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
Teaching Assistant:
Ananthan Nambiar
nambiar4@illinois.edu
Questions and Suggestions:

maslov@Illinois.edu
nambiar4@illinois.edu

Start subject with [BIOE310]
Homework and Exams

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

- **Midterm exam.** March either before or after the spring break

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

- **Grading:**
  - Homework 30%
  - Midterm 30%
  - Final 40%
Course Website

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

Grades will be on

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

### BIOE 310 - Computational Tools for Biological Data

Return to syllabus

<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 26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Jan 28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bring your iClickers to my lectures

• Who knows what is an iClicker?

• Show of hands: who has an iClicker?

• I would like you all to have an iClicker and bring it to every class. On amazon.com a new iClicker (1st generation is OK) costs around $40. It is also sold at UIUC Bookstore. The used ones are cheaper.

• An alternative solution is using a mobile app: https://www.iclicker.com/students/apps-and-remotes/apps

• Your answers WILL NOT be used for grading. I need them to see if I lost some of you and what could I rephrase to better explain the material
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 on my own laptop

B. Plan to use CITRIX

C. Plan to install on my laptop
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
  – Basics of genomics
  – Systems biology (gene expression, 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 at our library
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 at our library
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,

  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 is available for free in electronic form in our library
Another Bioinformatics/Statistics Textbook


- 2nd edition as PDF eBook
QUESTIONs
FONd IN GOOGLE AUnTOCoMPLETe

WHY ARE THERE SLAVES IN THE BIBLE
WHY ARE THERE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS IMPORTANT
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS CROSSED OUT IN RED
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY ARE THERE SO MANY AVENGERS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK-47S SO EXPENSIVE
WHY DO THE SPOOKS WHY ARE THERE GODS
WHY ARE THERE TWO SPOOKS
WHY DO CARS DRIVING ON THE WALL
WHY IS THERE AN OWL IN MY BACKYARD
WHY DO CARS DRIVING ON THE WALL
WHY ARE THERE two SPOOKS
WHY ARE CIGARETTES LEGAL
WHY ARE THERE GLACIERS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LION IN MY EAR
WHY ARE THERE BULLETS SHARP
WHY ARE THERE ANY FOREIGN MILITARY BASES IN AMERICA

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SQUIRRELS
WHY ARE THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNakes EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON PINE'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE SQUIRRELS
WHY DO DYING PEOPLE REACT
WHY AREN'T THERE VACCINE ATTRACTIONS
WHY ARE OUR HUMANS DIFFERENT
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E-GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGH
WHY IS LYING GOOD
WHY IS SEX IMPORTANT
WHY IS LIFE SO BORING
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
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>

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

<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>

<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)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(<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 (Lampitt, 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
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](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 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 **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 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 **HTML, JavaScript, etc.**
  - All neurons generate and receive **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
Metabolic pathway chart by ExPASy: 5702 reactions as of December 2015
Brain and nerves of a worm

- Worm (C. elegans) has 302 neurons
- Our brain has 100 billion \((10^{11})\) neurons
Food Web of Little Rock Lake

1st Tropic Level
Mostly Phytoplankton

2nd Tropic Level
Many Zooplankton

Leech

Smallmouth Bass (Cannibal)
Credit: XKCD comics