith Roberts, Peter Walter, Rebecca Heald (2023) Essential Cell Biology. Sixth Edition. W. W. Norton.
- David M. Hillis, Craig H. Heller, Sally D. Hacker, David W. Hall, Marta J. Laskowski, Lauren A. O'Connell, and David E. Sadava. (2023). Life: The Science of Biology. 12th ed. Macmillan. ISBN: 9781319440992
- MIT OpenCourseware - Introduction to Biology

Relationship to CSI 5126: Algorithms in Bioinformatics

Algorithms in Bioinformatics is centered around the data structures and algorithms essential for tackling classic problems in bioinformatics. It covers topics such as suffix trees and suffix arrays, which are crucial for solving various string problems efficiently in linear time. Furthermore, dynamic programming plays a significant role in bioinformatics, being widely used for tasks such as aligning biological sequences or reconstructing ancestral states in phylogenies, among other applications. The course outline is described below:

> Fundamental mathematical and algorithmic concepts underlying computational molecular biology; physical and genetic mapping, sequence analysis (including alignment and probabilistic models), genomic rearrangements, phylogenetic inference, computational proteomics and systemic modelling of the whole cell.

No prior knowledge of bioinformatics should be needed to succeed with **Machine Learning for Bioinformatics Applications** and the intersection between the content of two courses should be minimum.

- https://turcotte.xyz/teaching/csi-5126/lectures/

Bioinformatics resources

For a high-level, first encounter with bioinformatics, I am suggesting the following textbook. Beware, this is the fifth edition. If you can, I recommend accessing the latest edition since high-throughput technologies are fast evolving. Here is what the editor had to say regarding the latest edition: "A host of new material includes new content on next generation sequencing, function prediction, sequence assembly, epigenomics, the bioinformatics of gene editing, and the effects of single nucleotide variants."

- Arthur Lesk, *Introduction to Bioinformatics*. Fifth ed., Oxford University Press, 2019.

For a practical and economical introduction, you might have a look at **The Biostar Handbook: Bioinformatics data analysis guide, 2023**, which also gives you access to a number of online courses. Namely, the guide provides an introduction to Unix and Conda, which are both of importance for Bioinformatics, but also for Machine Learning!

- biostar.myshopify.com

For those who would like to explore the subject even further, the following monographs are part of my short list of essential Bioinformatics books.

- Wing-Kin Sung, *Algorithms in Bioinformatics: A Practical Introduction*. Chapman & Hall/CRC, 2010. (QH 324.2 .S86 2010)

- R. Durbin *et al*, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 2000. (QP 620 .B576 1998)
- D. Gusfield, *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press, 1997. (QA 76.9 .A43 G87 1997)

On the Web

- **Johns Hopkins University**: Algorithms for DNA Sequencing
- **University of California San Diego**: Bioinformatics Specialization.
- **MIT 7.91 J**: Foundations of Computational and Systems Biology.
- Learning bioinformatics through **problem solving** at rosalind.info/.

Academic integrity

Academic fraud is an act by a student that may result in a false evaluation (including papers, tests, examinations, etc.). It is not tolerated by the University. Any person found guilty of academic fraud will be subject to sanctions.

Here are some examples of academic fraud:

- Plagiarism or cheating of any kind;
- Present research data that has been falsified;
- Submit a work for which you are not the author, in whole or part;
- Submit the same piece of work for more than one course without the written consent of the professors concerned.
- Please consult this webpage: it contains regulations and tools to help you avoid plagiarism.

An individual who commits or attempts to commit academic fraud, or who is an accomplice, will be penalized. Here are some examples of possible sanctions:

- Receive an "F" for the work or in the course in question;
- Imposition of additional requirements (from 3 to 30 credits) to the program of study;
- Suspension or expulsion from the Faculty.
- You can refer to the regulations on this web page

Student services

Academic writing help

At the Academic Writing Help Centre you will learn how to identify, correct and ultimately avoid errors in your writing and become an autonomous writer. In working with our Writing Advisors, you will be able to acquire the abilities, strategies and writing tools that will enable you to:

- **Master** the written language of your choice
- **Expand** your critical thinking abilities
- **Develop** your argumentation skills
- **Learn** what the expectations are for academic writing

Further information is available here:

- www.uottawa.ca/study/academic-support/academic-writing-help

Career services

Career Services offers various services and a career development program to enable you to recognize and enhance the employability skills you need in today's world of work.

- www.uottawa.ca/current-students/career-experiential-learning/career-development

Counselling service

Counselling and therapy is a confidential service for students who are facing life challenges. It is a safe space to explore new perspectives and build resilience.

- www.uottawa.ca/campus-life/health-wellness/counselling-therap

Access Service

The Access Service acts as an intermediary between students, their faculty and other University offices to ensure that the special needs of these students are addressed and that the best possible learning conditions are being offered.

Note that the University of Ottawa is affiliated with AERO and ACE services for the adaptation of accessible academic materials for students with perceptual disabilities. If you have any questions, please contact the Accessibility Librarian or the Access services for textbooks.

- www.uottawa.ca/study/academic-support/accommodation-services-available

Support and prevention of sexual violence

The University of Ottawa will not tolerate any act of sexual violence. This includes acts such as rape and sexual harassment, as well as misconduct that take place without consent, which includes cyberbullying. The University, as well as various employees and student groups, offers a variety of services and resources to ensure that all uOttawa community members have access to confidential support and information, and to procedures for reporting an incident or filing a complaint. For more information, please visit www.uOttawa.ca/sexual-violence-support-and-prevention.

Information sharing and copyright

All documents prepared by the course instructor, including assignments, course notes, and exams, are protected by copyright. Copying, digitizing, or publishing on a Web site is therefore a violation of copyright and is illegal.