Computational Biology

Instructor: Prof. Michael Q. Zhang
(Co-instructor: Dr. Pradipta Ray)

BIOL 6385 / BMEN 6389
Spring (Jan. 9 – May. 1) 2018,
The University of Texas at Dallas
What the course teaches

• **Computational and statistical methods** for analyzing biological data and understanding the biological systems.

• Introduces computational aspects of:
  – Genomics
  – Evolution & phylogenetics
  – Gene regulation & gene networks

• **Focus on generic methods** and algorithms, NOT on specific protocols or tools
Course resources

• Instructors (contact details on website):

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Zhang</td>
<td>Pradipta Ray</td>
<td>Peng Xie</td>
</tr>
<tr>
<td>Instructor</td>
<td>Associate Instructor</td>
<td>Teaching Assistant</td>
</tr>
</tbody>
</table>
Course resources

• Mailing list:
  – biol6385@googlegroups.com

• Please email the instructors your convenient email (UTD email preferred) to join.

• This is a broadcast email list
  – Only instructors post
  – For students, it is best to directly email the instructors (email early, not late)
  – **Email is the preferred mode** of communication
Course resources

- Website: home page (dates, contacts, hours, news): [http://utdallas.edu/~prr105020/biol6385/](http://utdallas.edu/~prr105020/biol6385/)

---

**Computational Biology**

Course No: **BIOL6385.001.18S / BMEN6389.001.18S**

Spring 2018

Biology Department, The University of Texas at Dallas

Instructor: **Michael Zhang**

Director, Center for Systems Biology, The University of Texas at Dallas

Associate Instructor: **Pradipta Ray**

Research Scientist, Center for Systems Biology, The University of Texas at Dallas

Please consult this page for instructors' offices and hours, class/tutorial rooms and hours, and current announcements.

**Important details**

- **Class structure**: Jan 9 - May 1 (31 Units)
- **Class Date / Time**: Tue and Thu 2:30-3:45pm (see [calendar](#))
- **Class Location**: Building: **CB** Room: **1.214**
- **Recitation Date / Time**: To be announced
- **Recitation Location**: NSERL 4.420
- **Midterm Exam Date / Time**: Mar 1, Class hours
- **Mid Term Location**: In class
- **Final Exam Date / Time**: May 1, 2:30-4pm
- **End Term Location**: Normal classroom

**Instructors:**

- **Michael Q. Zhang**, (email: [Click for email](#)) NSERL 4.742, 972-883-2523, Office hours: By Appointment
- **Co-instructor**: **Pradipta Ray**, (email: [Click for email](#)) BSB 10.806, 972-883-7262, Office hours: By Appointment
- **Teaching Assistant**: Peng Xie, (email: pxx140230@utdallas.edu), NSERL 4.716 cubicle complex, 972-883-2528, Office

If stopped by security at NSERL, please call 972-883-2528 to have someone come and fetch you.
Course resources

• Website: course info tab (course policy)
Course resources

- Website: schedule tab (schedule, handouts, HW, solns)

Computational Biology
Course No: BIOI6385.001.18S / BMEN6389.001.18S
Spring 2018
Biology Department, The University of Texas at Dallas

Instructor: Michael Zhang
Director, Center for Systems Biology, The University of Texas at Dallas
Associate instructor: Pradipta Ray
Research Scientist, Center for Systems Biology, The University of Texas at Dallas

Consult this page for class, recitation and exam dates, handouts, and solutions. A printable version of course policy and syllabus is here. Updates to this document will be made on this website.

Course Schedule

Last day to drop the course without a "W" grade is Jan 24. Last day to drop a graduate course in any way is Mar 26. See the academic calendar for details.

<table>
<thead>
<tr>
<th>Lecture</th>
<th>Topic</th>
<th>Handouts / readings</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit 1 Class 1</td>
<td>Introduction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unit 1 Class 2</td>
<td>Introduction (cont'd): Probability Theory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unit 1</td>
<td></td>
<td>Slides</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Slides</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Slides</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Slides</td>
<td></td>
</tr>
<tr>
<td>Unit 1</td>
<td></td>
<td>Slides</td>
<td></td>
</tr>
</tbody>
</table>

Unit 1: Background and Statistical Inference

- CF gene discovery original paper, hosted at UNC here
- Mathematical Writing
  - See R tutorial here, PERL tutorials here, Python tutorials here, and MATLAB tutorial here.
  - A database of open source machine learning tools is at mloss.org, here.
  - The not-so-short introduction to LATEX.
Course policy

• Attendance and participation:
  – Active participation in classroom discussion is expected. **Attendance is mandatory except special permission from the instructor.**
Grading

• Grading:
  – midterm and final exams, and 3 problem sets
  – HWs (50%)
  – Midterm (25%)
  – Final exam (25%)
  – This is a graduate course: don’t focus on grades: the goal is to understand the subject matter!
  – Final letter grades will depend on clustering and relative quantile profiles, not on direct translation of numerical grades.
Examinations

• Exams
  – 75 minutes in duration.
  – open book and open notes. No Computers or communication devices allowed.
  – Mid term exam date: March 1, class hours, in class
  – Final exam date: May 1, class hours, in class
  – It is impossible for us to accommodate individual requests to reschedule the exams.
Homework

• To be done individually. Show all intermediate steps.

• Late homework: Homework is worth full credit at the beginning of class on the due date. It is worth 75% for the next 24 hours, 50% credit from 24 to 96 hours after the due date, 0% credit after that.

• Turn in all 3 HWs, even if for no credit, to pass the course. Late HW assignments must be turned in to the instructors.
• For how to access online, or from a library near you, check the class website.
Reference books

- For how to access online, or from a library near you, check the class website.

http://work.caltech.edu/lectures.htm
5 sections
• Unit 1: Modelling Uncertainty in Biology

– How to build a framework to rationally deal with uncertainty: probability (deductive/logical inference)
– How to estimate and infer parameters associated with such uncertainty: statistics (inductive learning, generalization/abstraction)
– How to proceed when there are many sources of uncertainty in a BIG DATA: bayes nets / deep neural networks
5 sections

• **Unit 2: Molecular Sequence Analysis**
  – Searching and **alignment** of sequences
  – Modelling composition of sequences and guessing their functionality: classification of subsequences and **annotation**
  – **Integrative analysis**: how to combine evidence from multiple and extra-sequential sources when analyzing sequences
5 sections

• Unit 3: Markovian models
  – Markov chains: The Markov condition among random variables, factoring the joint
  – Hidden Markov Models: What happens when the state of the system is unobserved?
  – Supervised and unsupervised inference: Forward-Backward type of algorithms, Baum-Welch / Expectation Maximization algorithm
  – Pair and profile HMMs: Engineering Markovian models to solve computational biological problems
5 sections

- **Unit 4: Evolution & Comparative Genomics**
  - **Evolutionary dynamics**: how DNA may change by mutations
  - **Multiple sequence alignment**: comparing sequences across individuals or species
  - **Phylogenetic trees**: clustering based on sequences, explicitly modelling evolution of sequences
5 sections

- **Unit 5: Generic Machine Learning Approaches for Comp Biologists**
  - Optimization techniques: greedy and more systematic optimization strategies
  - Markov Chain Monte Carlo: Algorithms to sample from probability distributions
  - Classification: identifying classes of observation, category prediction
  - Regression: estimating quantitative relationships among multiple variables, forecasting
  - Structure learning: how to learn the structure of data
  - Ensemble learning: combining learning machines
  - Neural Networks and Deep Learning: feature free learning
What’s computational biology?

**Bioinformatics** applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful. **Computational biology** uses mathematical models and computational approaches to address theoretical and experimental questions in biology. Although bioinformatics and computational biology are distinct, there is also significant overlap and activity at their interface.[1]

Learning: Information ➔ Knowledge, but what’s more important than Knowledge?
"Information is any difference that makes a difference."
Shannon/Turing/Bateson

"It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."
Watson&Crick

Digital revolution
“The new paradigm now emerging, is that all the genes will be known (in the sense of being resident in databases available electronically), and that the starting point of a biological investigation will be theoretical”

W. Gilbert (1991)
Gene finding and structure/function prediction
(Sequence → Structure → Function)

A typical vertebrate gene

DNA

mRNA

Splicing

Some sizes of human genes

<table>
<thead>
<tr>
<th>Name</th>
<th>Size (kb)</th>
<th>mRNA (kb)</th>
<th>Introns</th>
</tr>
</thead>
<tbody>
<tr>
<td>β-Globin</td>
<td>1.5</td>
<td>0.6</td>
<td>2</td>
</tr>
<tr>
<td>Insulin</td>
<td>1.7</td>
<td>0.4</td>
<td>2</td>
</tr>
<tr>
<td>Protein kinase C</td>
<td>11</td>
<td>1.4</td>
<td>7</td>
</tr>
<tr>
<td>Albumin</td>
<td>25</td>
<td>2.1</td>
<td>14</td>
</tr>
<tr>
<td>Catalase</td>
<td>34</td>
<td>1.6</td>
<td>12</td>
</tr>
<tr>
<td>LDL receptor</td>
<td>45</td>
<td>5.5</td>
<td>17</td>
</tr>
<tr>
<td>Factor VIII</td>
<td>186</td>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>Thyroglobulin</td>
<td>300</td>
<td>8.7</td>
<td>36</td>
</tr>
<tr>
<td>Dystrophin</td>
<td>&gt; 2000</td>
<td>17</td>
<td>&gt; 50</td>
</tr>
</tbody>
</table>

Example: alternative splicing of the fly sex determination gene
Positional cloning:
- Linkage analysis
- Physical mapping
- cDNA selection
- Sequencing
- Database search (alignment)
Single gene regulation

gene regulatory sequences

spacer DNA

(enhancer)

CTCF

(insulator/boundary)

gene regulatory proteins

general transcription factors

RNA polymerase

TATA box

(start of transcription)

upstream

(promoter)
GRN: Respiration Module

*(Segal et al., Nature Genetics ’03)*

- Module genes known targets of predicted regulators?

![Diagram showing regulatory network with the text: Hap4+Msn4 known to regulate module genes.](image)
Start filling in the gaps with your DNA

"Because I had given my doctor information from 23andme, he got to a diagnosis much faster. 23andme saved my life." Kirk C.

$99*  Our new low price for all!  Was $199

Order Now »

*Limits apply. New customers only. 23andme Personal Genome Service® at $6/month. Order for $399 without commitment.

Gain insight into your traits, from baldness to muscle performance. Discover risk factors for 97 diseases. Know your predicted response to drugs, from blood thinners to coffee. And uncover your ancestral origins. start tour »

Overview

Discover Health & Ancestry

Keep Your Doctor Informed

Participate In Research

LATEST SUBSCRIPTION UPDATES

NEW links for Hypothyroidism
Nov 03, 2011

UPDATED links for Coronary Heart Disease: Preliminary Research
Oct 27, 2011

NEW links for Scoliosis
Oct 27, 2011

23andMe RESEARCH FINDINGS

Don't it make your brown eyes blue
Oct 2011

Genes, geography, and socio-cultural traits
Oct 2011

Ancestry and disease risk
Oct 2011

WHAT PEOPLE ARE SAYING

kaihendry Exploring my @23andMe Genetic Profile. Having fun. :)

MistaBell I just met a 3rd cousin on 23andme found through DNA analysis then confirmed with our family trees. Amazing technology.
DNA predict:
- Face
- Age
- Behavior
- Make new species
- etc.

Yeast (Saccharomyces cerevisiae, 12.1Mb, 2017)

Microbial genome (Mycoplasma genitalium; 582,970 bp) (2008)
The Omics-cascade, nature is unity

What can happen

Environment

What appears to be happening

Bioinformatics

What makes it happen

Comp. Biol.

What actually happens

GENOME

TRANSCRIPTOME

PROTEOME

SYSTEMS BIOLOGY

METABOLOMOME

PHENOTYPE

WHY WE CARE!

What can be more interesting than understanding ourselves?
Two levels of modeling

- Statistical (Macroscopic) and Population models
  - Simple correlation: \( Y \sim X \)
  - Probabilistic/Predictive:
    - \( P(Y,X), \ P(Y|X) \)
    - \( \bar{Y} = f(x, \alpha) = E[Y|X] = \sum y \ P(Y=y \mid X=x) \)
      e.g. \( f = a \ x + b \) (linear regression);
    - Boyel’s law: \( V = C(T) / p \), Kinetic theory (Boltzmann);
    - Brown’s motion: \( \frac{\partial \rho}{\partial t} = D \frac{\partial \rho}{\partial x} \), \( \frac{x^2}{2t} = D = \frac{RT}{6 \pi \eta \nu N} \) (Einstein).

- Biophysical/Biochemical (Microscopic) and Evolutionary (dynamical) models
Chance-Life: Statistical Learning

• **Probabilistic Graphical** (chains/trees/DAGs) Models
  – Directed (Bayesian Networks, Phylogeny), Undirected (Markov Networks: HMM/generative, CRF/discriminative)
  – Representation (Conditional independence, H-C Thm: MN=Gibbs), Inference (DP/VP), Learning (MLE/BE, EM/MCMC, Sparsity, Regularization)

  [http://www.pgm-class.org/](http://www.pgm-class.org/)

• **Machine Learning & Learning Machines**
  – ANN, GA, Perceptron, SVM, Boosting, Boltzeman Machine

  [https://www.coursera.org/learn/machine-learning](https://www.coursera.org/learn/machine-learning)

  Belief, behavior, Boosting (Efron)
Machine Brain convergence

IBM's supercomputer Deep Blue (May 1997) beat chess master Garry Kasparov in a six-game match, in a dramatic reversal of their battle the previous year. Machine: extension of human being, replacing or beating man in specific functional task.

On March 15, 2016, the distributed version of AlphaGo won 4-1 against Lee Sedol, whose Elo rating is now estimated at 3,520. The distributed version of AlphaGo is now estimated at 3,586. It is unlikely that AlphaGo would have won against Lee Sedol if it had not improved since 2015.
Cognitive Computing
Welcome new and talented collaborating research students!