BIOE 505: Computational Bioengineering

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
Questions and Suggestions:

maslov@Illinois.edu

Start subject with [BIOE505]
Grading

- Midterm exam 40%
- Final exam 60%
- Homework (ungraded) will be posted online solutions in a week.
- It will build on topics covered in lectures and will consist of problem sets related to topics covered in lectures
- Useful to prepare for exams
Course Website

https://courses.engr.illinois.edu/bioe505

Grades will be on

https://my.bioen.illinois.edu/gradebook

BIOE 505 (Fall 2018) - Computational Bioengineering

Schedule

Instructor

Sergei Maslov: maslov@illinois.edu
Office: IGB 3406
Office hours: by appointment

Logistics

BIOE 505 (Fall 2018) - Computational Bioengineering

Schedule

Tuesdays: 11:00AM -
Thursdays: 11:00AM - [Course homepage]

3217 Everitt Laborator

<table>
<thead>
<tr>
<th>#</th>
<th>Date</th>
<th>Topic</th>
<th>Slides</th>
<th>Matlab</th>
<th>Homework and other materials to contemplate at home</th>
<th>Exams</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Aug 28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Aug 30</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?**

- $20 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 file with Matlab commands will be on the website**
What will you learn in this course?

• Basics of probability and statistics
  – Probability distributions
  – Multivariate statistics
  – Sampling distributions
  – Parameter estimation
  – Hypothesis testing
  – Regressions

• How it is applied to biological data
  – Genomics
  – Systems biology
The main Probability/Statistics Textbook

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

You can also use other editions from 4<sup>th</sup> (2007) to 6<sup>th</sup> (2014)

5<sup>th</sup> 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,
  2\textsuperscript{nd} edition [2009] **exists in electronic form**
  3\textsuperscript{rd} edition [2015] **has up-to-date information on NGS: RECOMMENDED**
  (about $60 on amazon)

- 2\textsuperscript{nd} 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 Biological Data/Statistics Textbook

- Ewens, WJ and Grant, GR

- 2nd edition as PDF eBook
  http://vufind.carli.illinois.edu/vf-uiu/Record/uiu_5590643
What biological data will we cover?

• **Genomic data**: 1D strings of letters ACGT (or U for RNA)
• **Proteomic data**: protein sequences (strings of 20 Amino Acids) and their abundances
• **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

**We will not cover:**

• Biological imaging data: e.g. microscopy, single fluorescent molecules, fMRI brain scans
• Brain, 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

National Human Genome Research Institute

genome.gov/sequencingcosts
<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>A, C, G, T = 2 bits</td>
</tr>
<tr>
<td>1000</td>
<td>1 kilobase pair</td>
<td>1 kb</td>
<td>Size of a typical coding region of a gene</td>
</tr>
<tr>
<td>1,000,000</td>
<td>1 megabase pair</td>
<td>1 Mb</td>
<td>Size of a typical bacterial genome</td>
</tr>
<tr>
<td>$10^9$</td>
<td>1 gigabase pair</td>
<td>1 Gb</td>
<td>The human genome is 3 billion base pairs</td>
</tr>
<tr>
<td>$10^{12}$</td>
<td>1 terabase pair</td>
<td>1 Tb</td>
<td>300 Human Genomes</td>
</tr>
<tr>
<td>$10^{15}$</td>
<td>1 petabase pair</td>
<td>1 Pb</td>
<td>300,000 Human Genomes</td>
</tr>
</tbody>
</table>

<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>
</tbody>
</table>
| Gigabytes | 1 GB         | $10^9$    | 600 GB: size of GenBank (uncompressed flat files)  
|           |              |           | 464 TB: Data generated by the 1000 Genomes Project (http://www.1000genomes.org/faq/how-much-disk-space-used-1000-genomes-project) (WebLink 2.86)                                                                 |
| Petabytes | 1 PB         | $10^{15}$ | 1 PB: size of dataset available from The Cancer Genome Atlas (TCGA)                                                                                                                                  |
|           |              |           | 5 PB: size of SRA data available for download from NCBI                                                                      |
|           |              |           | 15 PB: amount of data produced each year at the physics facility CERN (near Geneva) (http://home.web.cern.ch/about/computing) (WebLink 2.87)                                                        |
| Exabytes  | 1 EB         | $10^{18}$ | 2.5 exabytes of data are produced worldwide (Lampitt, 2014)                                                                                                                                          |
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

**STORAGE LIMITS**

Estimates based on bacterial genetics suggest that digital DNA could one day rival or exceed today's storage technology.

<table>
<thead>
<tr>
<th></th>
<th>Hard disk (μs per bit)</th>
<th>Flash memory</th>
<th>Bacterial DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Read–write speed</strong></td>
<td>~3,000–5,000</td>
<td>~100</td>
<td>&lt;100</td>
</tr>
<tr>
<td><strong>Data retention</strong></td>
<td>&gt;10</td>
<td>&gt;10</td>
<td>&gt;100</td>
</tr>
<tr>
<td><strong>Power usage</strong></td>
<td>~0.04</td>
<td>~0.01–0.04</td>
<td>&lt;10⁻¹⁰</td>
</tr>
<tr>
<td><strong>Data density</strong></td>
<td>~10¹³</td>
<td>~10¹⁶</td>
<td>~10¹⁹</td>
</tr>
</tbody>
</table>

WEIGHT OF DNA NEEDED TO STORE WORLD'S DATA
~1 kg
• 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
Why life is so random?

• Biomolecules are very small (nano- to micro-meters) → Brownian noise
• # molecules/cell is often small → Large cell-to-cell variations
• Genomic data comes from biological evolution – the Mother of all random processes
• Genomic data involves (random) samples – We have genomes of some (not all) organisms – We have tissue samples of some (not all) cancer patients
Why life is so complex?

Primer on complex system
Complex systems have many interacting parts

• All parts are different from each other
  – 10s thousands ($10^4$) types of proteins in an organism
  – 100 thousands ($10^5$) organizations (AS) in the Internet
  – 1 billion ($10^9$) people on Facebook
  – 10 billion ($10^{10}$) web pages in the WWW
  – 100 billion ($10^{11}$) neurons in a human brain
  – NOT $10^{23}$ electrons or quarks studied in physics: they are all the same!

• Yet they share the same basic design
  – All proteins are strings of the same 20 amino acids
  – All WWW pages use HTML, JavaScript, etc.
  – All neurons are specialized cells generating and receiving electric spikes
Example: a complex system with many parts

Justin Pollard,
http://www.designboom.com
Parts interact →
they need to be assembled to work
Intra-cellular Networks operate on multiple levels

Slides by Amitabh Sharma, PhD
Northeastern University & Dana Farber Cancer Institute
Sea urchin embryonic development (from endomesoderm up to 30 hours) by Davidson’s lab
Protein-Protein binding
IntAct Database (Dec 2015)
Interactions: 577,297  Proteins: 89,716

Baker’s yeast *S. cerevisiae* (only nuclear proteins shown)
From S. Maslov, K. Sneppen, Science 2002

Worm *C. elegans*
From S. Lee et al, Science 2004
Brain and nerves of a worm

• Worm (C. elegans) has 302 neurons

• Our brain has 100 billion ($10^{11}$) neurons
by Neo Martinez and Richard Williams

Food Web of Little Rock Lake

Smallmouth Bass (Cannibal)

1st Tropic Level
Mostly Phytoplankton

2nd Tropic Level
Many Zooplankton
Networks in technology & engineering
153 Linux packages required for installation of Firefox

TY Pang, S. Maslov, PNAS (2013)
Network of highways has no hubs
Airline networks have hubs
Physical backbone of the Internet