Basic Information

- **Instructor: Haris Vikalo**
  - Contact information: hvikalo@ece.utexas.edu, (512) 232-7922
  - Office hours: POB 3.110, Tue/Thus 2:00pm-3:00pm

- **Teaching Assistant: Somsubhra Barik**
  - Contact information: barik4u@gmail.com
  - Office hours: TBD

- **Electronic course site: Canvas**
  - http://canvas.utexas.edu/
  - distribution of homework assignments, solutions, and class slides/notes
  - should be able to access it if you have UT EID and are registered

- **Course website: http://users.ece.utexas.edu/~hvikalo/ee381v.html**
  - class notes (mirrored from Canvas) and suggested reading
  - final project guidelines
Basic Information Cont’d

- **Textbook:** none
  - class notes, reading assignments will be distributed via course website, Canvas

- **Suggested reading:**

- **Grading (tentative):**
  - homeworks (30%), midterm (30%), final project (40%)

- **Homeworks and exams:**
  - 4-5 assignments (theory component + programming component)
  - midterm (take-home)

Prerequisites and Target Audience

- **Final project:** either expository (survey) or innovative (research)
  - up to two students can collaborate on a project
  - required written documents: (1) proposal and (2) final report
  - a list of possible projects will be provided shortly

- **Prerequisites:**
  - an undergraduate course in probability
  - programming experience (Matlab or Python)
  - no biology background required

- **Target audience:**
  - students specializing in signal processing / machine learning / algorithms / information theory who want to learn of applications in biology / genomics and get exposure to real data
  - students specializing in computational biology, who want to strengthen their knowledge of basic signal processing / machine learning / information theory
Course Description

- **Course Description**: An exploration of signal processing and data sciences problems encountered in the analysis of high-throughput genomics data
  - applications to diagnostics (e.g., viral strain recognition), studies of complex diseases (e.g., cancer), studies of immune system, phenotype prediction, non-invasive pre-natal testing

- **Topics include**:
  - DNA sequencing and sequence alignment;
  - base calling in high-throughput sequencing systems
  - reference-guided and reference-free (*de novo*) genome assembly
  - genotyping and single individual haplotyping (haplotype assembly);
  - RNA sequencing and ChiP-Seq;
  - DNA microarrays and quantitative polymerase chain reaction systems;
  - modeling and inference for genetic regulatory networks;
  - population haplotyping; phylogeny;
  - future sequencing technologies

Goals for the Term

- **Signal processing and “big data” challenges in genomics**
  - formulating problems, presenting solutions

- **Duality: computation and biology**
  - provide a biology/technology background to motivate a computational task
  - overview relevant computational techniques, derive solutions, analysis

- **Foundations and frontiers**
  - well defined conventional problems and general methodologies
  - contemporary challenges, future research directions, etc.

- **Major themes**:
  - **enabling biotechnologies**: modeling, algorithms, analysis of performance
  - **cellular systems**: computational methods for inferring their structure and understanding how they function
Theme #1: Enabling Technologies Cont’d

- **Detection and quantification of molecules:** high precision (quantitative polymerase chain reaction -- QPCR) or high throughput (DNA microarrays)

  - **QPCR:** high precision (quantifies small # of DNA molecules)
    - *in vitro* replication (amplification) of DNA molecules
    - applications to diagnostics (viral and bacterial detection), cancer markers identification, genetic fingerprinting (as in forensics), etc.

- **DNA Microarrays:** high throughput (screens 10,000s of molecules)
  - massively parallel biosensor arrays
  - used for studies of genetic diseases, drug discovery, genotyping (the specific genome of an individual), genetic pathway discovery, etc.
Theme #1: Enabling Technologies Cont’d

- **QPCR, DNA microarrays**: detect/quantify DNA molecules of known structure
- **DNA sequencing systems**: identify unknown structure

- Sanger sequencing: 1977 – 1990s
- 2nd generation sequencing: since 2007
- 3rd generation sequencing: since 2010

- High-throughput sequencing is revolutionizing research and medicine
  - routine sequencing tasks generating massive amounts of data
  - computationally challenging “big data” problems

Theme #1: Enabling Technologies Cont’d

- Dramatic improvement in affordability:
Theme #2: Cellular Systems

- Information flow in a cell (traditional view: Central Dogma):

  ![Diagram of DNA to Protein]

- Information (signal) is carried by molecules.

- Previously mentioned biotechnologies interrupt the information flow and so provide insight into the cellular structure and functions

  ![Diagram of Sequences to DNA, RNA, Protein]

- Moreover, study the temporal changes in the information flow; gives insight in regulation mechanisms, biological network structure, etc.
Signal Processing and Data Science Tasks

- Data science tasks on sequencing data can be categorized as follows:

  - **process data**
    - Assembly (de Novo)
    - Variant calling
    - Phasing
    - Quantification

  - **manage data**
    - Compression
    - Privacy

  - **utilize data**
    - Genome wide association studies
    - Multi-omics analysis
    - Phylogenetic tree reconstruction
    - Single-cell analysis

- To complete those tasks, we rely on a variety of tools:
  - statistical signal processing and machine learning
  - combinatorial algorithms
  - information theory
Example Application #1: Sequence Assembly

- Sequencing: determining the order of nucleotides in a target DNA string

- Shotgun sequencing: assemble the target from overlapping short reads
  - *de novo*: no side information, only the reads are available
  - reference-guided: rely on a pre-existing reference sequence

- Reference-guided assembly relies on mapping the reads onto a reference; sequence alignment/mapping is a fundamental first step
  - dynamic programming solutions (Viterbi, forward-backward algorithms)
  - estimation in Hidden Markov Models (EM algorithm)
  - data compression concepts (Burrows-Wheeler transform)

- Reference-free (*de novo*) assembly
  - greedy merging+extension of the overlapping fragments
  - finding Eulerian path in the de Bruijn graph
  - conditions for error-free reconstruction
Example Application #2: Haplotype Assembly

- In many applications, there are multiple target sequences of interest that cannot be separated prior to sequencing
  - haplotype assembly, viral quasispecies reconstruction, bacterial communities, immune cell repertoire
- The simplest one: haplotype assembly for diploids
  - reconstruct variable parts of chromosome pairs

```
AGGATTCC AAGTTAC CGAAATTCAGGATTCA GCTTAATGGCTT
AGGATAACC GAGTTAG CGAAATTCAGGATTCA AGCTTAATGGCTT
```

Example Application #2: Haplotype Assembly

- Shotgun sequencing for haplotype assembly:

```
ch1a AGGATTCC AAGTTAC CGAAATTCAGGATTCA GCTTAATGGCTT
ch1b AGGATAACC GAGTTAG CGAAATTCAGGATTCA AGCTTAATGGCTT
```

- Data model: short reads obtained by sampling (with replacement) from a complementary pair of binary strings
  - the task is to reconstruct the pair of strings
Example Application #2: Haplotype Assembly

- Methods for solving the haplotype assembly problem
  - (correlation) clustering
  - communication-theoretic techniques: decoding noisy codewords transmitted over a binary erasure channel
  - low-rank sparse matrix completion/factorization

- Analysis of fundamental limits of performance (accuracy, data redundancy)
  - Information-theoretic tools

Recent Special Issues in EE/CS Community

- Recent IEEE special issues (can be accessed via IEEE Xplore):