BIOE 310: Computational Tools for Biological Data

What this class is all about?
Instructor

- Name: Sergei Maslov

- Professor of Bioengineering, Physics, Carl R. Woese Institute for Genomic Biology, and National Center for Supercomputing Applications

- Office: 3406 Carl Woese Institute for Genomic Biology and sometimes 3146C Everitt Laboratory (both by appointment)

- E-mail: maslov@illinois.edu

- Phone: 217-265-5705
Co-Instructor

- Name: Roy Dar

- **Assistant Professor of Bioengineering, Carl R. Woese Institute for Genomic Biology**

- Office: 2240 Everitt Laboratory (by appointment)

- E-mail: roydar@illinois.edu

- Phone: (217) 265-0708
Questions and Suggestions:

maslov@Illinois.edu

roydar@Illinois.edu

Start subject with [BIOE310]
Homework and Exams

- **Homework assignments.** Due at the beginning of the class on the designated day

- **Midterm exam.** Most likely on Tuesday, March 12

- **Final exam.** Date will be decided by the College of Engineering

- **Grade:**
  - Homework 30%
  - Midterm 30%
  - Final 40%

- I will have one grader to help me with grading homework and Roy with grading exams
Course Website

https://courses.engr.illinois.edu/bioe310

Grades will be on
https://my.bioen.illinois.edu/gradebook

BIOE 310 (Spring 2019) - Computational Tools for Biological Data

- Home
- Schedule

<table>
<thead>
<tr>
<th>#</th>
<th>Date</th>
<th>Topics</th>
<th>Slides</th>
<th>Matlab</th>
<th>Homework</th>
<th>Exams</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jan 15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Jan 17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bring your i-clickers to my classes

• Who knows what is i-clicker?

• Show of hands: who has an i-clicker?

• I would like you all to have an i-clicker and bring it to every class. On amazon.com a new i>clicker costs $38. Also sold at UIUC Bookstore. Used are cheaper.

• Your answers WILL NOT be used for grading. I need it to see if I lost you and what do I need to explain better
We will use Matlab in class

• Bring your own laptops.
• Need to have Matlab installed and know the basic interface (inline commands, plotting)

Show of hands: who has Matlab?
  – Using CITRIX for UIUC students
    We will be using: Statistics and Machine Learning Toolbox and Bioinformatics Toolbox

• If you miss it in class or want to refresh your memory .m files and .mat with Matlab commands and data will be on the website
What will you learn in this course?

• Basics of probability and statistics
  – Basic concepts of probability, Bayes theorem
  – Discrete and continuous probability distributions
  – Multivariate statistics
  – Sampling distributions
  – Parameter estimation
  – Hypothesis testing
  – Regression

• How it is applied to biological data
  – Genomics, trees
  – Systems biology (gene expression, networks)
  – Imaging data (in a special unit by Prof. Dar)
The main Probability/Statistics Textbook

*D. C. Montgomery and G. C. Runger*
John Wiley & Sons, Inc. (2011)

You can also use other editions from 4th (2007) to 6th (2014)

5th edition is available for free as a Safari eBook at
https://vufind.carli.illinois.edu/vf-uiu/Record/uiu_8485760
Probability/Statistics for Bioengineering with Matlab exercises

Statistics for Bioengineering Sciences
with MATLAB and WinBUGS Support
Brani Vidakovic
Department of Biomedical Engineering, Georgia Tech
It is constantly updated with the newest version at the link below.

Free as a PDF eBook at
http://statbook.gatech.edu
Matlab exercises and datasets are at
http://springer.bme.gatech.edu
Genomics/Systems Biology Textbook

- J Pevsner
  *Bioinformatics and functional genomics*
  Wiley-Blackwell,
  *2nd edition [2009] exists in electronic form*
  *3rd edition [2015] has up-to-date information on NGS: RECOMMENDED (about $60 on amazon)*

- *2nd edition as PDF in electronic form*  
  [https://vufind.carli.illinois.edu/vf-uiu/Record/uiu_6212694](https://vufind.carli.illinois.edu/vf-uiu/Record/uiu_6212694)
Another Bioinformatics/Statistics Textbook

- *Ewens, WJ and Grant, GR*  

- 2nd edition as PDF eBook  
  [http://vufind.carli.illinois.edu/vf-uiu/Record/uiu_5590643](http://vufind.carli.illinois.edu/vf-uiu/Record/uiu_5590643)
What biological data will be covered?

Will be covered in lectures and/or Matlab class exercises:

- **Genomic data**: 1D strings of letters ACGT (or U for RNA)
- **Proteomic data**: protein sequence, abundances, and interaction strengths
- **Gene Expression data**: messenger RNA copy numbers expressed from 1000s of genes
- **Network data**: up to 1,000,000 pairs of interacting genes or proteins

Will be covered in some Matlab class exercises:

- Biological imaging data: e.g. microscopy, fluorescent molecules

Will not be covered:

- fMRI brain scans, Brain connectome, Ecosystem dynamics data
Why do you need probability and statistics to analyze modern biological data?
Definition of *statistics* by Merriam-Webster

1: a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data ...
Why do you need probability and statistics to analyze modern biological data?

**Reason 1:** Biology now has **Lots of Data**

If data was money: $1 investment in 1985 would bring you $1 billion in 2015
Cost per Megabase of DNA Sequence

Moore's Law

National Human Genome Research Institute

genome.gov/sequencingcosts
Cost per Genome Sequenced

The cost of sequencing a human genome compared with the reductions that would be expected at the rate Moore’s law predicts for computer chips. Over the past decade, next-generation sequencing and cloud computing drove the figure down. The average bumped higher in recent years because of brief slowdowns in production.

Source: NIH
### Base pairs

<table>
<thead>
<tr>
<th>Base pairs</th>
<th>Unit</th>
<th>Abbreviation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 base pair</td>
<td>1 bp</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>1 kilobase pair</td>
<td>1 kb</td>
<td></td>
</tr>
<tr>
<td>1,000,000</td>
<td>1 megabase pair</td>
<td>1 Mb</td>
<td></td>
</tr>
<tr>
<td>$10^9$</td>
<td>1 gigabase pair</td>
<td>1 Gb</td>
<td></td>
</tr>
<tr>
<td>$10^{12}$</td>
<td>1 terabase pair</td>
<td>1 Tb</td>
<td></td>
</tr>
<tr>
<td>$10^{15}$</td>
<td>1 petabase pair</td>
<td>1 Pb</td>
<td></td>
</tr>
</tbody>
</table>

### Size

<table>
<thead>
<tr>
<th>Size</th>
<th>Abbreviation</th>
<th>No. bytes</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bytes</td>
<td>–</td>
<td>1</td>
<td>1 byte is typically 8 bits, used to encode a single character of text</td>
</tr>
<tr>
<td>Kilobytes</td>
<td>1 kb</td>
<td>$10^3$</td>
<td>Size of a text file with up to 1000 characters</td>
</tr>
<tr>
<td>Megabytes</td>
<td>1 MB</td>
<td>$10^6$</td>
<td>Size of a text file with 1 million characters</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>464 TB: Data generated by the 1000 Genomes Project <a href="http://www.1000genomes.org/faq/how-much-disk-space-used-1000-genomes-project">http://www.1000genomes.org/faq/how-much-disk-space-used-1000-genomes-project</a> (WebLink 2.86)</td>
</tr>
<tr>
<td>Petabytes</td>
<td>1 PB</td>
<td>$10^{15}$</td>
<td>1 PB: size of dataset available from The Cancer Genome Atlas (TCGA)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5 PB: size of SRA data available for download from NCBI</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>15 PB: amount of data produced each year at the physics facility CERN (near Geneva) <a href="http://home.web.cern.ch/about/computing">http://home.web.cern.ch/about/computing</a> (WebLink 2.87)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.5 exabytes of data are produced worldwide (Lampitt, 2014)</td>
</tr>
</tbody>
</table>

Who will have *bigger data* by 2025?

<table>
<thead>
<tr>
<th>Data Phase</th>
<th>Astronomy</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acquisition</td>
<td>25 zetta-bytes/year</td>
<td>0.5–15 billion tweets/year</td>
</tr>
<tr>
<td>Storage</td>
<td>1 EB/year</td>
<td>1–17 PB/year</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>YouTube</th>
<th>Genomics</th>
</tr>
</thead>
<tbody>
<tr>
<td>500–900 million hours/year</td>
<td>1 zetta-bases/year</td>
</tr>
<tr>
<td>1–2 EB/year</td>
<td>2–40 EB/year</td>
</tr>
</tbody>
</table>

One more reason to study hard in this class

Glassdoor average salary:
- Data scientist $118,709
- Programmer only $64,537

SOURCE: LinkedIn’s 20 Hottest Skills of 2014 Report
What makes genomic data so big?

- There are ~9 millions species each with its own genome
- Each of us humans (7.5 billions and counting) has a unique DNA: we want to compare them to each other
- Each cell has just 1 genome (DNA) but multitude of transcriptomes (RNA levels) and proteomes (protein levels)
- Cancer cells acquire mutations in their genomes: need to track multiple lineages in a tumor vs time to understand cancer
- DNA was proposed as a long-term storage medium of information
How DNA could store all the world’s data

Modern archiving technology cannot keep up with the growing tsunami of bits. But nature may hold an answer to that problem already.

Andy Extance

31 August 2016
Farfetched? Storage standards evolve fast

Note: Nature started the comparison with hard disk skipping the floppy disk
• Prof Olgica Milenkovic from Electrical and Computer Engineering UIUC is a local expert on this topic

• Profs. George Church and Sri Kosuri (Harvard Medical School) explains a potential use of DNA as storage medium

• https://www.youtube.com/watch?v=IJAdqAVjQqY
Why do you need probability and statistics to analyze modern biological data?

Reason 2: Life is random and messy
Show video
“Cell organelles”

• Made at the Walter and Eliza Hall Institute of Medical Research at Victoria, Australia
• Animated by award-winning artist Dr. Drew Berry
• Go to https://www.wehi.edu.au/wehi-tv for other videos
Life is messy, random, and noisy

Yet it is beautifully complex