CSI5126. Algorithms in bioinformatics

Overview of the course content and expectations

Marcel Turcotte

School of Electrical Engineering and Computer Science (EECS)
University of Ottawa

Version September 6, 2018
Motd: scholarships

- www.uottawa.ca/graduate-studies/students/awards
- Intellectual independence
- Building up your curriculum vitae
- Natural Sciences and Engineering Research Council of Canada
- NSERC, CIHR, OGS, ...

Marcel Turcotte  CSI5126. Algorithms in bioinformatics
I want to learn about you.

- What is your name?
- Are you an undergraduate or a graduate student?
- If you are a graduate student:
  - Who is your supervisor?
  - Give us two or three sentences about your research topic.
- Where are you from?
- What background do you have in biology?
- Do you program in Java (at least the basics)?
- What are you expecting from this course?
1989, **Honours project**, implementation of a graphical user interface for a protein folding/unfolding system.
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
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

2000–, **University of Ottawa**, work on nucleic acids’ secondary structure determination, motifs inference and pattern matching
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
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
What is bioinformatics?
**TED**: Juan Enriquez on genomics and our future

http://www.ted.com/talks/juan_enriquez_on_genomics_and_our_future.html
“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 (…)

SIB - Swiss Institute of Bioinformatics

https://youtu.be/182AzhLiwxo
Atul Butte/Stanford at TEDMED 2012

https://youtu.be/dtNMA46YgX4
“Computers and specialized software have become an essential part of the biologist's 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.”

Leonard **Adleman** (*Science*, December 1994) solved a particular instance of the Hamiltonian Path problem using DNA molecules!

⇒ An Hamiltonian path visits every node of a graph exactly once.
What it’s not! (contd)

**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.
Other bioinformatics courses on campus

- [http://www.bioinformatics.uottawa.ca](http://www.bioinformatics.uottawa.ca)
- **BNF5106** Bioinformatics*
- **BCH5101** Analysis of -omics data
- Ottawa Bioinformatics User Group (**OttBUG**)  
  - See [reddit conversation](http://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**.
Course learning outcomes

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

- **List** and **describe** the fundamental algorithms in bioinformatics
- **Articulate** the trade-offs behind algorithms in bioinformatics
- **Write** computer programs for solving large scale bioinformatics problems
- Critically **review** scientific publications in this field
- **Locate** and critically **evaluate** scientific information
- **Apply** one of the paradigms presented in class to solve real-world problems
- **Present** scientific content to a small technical audience
Outline

- Essential cell biology
- Suffix trees, lowest common ancestor
- Suffix trees applications
- Molecular sequence alignment
- Students presentation (review paper)
- Phylogeny
- RNA secondary structure
- Sequence motifs (deterministic and probabilistic)
- Students presentation (final project)
Evaluation

- Programming assignments (20%)
- Review and oral presentation of a scientific publication (10%)
- Midterm examination (20%)
- Project (50%) - (proposal 10%, presentation 10%, report 40%)
The following textbooks can be downloaded freely as PDF (access restricted to uOttawa IP addresses).

Other excellent textbooks.


  - QA 76.9 .A43 G87 1997

  - http://bioinformaticsalgorithms.com
If you want to compete in bioinformatics, first you need to compete for really smart people. You need really smart people who understand how to manipulate nanomolecules.

Juan Enriquez
**Jobs:** [http://www.bioinformatics.ca/jobs](http://www.bioinformatics.ca/jobs)

**DISCLAIMER:** OICR and the Canadian Bioinformatics Workshops are not affiliated with and have not investigated the companies listing jobs on this site. OICR is not making any representations with respect to the positions and is not acting as an agent for the companies listed. The OICR and CBW reserve the right to select and edit job postings which are added to this site to ensure Canadians are eligible to apply and that positions are bioinformatics related.

**HOW TO POST A JOB:** Job posting is only available to Bioinformatics.ca members. To become a member please sign up for an account [here](http://www.bioinformatics.ca/jobs). Job postings are moderated and must be approved before posting becomes public.

For full-time positions outside of Canada, please include the statement: “Applications from Canadian citizens are welcome”.

<table>
<thead>
<tr>
<th>Job Title</th>
<th>Institution/Company</th>
<th>Location</th>
<th>Date Posted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hong Kong PhD Fellowship Scheme 2017/18</td>
<td>City University of Hong Kong</td>
<td>Hong Kong</td>
<td>2016-08-31 06:29</td>
</tr>
<tr>
<td>Post Doctoral Position in Tuberculosis</td>
<td>University of British Columbia</td>
<td>Vancouver, BC</td>
<td>2016-08-30 18:36</td>
</tr>
<tr>
<td>Postdoctoral fellowship in Computational Biology</td>
<td>University of British Columbia</td>
<td>Vancouver, BC, Canada</td>
<td>2016-08-30 08:08</td>
</tr>
<tr>
<td>Post-doctoral Fellow</td>
<td>Ontario Institute for Cancer Research</td>
<td>Toronto, ON</td>
<td>2016-08-29 13:48</td>
</tr>
<tr>
<td>Postdoctoral Researcher - Community Genome Database Development</td>
<td>University of Tennessee, Knoxville</td>
<td>Knoxville, TN, United States</td>
<td>2016-08-26 14:08</td>
</tr>
<tr>
<td>Research Associate</td>
<td>University of Saskatchewan</td>
<td>Saskatoon, SK</td>
<td>2016-08-23 16:40</td>
</tr>
<tr>
<td>Bioinformatics specialist and Web Database Developer</td>
<td>McGill Centre for Integrative Neuroscience</td>
<td>Montreal, QC</td>
<td>2016-08-16 16:02</td>
</tr>
<tr>
<td>Postdoctoral Fellowship in Bioinformatics</td>
<td>McGill University</td>
<td>Montreal</td>
<td>2016-08-16 12:22</td>
</tr>
<tr>
<td>Post-Doctoral fellow in High Dimension Flow Cytometry Bioinformatics</td>
<td>Simon Fraser University / University of British Columbia</td>
<td>Vancouver, BC</td>
<td>2016-08-15 01:47</td>
</tr>
</tbody>
</table>
The global bioinformatics market, valued at nearly $3.2 billion in 2012, is forecast to grow to nearly $7.5 billion by 2017, according to Wellesley, Mass.-based BCC Research.

*Healthcare IT News, April 30, 2013*

Bioinformatics grows by billions by Bernie Monegain

Market Size

- Bioinformatics Market worth **16.18 Billion** USD by 2021

- Bioinformatics Market Size Worth **US$ 16 Billion** By 2022
What/Who is a bioinformatician?

According to a (dated?) survey on www.bioinformatics.org (540)

- Biology (192) 36%
- Computer Science (133) 25%
- Engineering (72) 13%
- Mathematics (26) 5%
- Physics (20) 4%
- Chemistry (34) 6%
- Other (54) 10%
Professional associations

ISCB — International Society for Computational Biology
(www.iscb.org)
SMB — Society for Mathematical Biology
(www.smb.org)
CSSB — Canadian Society for Systems Biology
(www.sysbiosociety.ca)
Essential Cellular Biology: Molecules

- Deoxyribonucleic acid (DNA)
- Ribonucleic acid (RNA)
- Proteins
Essential Cellular Biology: Deoxyribonucleic acid (DNA)

- DNA is a **polymer**
Essential Cellular Biology: Deoxyribonucleic acid (DNA)

- DNA is a **polymer**
- Can be seen as a **string** over four letters: A, C, G, T
Essential Cellular Biology: Deoxyribonucleic acid (DNA)

- DNA is a **polymer**
- Can be seen as a **string** over four letters: A, C, G, T
- **Code of instructions** for life
DNA is a **polymer**

- Can be seen as a **string** over four letters: **A**, **C**, **G**, **T**
- **Code of instructions** for life
  - **A list of parts** and a **user manual**
Essential Cellular Biology: Deoxyribonucleic acid (DNA)

- DNA is a **polymer**
- Can be seen as a **string** over four letters: A, C, G, T
- **Code of instructions** for life
  - A **list of parts** and a **user manual**
- Each one of your cell has an **identical** copy of your **DNA**
Essential Cellular Biology: Deoxyribonucleic acid (DNA)

- DNA is a **polymer**
- Can be seen as a **string** over four letters: A, C, G, T
- **Code of instructions** for life
  - A **list of parts** and a **user manual**
- Each one of your cell has an **identical** copy of your DNA
- Different regions of your DNA are **active** in each cell
## Essential Cellular Biology: Deoxyribonucleic acid (DNA)

<table>
<thead>
<tr>
<th>Species</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potato spindle tuber viroid (PSTVd)</td>
<td>360</td>
</tr>
<tr>
<td>Human immunodeficiency virus (HIV)</td>
<td>9,700</td>
</tr>
<tr>
<td>Bacteriophage lambda ((\lambda))</td>
<td>48,500</td>
</tr>
<tr>
<td>Mycoplasma genitalium (bacterium)</td>
<td>580,000</td>
</tr>
<tr>
<td>Escherichia coli (bacterium)</td>
<td>4,600,000</td>
</tr>
<tr>
<td>Drosophila melanogaster (fruit fly)</td>
<td>120,000,000</td>
</tr>
<tr>
<td>Homo sapiens (human)</td>
<td>3,000,000,000</td>
</tr>
<tr>
<td>Lilium longiflorum (easter lily)</td>
<td>90,000,000,000</td>
</tr>
<tr>
<td>Amoeba dubia (amoeba)</td>
<td>670,000,000,000</td>
</tr>
</tbody>
</table>
Essential Cellular Biology: Ribonucleic acid (RNA)

RNA is a polymer
Essential Cellular Biology: Ribonucleic acid (RNA)

- RNA is a **polymer**
- Can be seen as a **string** over four letters: A, C, G, U
Essential Cellular Biology: Ribonucleic acid (RNA)

- RNA is a polymer
- Can be seen as a string over four letters: A, C, G, U
- Tens, hundreds, thousands nucleotides (letter) long
Essential Cellular Biology: Ribonucleic acid (RNA)

- RNA is a **polymer**
- Can be seen as a **string** over four letters: A, C, G, U
- Tens, hundreds, thousands nucleotides (letter) long
- Gene transcription and translation, but also regulation, editing, etc.
Essential Cellular Biology: Proteins

🌱 Protein is a polymer
Essential Cellular Biology: Proteins

- Protein is a **polymer**
- Can be seen as a **string** over twenty letters: A, C, D, ... Y
Essential Cellular Biology: Proteins

- Protein is a **polymer**
- Can be seen as a **string** over twenty letters: A, C, D, ... Y
- **Hundreds** or **thousands** amino acids (letter) long
Essential Cellular Biology: Proteins

- Protein is a **polymer**
- Can be seen as a **string** over twenty letters: A, C, D,...Y
- **Hundreds** or **thousands** amino acids (letter) long
- **Catalytic** activity, **transporter** activity, **binding**, etc.
Essential Cellular Biology: Central Dogma

Replication

DNA → Transcription → RNA → Translation → Protein
What is a gene?

- A locatable **region** of genomic sequence [DNA], corresponding to a unit of inheritance, which is associated with **regulatory** regions, **transcribed** regions, and or other **functional** sequence regions.
What is a gene?

- A locatable region of genomic sequence [DNA], corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and/or other functional sequence regions.

PEARSON H.
Genetics: what is a gene?.

Gene Regulation

- Ensemble of mechanisms by which the cell increases or decreases the production of (protein or RNA) genes
Essential Cellular Biology: Gene Regulation

Molecular sequences and structures
Hundreds of ontologies describe the parts and processes
High-throughput experiments
- ChIP-Seq informs about protein-DNA interactions
- DNA microarrays measure the expression levels of genes
- And many more
Essential Cellular Biology: Resources

!xe unlockinglifescocode.org
See also:

https://www.ted.com/talks/james_watson_on_how_he_discovered_dna.
References


Pensez-y!

L’impression de ces notes n’est probablement pas nécessaire!