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 will be posted in a week.

• Homework 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 - Computational Bioengineering

Schedule

Instructor

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

Logistics

Tuesdays: 12:00AM - 1:50AM
Thursdays: 12:00AM - 1:50PM
106RR Engineering Hall

I WANT YOU TO BRING YOUR i-CLII
I WANT YOU TO BRING YOUR OWN
LAPTOPS SHOULD HAVE MATLAB
STATISTICS AND MACHINE LEARN

Description

<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>Aug 27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Aug 29</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 laptops to class
• Poll: who has Matlab?
• Need to have Matlab installed and know the basic user interface (inline commands, plotting)
• We will use Statistics and Machine Learning Toolbox and Bioinformatics Toolboxes
• You can use CITRIX for UIUC students and connect to EWS Windows Lab Software
• .m files and .mat with Matlab commands and data will be on the website after the lecture
Who has Matlab?

A. Have it installed on my own laptop
B. Plan to use CITRIX
C. I don’t know yet
D. I plan to use Python or other language
   (cannot use my templates then)

Get your i-clickers
Possible alternative to purchasing Matlab and toolboxes is to use campus resources.

Both Engineering Workstations (EWS) and ACES computers have Matlab. I don’t think all of them offer the statistics and bioinformatics toolboxes (EWS should, ACES computers may not..).

See the following to access:

**Citrix for EWS, Matlab, and ACES computers -- links for all**
https://it.engineering.illinois.edu/ews/lab-information/remote-connections/connecting-citrix
https://it.engineering.illinois.edu/services/instructional-services/remote-connections-citrix

**Accessing Engineering Workstations (EWS)**
https://it.engineering.illinois.edu/ews

**Accessing ACES Academic Computing Workstations**
http://acf.aces.illinois.edu/remote/
http://acf.aces.illinois.edu/remote/pc.html

To access off campus use:
**CISCO Virtual Private Network -- For off-campus access to campus computer and network resources (software programs, files saved on the network, etc.)**
https://techservices.illinois.edu/services/virtual-private-networking- vpn/download-and-set-up-the-vpn-client

**CISCO VPN CLIENT**
https://webstore.illinois.edu/shop/product.aspx?zpid=2600

**CISCO AnyConnect VPN**
https://webstore.illinois.edu/shop/product.aspx?zpid=1222
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 (genome assembly)
  – Systems biology (gene expression, PPI networks)
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://i-share-uiu.primo.exlibrisgroup.com/permalink/01CARLI_UIU/gpjosq/alma99943707412205899
Problems for our main Probability/Statistics Textbook

D. C. Montgomery and G. C. Runger
John Wiley & Sons, Inc. (2010)

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

5th edition is available for free as a Safari eBook at https://i-share-uiu.primo.exlibrisgroup.com/permalink/01CARLI_UIU/gpjosq/alma99947443312205899
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/statb4.pdf
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://i-share-uiu.primo.exlibrisgroup.com/permalink/01CARLI_UIU/gpjosq/alma99621269412205899](https://i-share-uiu.primo.exlibrisgroup.com/permalink/01CARLI_UIU/gpjosq/alma99621269412205899)
Another Bioinformatics/Statistics Textbook


- 2nd edition as PDF eBook
  https://i-share-uiu.primo.exlibrisgroup.com/permalink/01CARLI_UIU/gpjosq/alma99784955712205899
Why are there ants in my laptop?

Why are there so many crows in Rochester, MN?

Why are there so many avengers?

Why is there ice in space?

Why is there a line through HTTPS?

Why aren't my arms growing?

Why aren't my quail eggs hatching?

Why is there a red line through HTTPS on Facebook?

Why is there a red line through HTTPS on Facebook?

Why is HTTPS important?
This course is about **biological data** and **probability theory** and **statistics** concepts needed for its analysis.
What biological data will be discussed?

**Will be covered in lectures or Matlab exercises:**

- **Genomic data**: strings of letters ACGT
- **Gene Expression data**: messenger RNA copy numbers transcribed from genes
- **Proteomic data**: protein abundances
- **Network data**: pairs of interacting genes or proteins and protein-protein interaction strengths

**Will not be covered:**

- Imaging data such as e.g. fMRI brain scans, Brain connectome data, Ecosystem dynamics data
Why do you need probability and statistics to analyze modern biological data?
Definition of **probability theory** by Encyclopedia Britannica

a branch of mathematics concerned with the analysis of **random** phenomena

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 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
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>

Peta=$10^{15}$  
Exa=$10^{18}$  
Zetta=$10^{21}$

<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>

### Base Pairs Table

<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>

A, C, G, T = 2 bits = 0.25 bytes

---

### Size 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>
<tr>
<td>Gigabytes</td>
<td>1 GB</td>
<td>$10^9$</td>
<td>600 GB: size of GenBank (uncompressed flat files)</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>Exabytes</td>
<td>1 EB</td>
<td>$10^{18}$</td>
<td>2.5 exabytes of data are produced worldwide (Lampttt, 2014)</td>
</tr>
</tbody>
</table>
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 unique DNA: we want to compare them all 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 may hold an answer to that problem.

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</th>
<th>Flash memory</th>
<th>Bacterial DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read–write speed (µs per bit)</td>
<td>~3,000–5,000</td>
<td>~100</td>
<td>&lt;100</td>
</tr>
<tr>
<td>Data retention (years)</td>
<td>&gt;10</td>
<td>&gt;10</td>
<td>&gt;100</td>
</tr>
<tr>
<td>Power usage (watts per gigabyte)</td>
<td>~0.04</td>
<td>~0.01–0.04</td>
<td>&lt;10–10</td>
</tr>
<tr>
<td>Data density (bits per cm³)</td>
<td>~10^{13}</td>
<td>~10^{16}</td>
<td>~10^{19}</td>
</tr>
</tbody>
</table>

WEIGHT OF DNA NEEDED TO STORE WORLD’S DATA

~1 kg
Farfetched? Storage standards evolve fast but DNA standard remained unchanged for 4 billion years

Note: Nature article started the comparison with a hard drive and flash memory 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?

Reason 1.5: It is necessary to rationally manage the COVID-19 epidemic
I was one of 28 people who received the University of Illinois Presidential Award and Medallion

Past recipients include:
- William Daley, 2011
  chief of staff to President Barack Obama
- Thomas M. Siebel, 2001
  Founder, chairman and CEO of Siebel Systems, Inc., and a major donor
- Arthur C. Clarke, 1997
  Science fiction author of "2001: A Space Odyssey"
- David W. Grainger, 1989
  Friend of the university and generous donor
During March and April 2020, testing was inadequate to track COVID-19 in Illinois ...
During March and April 2020, testing was inadequate to track COVID-19 in Illinois ...

... The only way to see the disease was through mathematics and modeling
Even through the number of reported cases was small, we could predict the dramatic effect of a pre-emptive stay-at-home order.
Window of Opportunity for Mitigation to Prevent Overflow of ICU capacity in Chicago by COVID-19

Sergei Maslov and Nigel Goldenfeld

Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign

March 18 2020

Please note: this is a working document and has not been submitted for journal publication. It is planned that a later version of this document will be submitted for peer-reviewed publication, but in the interests of sharing information during a rapidly changing epidemic landscape, we are making this early version available.

Executive Summary

We estimate the growth in demand for ICU beds in Chicago during the emerging COVID-19 epidemic, using state-of-the-art computer simulations calibrated for the SARS-CoV-2 virus. The questions we address are these:

(1) Will the ICU capacity in Chicago be exceeded, and if so by how much?

(2) Can strong mitigation strategies, such as lockdown or shelter in place order, prevent the overflow of capacity?

(3) When should such strategies be implemented?

https://www.medrxiv.org/content/10.1101/2020.03.20.20040048v1.full.pdf+html
Wuhan-level implemented at different times

Early mitigation makes a big difference!
What does it have to do with probability and statistics?

• I was a member of the COVID-19 modeling taskforce providing the Governor of Illinois with weekly updates:
  – Model: **probabilistic time-of-infection model** using probability distributions of various delays infection-> symptoms -> testing -> hospitalizations -> ICU -> deaths

• I was a member of the modeling group for UIUC advising management on mitigation strategies, testing, etc.
  – Method: **Agent-Based Model (ABM)** probabilistically describing the spread of the epidemic on campus
  – 45,000 students/faculty/staff assigned to zones, etc.
SEEING THE COVID-19 EPIDEMIC WITH MATHEMATICS AND COMPUTER SIMULATIONS

Sergei Maslov (UIUC Bioengineering)
Nigel Goldenfeld (UIUC Physics)

COVID-19 MODELLING GROUP

Ahmed Elbanna (Civil Eng.), Zach Weiner (Physics), George Wong (Physics), Tong Wang (Physics),

Alexei Tkachenko

Brookhaven National Laboratory
Hospitals at Risk: Sothern Illinois, Region 4

50% risk of exceeding ICU COVID availability at the beginning of October 2020
Time-of-infection model

• Model calibration
  – The equations of the epidemic need to know what the starting condition is
  – The equations of the epidemic need certain parameters that describe the disease transmission
  – One needs good data to estimate parameters. We get ours from IDPH through Data Use Agreement
  – We use a sophisticated process of estimating these parameters using a Bayesian inference algorithm known as Markov Chain Monte Carlo
  – We need supercomputers to do this, because we have to try many combinations of parameters and see which have the highest statistical support

• Forward simulation
  – Once these parameters are known, together with their ranges, we can run the equations forward in time.
  – We estimate uncertainties due to fitting of parameters
All State and Chicago

Illinois, Baseline

Region 11, Baseline
Central regions
Agent Based Model of COVID-19 epidemic at our University

• Special feature of the University is that students and professors come together at specific times and places for classes
  – Agents (40,000 or more = students + workers)
  – Zones (classrooms, bars, restaurants, dorms, coffee shops, library, parties)

• Data input: we constructed the network of students and classroom zones from anonymized data of all students at UIUC in Fall 2019
Mandatory fuzzball image

- Nodes are students
- Edges are students
- Small-world network with ~2.5 “degrees of separation”
  - Students cluster by major, common classes
- Information spreads rapidly in network
- Social bubbles are connected by social activities outside of class scheduled time
Background Transmission model

- Each infected person produces “quanta” of virus, depending on purpose of room and agent’s activity, e.g.
  - Lecture: passive = baseline
  - Lecturer: 5 X baseline
  - Restaurant: 5X baseline
  - Bar: 50 X baseline

- Concentration of quanta depends on room volume (use literature for typical values)

- Quanta advected by air-currents lead to a background level which can decay or remain airborne depending on droplet size and ventilation efficiency

- Ventilation capacity is zone-dependent
  - Classroom: air exchange rate = 3-4 X per hour
  - Bar: AER = 15 X classroom
  - Restaurant: AER = 10 X

Background Transmission model

- Susceptible agents inhale the quanta and accumulate them
- Infection risk is computed from time of exposure to accumulated quanta
- Quanta decay in body if not infected, once you leave the zone
- New quanta start to accumulate when agent enters new zone with infected agents
- Quanta remain in room with half-life (dependent on airflow) ~ 15-30 minutes

Mitigation including the return to campus

- Masks in university bldgs.
- Twice-weekly testing
- No in-person class > 50
- Contact tracing app
- Social life in bars/restaurants allowed
Second shameless bragging slide: Nigel and I inspired an **XKCD cartoon**

<table>
<thead>
<tr>
<th>Frame 1</th>
<th>Frame 2</th>
<th>Frame 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apparently some university reopened based on a COVID model developed by two physicists.</td>
<td>But even their worst-case model underestimated the number of student parties and they had to shut down.</td>
<td>Can't understand why someone with a physics degree would be bad at judging how often college students get invited to parties.</td>
</tr>
<tr>
<td>Uh oh.</td>
<td></td>
<td>Excuse me, I was invited to multiple parties, and attended both of them!</td>
</tr>
</tbody>
</table>
UIUC: ~2.2 million tests!

https://go.illinois.edu/COVIDTestingData
New variants like Delta are more transmissible.

http://91-divoc.com/pages/covid-visualization/
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 and has many parts
(see statistics)
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 \textit{proteins} in an organism
  – 100 thousands \((10^5)\) \textit{organizations (AS)} in the Internet
  – 1 billion \((10^9)\) people on Facebook
  – 10 billion \((10^{10})\) \textit{web pages} in the WWW
  – 100 billion \((10^{11})\) \textit{neurons} in a human brain
  – NOT \(10^{23}\) electrons or quarks studied by physics: they are all the same and boring!

• Yet they share the same basic design
  – All proteins are strings of the same 20 amino acids
  – All WWW pages use \textit{HTML}, JavaScript, etc.
  – All neurons generate and receive \textit{electric spikes}
Example: a complex system with many parts

Justin Pollard,
http://www.designboom.com
Parts interact →

they need to be assembled to work

1

2

30x

Justin Pollard,
http://www.designboom.com