Computational Systems Biology: Biology X

Bud Mishra

Room 1002, 715 Broadway, Courant Institute, NYU, New York, USA

L#1:(Jan-24-2011)
Genome Wide Association Studies

Outline

"The curse of the human race is not that we are so different from one another, but that we are so alike."

–Salman Rushdie, *The Enchantress of Florence*, 2008.

Outline

Administrivia

- **Instructor:** Bud Mishra
- Room 1002, 715 Broadway
- email: mishra@nyu.edu
- phone: 212-998-3464
- Office Hours: Mondays, 2:00 pm – 2:45 pm

Administrivia

- BIOLOGY X
- **Course Details:** G22.3033-003
  || Computational Systems Biology
- **Time and Place:** 5:00-6:50 pm EST
  || Room 1221, 719 Broadway
- **Number of Credits:** 3 credits
- **Course Work:** Software Project, Analyzing Genetics Data
- **Languages of Choice:** R (May be Python, Matlab, Mathematica — But no Perl please)

Text Books

- **Required Textbook**: Andrea S. Foulkes || Applied Statistical Genetics with R: For Population-based Association Studies (Use R) || Springer; 1st edition (April 17, 2009).
  - **Recommended textbook (1):** Kenneth Lange || Mathematical and Statistical Methods for Genetic Analysis || Springer; 2nd edition (June 3, 2003).
  - **Recommended textbook (2):** Rongling Wu, Changxing Ma and George Casella || Statistical Genetics of Quantitative Traits: Linkage, Maps and QTL || Springer; 1st edition (July 31, 2007).
  - **Recommended textbook (3):** Geoffrey S. Ginsburg and Willard Huntington || Essentials of Genomic and Personalized Medicine || Academic Press; 1st edition (October 8, 2009).
  - **Recommended textbook (4):** Daniel Hartl and Elizabeth Jones || Genetics: Analysis of Genes and Genomes || Jones & Bartlett Publishers; 7th edition (August 1, 2008).

Outline

Genomics from a Population View-point

- **Main Thesis**
- Assume that in the not-so-distant future, we face no computational, technological or biological obstacles to gathering a large amount genomic (+epigenomic, transcriptomic, proteomic, etc.) data ... We may also have large amount EHR (Electronic Health Record) data
  1. *How would such data be anlayzed? Mathematical Models? Faster Algorithms?*
  2. *How can these data be put to use for better, cheaper and more universal health care?*
  3. *What is the analog of GOOGLE for biological information?*

Areas we wish to touch on...

- Ancestry and Population Models
- Genome Wide Association Studies
- Complex and Mendelian Diseases
- Common and Rare Diseases

Let us think about these inter-connected questions from a single global perspective...

Example 1: Sickle Cell Anemia

- *Mendelian Disorder*... Affects red blood cells, leading to hemolytic anemia and infection
- *Inherited Disorder*... Understood for millennia by the population of sub-Saharan Africa
- *Molecular Disorder*... Resulting from a genetic mutation in the hemoglobin gene chromosome 11.
- *Microscopic Investigation*.. “sickling” of red blood cells... genetic variant changes the shape of hemoglobin.
- *Heterozygotes’ Advantage*... Redistance against Malaria parasite, *Plasmodium falciparum*

Example 2: Alzheimer's Disease (AD)

- *Complex Disorder*: With a strong genetic component; multiple genes explaining AD risks. Four Genes: Large number of Rare Variants in three genes: Chromosomes 1, 14, & 21, or a single gene (Mendelian): Chromosome 19.
- *Brain Disorder*: Progressive destruction of brain cells... Dementia, Loss of memory, Social impairment & Death
- *Familial Disorder*: FAD... Early onset (prior to age 50)... Mendelian.
- *Environmental Risk Factors*: Head injury... Cardio-vascular components (high BP & T2D)

Early Onset Alzheimer's Disease (AD)

- Multiple (three) genes on different chromosomes. Rare disease variants responsible for early onset of familial AD.
  - APP (Amyloid beta Precursor Protein) on chr. 21.
  - PSEN 1 (Presenilin 1) on chr. 14.
  - PSEN 2 (Presenilin 2) on chr. 1.

Late Onset Alzheimer's Disease (AD)

- Single gene responsible for late onset of AD.
  - APOE (Apoliprotein E) on chr. 19.

Human Diseases

- How to think about them?

A Tentative Syllabus

I would like to focus this course on four basic questions...

1. Who are we (humans)?
2. Why are there diseases?
3. Why do we suffer?
4. Why do we die?

Possible Sets of Lectures

- **Lecture 1:** Introduction to Biology (Genomics)
- **Lecture 2:** Probability/Statistics/Information Measure, Causality and Correlation
- **Lecture 3:** Statisitical Analysis and Multiple Hypotheses Testing
- **Lecture 4:** Population Genetics
- **Lecture 5:** Neutral Model: Experiment Design (Capture/Recapture)
- **Lecture 6:** Population Structure: STRUCTURE/Mstruct, GeneFlow, Indian Population
- **Lecture 7:** Ancestry, Coalescence, Sufficient Statistics, ICA

Possible Sets of Lectures (Contd.)

- **Lecture 8:** Equilibria: Hardy-Weinberg, Sex-Ratio, Stability, Multiple Equilbria
- **Lecture 9:** Models of Selection: Detecting Selection and CoSelection
- **Lecture 10:** Sex-Linkage, Heterozygous Advantage
- **Lecture 11:** Genetic Diseases: Why do they exist: Cancer, Autism, Thalassamia
- **Lecture 12:** Evolution of Complex Diseases: CD-CV Hypothesis
- **Lecture 13:** GWAS for Rare Mendelian Disease
- **Lecture 14:** GWAS for Complex Diseases
- **Lecture 15:** The Future Challenges

Questions???

Heated Discussions on the Suggested Topics...
Resulting in a New and Better Syllabus...
That EVERYONE Loves!

Projects

- Indian Population: Structure and Gene Flow:
- AGRE data set (Autism)
- Rare Mendelian Disorder (Miller's Syndrome)
- Neural data analysis (Partha Mitra)
- Network Analysis (Laxmi Parida)

[End of Lecture #1]