CSI5180. Machine Learning for Bioinformatics Applications

Course overview

by Marcel Turcotte

Version November 6, 2019
Preamble
Course overview

**Machine Learning for Bioinformatics Applications** is about the analysis of complex biological data using modern machine learning methods. No prior machine learning knowledge is assumed. However, a basic understanding of probability and statistics is needed, as well as, calculus and linear algebra. Also, I am expecting that you can write programs in Python. Now, **what about biology?** Biology is important as bioinformatics strives to solve “real-world” problems. There will be at least two lectures introducing essential concepts of the molecular biology of the cell. Inevitably, we will revisit these concepts each time that a new problem will be introduced. At the very least, I am expecting a desire to learn more about biology.

**General objective :**

- **Summarize** the learning objectives and the expectations for this course
Learning objectives

- **Clarify** the proposition
- **Summarize** what bioinformatics is about
- **Give** an overview of the instructor’s background
- **Discuss** the syllabus
- **Articulate** the expectations

**Reading:**

Plan

1. Preamble
2. Proposition
3. About the course
4. About me
5. What is Bioinformatics?
6. Syllabus
7. What is Machine Learning?
8. Prologue
Proposition
“Using artificial intelligence, a Princeton University-led team has decoded the functional impact of such mutations in people with autism.”


Interpreting a genome: predicting the effect of any mutation

DeepSEA

Histone marks + DNase accessibility + Transcription factors

Interpret noncoding mutation's regulatory impact

https://oreilly.com/go/ainy19
“We address the challenge of detecting the contribution of noncoding mutations to disease with a deep-learning-based framework that predicts the specific regulatory effects and the deleterious impact of genetic variants.”

“Our predictive genomics framework illuminates the role of noncoding mutations in ASD [autism spectrum disorder] and prioritizes mutations with high impact for further study, and is broadly applicable to complex human diseases.”

“Together, the HMP1 and HMP2 phases have produced a total of 42 terabytes of multi-omic data.”

Integrative HMP (iHMP) Research Network Consortium.
The Integrative Human Microbiome Project.

Nature 569, 641648 (2019).
“MyExome, a new DNA test designed by Toronto entrepreneur Zaid Shahatit, claims to be able to provide a little insight into our personal quirks by testing 57 different genes that could determine our ability to metabolize certain things, sleep patterns and physical performance.”

Yuval Noah Harari argues that artificial intelligence and genetic engineering will play a central role shaping the future of society.
About the course
What this course is not

Although the following are of paramount importance, this is not what this course is about:

- **Computational Learning Theory:**
  - **Probably approximately correct learning** (PAC Learning) proposed by Leslie Valiant;
  - **VC theory** proposed by Vladimir Vapnik and Alexey Chervonenkis;
  - **Bayesian inference** influenced by Judea Pearl;
  - **Algorithmic learning theory** from E. Mark Gold;
  - **Online machine learning** from Nick Littlestone.

- Compression bounds and learnability in general.
The course focuses on practical applications of machine learning to **biological sequence data**, gene expression, **genomics** and **proteomics**.

- **Aurélien Géron.**
  *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow.*

- **Andriy Burkov.**
  *The Hundred-Page Machine Learning Book.*
  Andriy Burkov, 2019.
What I would like the course to be... 

- In future editions of this course:
  - Extensive set of examples
  - **Practical Machine Learning Applications in Bioinformatics** (textbook)
  - Hackathon, hackfest, codefest, and (friendly) **competitive challenges**;
  - Participation to **international competitions**:

- Activity in the **bioGARAGE**;
- **Guests** lectures.
Predicting protein stability changes upon mutation, intrinsically disordered protein region
Protein secondary and tertiary structure prediction
Prediction of anti-hypertensive peptides
Genome assembly, gene prediction, genome annotation
Identifying DNA landmark sites: methylation, splice site, promoters, protein binding sites, etc.
Prediction and prioritization of gene functional annotations.
Clustering and classification of non-coding RNA genes
Subtypes cancer classification
Toxicity, carcinogenicity, structure activity relationships
Predicting disease associations, identify robust prognostic gene signatures
Sub-cellular localization
Machine Learning Concepts

- Feature Engineering, Data Imputation, Dimensionality Reduction
- Unsupervised Learning
- Linear and Logistic Regression
- Decision Trees, Random Forests and eXtreme Gradient Boosting, Ensemble
- Hidden Markov Models
- Kernel Methods, Support Vector Machines
- Deep Learning: Fundamentals, Embeddings, Architectures
- Concept and Rule-based
- Learning Graphs
- Semi-supervised Learning
- Automated Scientific Discovery
Learning objectives

- **Encode** and **clean** biological data for machine learning applications
- **Apply** modern machine learning methods to solve bioinformatics problems
- **Find** optimal values for the hyperparameters of 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
About me
Professional experience

- 1989, Honours project, implementation of a graphical user interface for a protein folding/unfolding system

- 1989–95, Université de Montréal, graduate studies under the direction of Guy Lapalme (IRO), Robert Cedergren (Biochemistry), work on methods for building nucleic acids’ 3-D structures

- 1995–97, University of Florida, work with Steven A. Benner (Chemistry) on evolutionary-based approaches to predict protein secondary structure

- 1997–00, Imperial Cancer Research Fund (London/UK), work with Michael J.E. Sternberg and Stephen H. Muggleton (York) on the application of Inductive Logic Programming to discover automatically protein folding rules

- 2000–, University of Ottawa, work on nucleic acids secondary structure determination, motifs inference and pattern matching
Professional experience

- 1989, **Honours project**, implementation of a graphical user interface for a protein folding/unfolding system
- 1989–95, **Université de Montréal**, graduate studies under the direction of Guy Lapalme (IRO), Robert Cedergren (Biochemistry), work on methods for building nucleic acids’ 3-D structures
- 1995–97, **University of Florida**, work with Steven A. Benner (Chemistry) on evolutionary-based approaches to predict protein secondary structure
- 1997–00, **Imperial Cancer Research Fund** (London/UK), work with Michael J.E. Sternberg and Stephen H. Muggleton (York) on the application of Inductive Logic Programming to discover automatically protein folding rules
- 2000–, **University of Ottawa**, work on nucleic acids secondary structure determination, motifs inference and pattern matching
Professional experience

- **1989**, Honours project, implementation of a graphical user interface for a protein folding/unfolding system

- **1989–95**, Université de Montréal, graduate studies under the direction of Guy Lapalme (IRO), Robert Cedergren (Biochemistry), work on methods for building nucleic acids’ 3-D structures

- **1995–97**, University of Florida, work with Steven A. Benner (Chemistry) on evolutionary-based approaches to predict protein secondary structure
Professional experience

- 1989, **Honours project**, implementation of a graphical user interface for a protein folding/unfolding system
- 1989–95, **Université de Montréal**, graduate studies under the direction of Guy Lapalme (IRO), Robert Cedergren (Biochemistry), work on methods for building nucleic acids’ 3-D structures
- 1995–97, **University of Florida**, work with Steven A. Benner (Chemistry) on evolutionary-based approaches to predict protein secondary structure
- 1997–00, **Imperial Cancer Research Fund** (London/UK), work with Michael J.E. Sternberg and Stephen H. Muggleton (York) on the application of Inductive Logic Programming to discover automatically protein folding rules
Professional experience

- 1989, **Honours project**, implementation of a graphical user interface for a protein folding/unfolding system

- 1989–95, **Université de Montréal**, graduate studies under the direction of Guy Lapalme (IRO), Robert Cedergren (Biochemistry), work on methods for building nucleic acids’ 3-D structures

- 1995–97, **University of Florida**, work with Steven A. Benner (Chemistry) on evolutionary-based approaches to predict protein secondary structure

- 1997–00, **Imperial Cancer Research Fund** (London/UK), work with Michael J.E. Sternberg and Stephen H. Muggleton (York) on the application of Inductive Logic Programming to discover automatically protein folding rules

- 2000–, **University of Ottawa**, work on nucleic acids secondary structure determination, motifs inference and pattern matching
Application of inductive logic programming to discover rules governing the three-dimensional topology of protein structure.  

Exploiting protein structure in the post-genome era.  
Learning protein structure principles.
Oral Presentation.

Generating protein three-dimensional folds signatures using inductive logic programming.
Oral Presentation.


Mikhail Jiline, Stan Matwin, and Marcel Turcotte. 
Annotation Concept Synthesis and Enrichment Analysis. 
*Canadian AI 2010: Advances in Artificial Intelligence, 304–308, 2010.*

Mikhail Jiline, Stan Matwin, and Marcel Turcotte. 
Annotation Concept Synthesis and Enrichment Analysis: a Logic-Based Approach to the Interpretation of High-Throughput Experiments. 
Oksana Korol and Marcel Turcotte

Learning relationships between over-represented motifs in a set of DNA sequences.

*2012 IEEE Symposium on Computational Intelligence and Computational Biology, CIBCB 2012, 2012.*
Alexander R. Gawronski and Marcel Turcotte.
RiboFSM: Frequent subgraph mining for the discovery of RNA structures and interactions.
*BMC bioinformatics, 15*(S2), 2014.
Aseel Awdeh, Marcel Turcotte, and Theodore J. Perkins.
WACS: Improving peak calling by optimally weighting controls.

Aseel Awdeh, Marcel Turcotte, and Theodore J. Perkins.
WACS: Improving Peak Calling by Optimally Weighting Controls.
biorxiv.org.
What is Bioinformatics?
“Computers and specialized software have become an essential part of the biologists toolkit. Either for routine DNA or protein sequence analysis or to parse meaningful information in massive gigabyte-sized biological data sets, virtually all modern research projects in biology require, to some extent, the use of computers. (...) the very beginnings of bioinformatics occurred more than 50 years ago, when desktop computers were still a hypothesis and DNA could not yet be sequenced.”

“Broadly speaking, bioinformatics can be defined as a collection of mathematical, statistical and computational methods for analyzing biological sequences, that is, DNA, RNA and amino acid (protein) sequences.”

“Bioinformatics is the design and development of computer-based technology that supports life sciences. Using this definition bioinformatics tools and systems perform a diverse range of functions including: data collection, data mining, data analysis, data management, data integration, simulation, statistics, and visualization. Computer-aided technology directly supporting medical applications is excluded from this definition and is referred to as medical informatics.”

“Biologists that reduce bioinformatics to simply the application of computers in biology sometimes fail to recognize the rich intellectual content of bioinformatics. Bioinformatics has become a part of modern biology and often dictates new fashions, enables new approaches, and drives further biological developments.”

“In bioinformatics, so much is to be done, the raw material to hand is already so vast and vastly increasing, and the problems to be solved are so important (perhaps the most important of any science at present) we may be entering an era comparable to the great flowering of quantum mechanics in the first three decades of the twentieth century (…)

What is Bioinformatics?

https://youtu.be/182AzhLiwxo
Leonard **Adleman** (*Science*, December 1994) solved a particular instance of the Hamiltonian Path problem using DNA molecules!

⇒ A Hamiltonian path visits every node of a graph exactly once.
DNA computing is the theoretical study of the use of DNA molecules to solve challenging problems or as a new architecture (what class of problems can be solved, what are the properties, limits, etc.).
Biotechnology and biomedical engineering apply engineering approaches to problems dealing with biological systems.

Examples of biomedical engineering include developing biomedical devices for human implantation, drug delivery systems, simulation of organs and micro-fluids, medical imaging, and many more.
Bioinformatics courses on campus

- http://www.bioinformatics.uottawa.ca
- **CSI 5126.** Algorithms in bioinformatics
- **BNF5106** Bioinformatics\(^1\)
- **BCH5101** Analysis of -omics data

\(^1\)www.bioinformatics.uottawa.ca/stephane/bnf5106.syllabus.pdf
Starting from January 2008, Carleton University and the University of Ottawa offers a Collaborative Program leading to an **MSc degree with Specialization in Bioinformatics** or **MSc of Computer Science degree with Specialization in Bioinformatics**;

A proposal for a **Ph.D.** program is under review.

Wren, J. D. Bioinformatics programs are 31-fold over-represented among the highest impact scientific papers of the past two decades. *Bioinformatics* **32**(17):2686-91, September 2016.
Search Job Postings:

Postdoc: Critical Assessment of Function Annotation

Institution/Company: Iowa State University
Location: Ames, Iowa, United States
Job Type: Postdoctoral
Degree Level Required: PhD

The Friedberg Lab is seeking to fill a postdoc position in the Critical Assessment of Function Annotation. The Friedberg Lab is located at Iowa State University in Ames, Iowa. The lab is equipped with high performance computers, including GPU machines suitable for machine learning. The successful candidate will be offered a competitive salary. Applications accepted exclusively via the Iowa State University job site. The successful candidate must have excellent bioinformatics programming skills. More on the Friedberg Lab can be found here: https://iddo-friedberg.net/

Critical Assessment of Function Annotation: The Critical Assessment of Function Annotation or CAFA is an experiment designed to provide a large-scale assessment of computational methods that are dedicated to predicting protein function, using a time challenge. The successful candidate will need to write and implement assessment algorithms for the Critical Assessment of Function Annotation (CAFA), as well as assess the methods competing in the 4th CAFA challenge which will take place late 2019-2020. The postdoc will gain experience in working with cutting-edge machine learning software in bioinformatics, produce robust software for the ongoing CAFA work, and interact with a large and diverse international community of students and researchers participating in CAFA. They will be...
Course information

Web sites
- https://www.eecs.uottawa.ca/~turcotte/teaching/csi-5180/
- https://piazza.com/uottawa.ca/fall2019/csi5180
- https://uottawa.brightspace.com

Schedule
- Lectures: Tuesday, 13:00 to 14:30, and Thursday, 11:30 to 13:00, MNT 103
- Office hours: Tuesday from 14:30 to 16:00 at STE 5-106
- Official schedule: www.uottawa.ca/course-timetable

Evaluation
- 30% — assignments (3)
- 10% — presentation (1)
- 20% — project (1)
- 40% — examinations (2)
Deadlines

-Assignments
  - **A1** - October 10, 2019, 18:00
  - **A2** - October 31, 2019, 18:00
  - **A3** - November 21, 2019, 18:00

-Presentation
  - Schedule will be published on September 19, 2019
  - Presentations between October 1, 2019 and December 3, 2019

-Project
  - Outline - October 1, 2019
  - Report - December 3, 2019

-Examinations
  - Midterm - October 24, 2019
  - Final - December 5 to 18, 2019
What is Machine Learning?
“Let's start by telling the truth: *machines don't learn*.

(...)

just like *artificial intelligence is not intelligence*, machine learning is not learning.”

The Hundred-Page Machine Learning Book, Andriy Burkov, 2019

“A computer can be programmed so that it will learn to play a better game of checkers than *can be played by the person who wrote the program.*”

“Programming computers *to learn from experience should eventually eliminate the need for much of this detailed programming effort.*”

“Inspired by a radio talk given by Turing in 1951, Christopher Strachey went on to implement the world’s first machine learning program.”

“A computer program is said to **learn** from experience $E$ with respect to some class of tasks $T$ and performance measure $P$, if its performance at tasks in $T$, as measured by $P$, improves with experience $E$. ”

“A machine learning algorithm is a computational method based upon statistics, implemented in software, able to discover hidden non-obvious patterns in a dataset, and moreover to make reliable statistical predictions about similar new data.”

“The ability [of machine learning] to automatically identify patterns in data [...] is particularly important when the expert knowledge is incomplete or inaccurate, when the amount of available data is too large to be handled manually, or when there are exceptions to the general cases.”
Prologue
Summary

- A practical application of machine learning to biological data
A practical application of machine learning to biological data
Python programming skills and a love of biology are both expected
https://youtu.be/dtNMA46YgX4
Next module

- **Essential** Cell Biology (two lectures)
Chunming Xu and Scott A Jackson.
Machine learning and complex biological data.

Jian Zhou, Christopher Y Park, Chandra L Theesfeld, Aaron K Wong, Yuan Yuan, Claudia Scheckel, John J Fak, Julien Funk, Kevin Yao, Yoko Tajima, Alan Packer, Robert B Darnell, and Olga G Troyanskaya.
Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk.

Tom M Mitchell.
*Machine Learning*.