I. Administrivia

Introduction to the course and its goals
Course organization and content
Homework and Quiz
Term Project

Introductions

• Lecturers
  – Manolis Kellis
    (MIT CSAIL, Computational Biology, Broad Institute)
  – My own research:
    Comparative genomics, Gene Regulation, Evolution,
    Epigenomics, Phylogenomics, etc

• TAs
  – Rachel Sealfon
    (MIT Ph.D. candidate, Kellis Lab / Sabeti Lab)
  – Rachel’s research:
    Genomics of pathogens and immunity response

Course Information

• Lectures
  – TR 9:30 – 11:00, Room 4-159

• Recitations:
  – On Friday at 3pm in TBD
  – Recitations at MIT (HST/HSPH students can join)
  – Please indicate your availability on the sign up sheet

• Course Website
  – or simply: compbio.mit.edu/6.047 (redirects to stellar)
  – All handouts, lectures, notes, etc will be posted here.

• Course calendar:
  – On Google, public add: “6.047 Lectures”

Goals for the term

• Introduction to computational biology
  – Fundamental problems in computational biology
  – Algorithmic/machine learning techniques for data analysis
  – Research directions for active participation in the field
  – Understanding how methods work

• Ability to tackle research
  – Problem set questions: algorithmic rigorous thinking
  – Programming assignments:
    → hands-on experience w/ real datasets
  – Final project experience:
    → propose and carry out independent original research
    → present findings in conference format (written, oral)
Course content

Computation & Biology | Foundations & Frontiers

* Duality #1 (x-axis): Computation and Biology
  - Important, relevant, current biology:
    - Important biological problems
  - Fundamental computer science:
    - General techniques, principles

* Duality #2 (y-axis): Foundations and Frontiers
  - Foundations:
    - well-defined problems, general methodologies
    - ‘The classics’ of the field
  - Frontiers:
    - in-depth look at complex, current problems, open questions
    - combine techniques learned
    - opens to projects, research directions

Course organized around bio/comp modules

* Each module corresponds to an active area of research
  - 1: Comparative genomics: Alignment, Signatures, Assembly
  - 2: Genes and Transcripts: HMMs, RNA folding and function
  - 3: Regulation: Expression, Epigenomics, Motifs, Networks
  - 4: Variation: Disease/Trait mapping, Phylogeny, PopGen
  - 5: Frontiers: Personal/Disease, 3D genomes, Pharma, Synth

  For each module:
  - First half ⇝ the foundations
    - Dynamic programming, string matching, hashing, HMMs, EM, Gibbs Sampling, Clustering, Classification, Feature selection, SVMs, CRFs, Context-Free Grammars, phylogenetics, gene/species trees, evolutionary models, disease mapping

  For each module:
  - Second half ⇝ the frontiers
    - Evolutionary signatures, Transcript analysis, lincRNAs, Network inference and analysis, Epigenomics, Recent human selection and ancestry, chromatin regulation, Missing heritability,

Textbook / class notes / resources

(Optional) Books for the Course

Durbin, Eddy, Krogh, Mitchison
Jones, Pevzner
Duda, Hart, Stork

Availability: BU Coop, MIT Coop, amazon.com (~$40-60)
All three books on reserve at the MIT and BU Engineering libraries
New this year!! Book for the Course

Computational Biology: Genomes, Networks, Evolution
MIT Course 6.047/6.878
Manolis Kellis & all of you!

Availability: Online PDF, for free

Lectures and Scribing

- Each lecture will have a dedicated scribe who will take notes on the lecture
  - Please sign up to scribe for lecture on the sheet being passed around
- Build on notes from previous years
  - Available on course website
- Complete draft of scribe notes: before prev. lecture
  - Unless it’s not there from previous year (this is rare)
- Final draft of scribe notes due 6 days after lecture
  - Your grade depends on the improvement from previous year and completeness
- Some lectures need more work: multiple scribes
- Some tasks are better-suited to you than just scribing
  - E.g. figures, references, layout, macros, let us know!

Lectures and Scribing

Homeworks and quiz

Details on Problem sets

- Each problem emphasizes one lecture (or two)
  - Practical problem: gain experience in techniques, write code, download datasets, carry out analysis, interpret your results, learn about behavior of problem/method
    (Typically additional advanced problem for 6.878)
- Due Mon at 8pm (except Columbus Day → Wed)
  - Late policy: we are flexible, give or take a few hours
  - If more than a few hours, need prior arrangements, extensions typically not granted, except special circ.
- Submit all homeworks online from stellar page
  - No solutions distributed. If you’ve solved them, you know what you needed to learn/discover/achieve.

Details on the in-class quiz

- It’s not a midterm, and it’s not a final exam
  - It’s a quiz, friendly, fun, interesting, cute, fuzzy
- Demonstrate mastery of the material in 4 modules
  - Understand key points emphasized in lecture
  - Understand subtleties revealed in the psets
  - Ability to apply new skills to solve practical problems
- Types of questions
  - Knowledge questions: T/F justify, multiple choice
  - Deeper understanding questions: short answers
  - Practical problems: work through simple algorithm
  - Design problem(s): new/modified algorithm, need both knowledge and new idea, argue correctness
Final Project: Original Research in Comp Bio

- A major aspect of the course is preparing you for original research in computational biology.
  - Framing a biological problem computationally
  - Gathering relevant literature and datasets
  - Solving it using new algorithms, machine learning
  - Interpreting the results biologically
- Also ability to present your ideas and research
  - Crafting a research proposal (fellowships/grants)
  - Working in teams of complementary skill sets
  - Review peer proposals, find flaws, suggest improvements
  - Receiving feedback and revising your proposal
  - Writing up your results in a scientific paper format
  - Presenting a research talk to a scientific audience
- Term project experience mirrors this process

Details on the final project

- Milestones ensure sufficient planning / feedback
  - Set-up: find project matching your skills and interests
  - Team: common interests and complementary skills
  - Inspiration: last year’s projects, and recent papers
  - Proposal: establish milestones, deliverables, expectations
  - Midcourse: see endpoint, outline report, methods, figures
- Periodic mentoring sessions
  - Senior students and postdocs can serve as your mentors
  - Group discussions to share ideas, guidance, feedback
  - Peer-review: think critically about peer proposals, receive feedback/suggestions, respond to critiques, adjust course
- Real-world experience, condensed in a single term
  - Grant/fellowships proposals, peer review, yearly reports, budget time/effort, collaboration, paper writing, give talk

Finding a research mentor / research advisor

- Chance to meet faculty at MIT/Broad/Harvard/BU:
  - Through guest lectures and mentoring
  - Topics and papers covered in the lectures
  - Experts on: (1) human comparative genomics, (2) lincRNAs, (3) metabolic modeling, (4) disease mapping, selection, evolution and ecology (following four modules)
- Chance to meet senior students and postdocs:
  - On: coding genes, ncRNAs, regulatory motifs, networks, epigenomics, phylogenomics (again on each module)
  - Mentorship sessions with entire MIT CompBio group
- Your own personal research experience:
  - collaborators, datasets
  - learn active research directions, frontiers
  - living, breathing changing field

Putting it all together
Course Grading

- Grading:
  - Problem sets 30%
  - Final Project 40%
  - Midterm 20%
  - Scrib 10%

- 4 problem sets:
  - Each problem set: 7-10%, covers 3-4 lectures, contains 3-4 problems.
  - Algorithmic problems and programming assignments (PS1 out now)
  - Graduate version includes additional problem on current research

- Final project
  - Introduction to research in computational biology (7 weeks!)
  - Includes peer-reviewed NIH-style proposal and much feedback

- Quiz
  - In-class quiz (Tue Nov 20). No final exam.

- Collaboration policy
  - Collaboration allowed, but you must:
    - Work independently on each problem before discussing it
    - Write solutions on your own
    - Acknowledge sources and collaborators. No outsourcing.

Why Computational Biology: Last year’s answers

- Lots of data (* lots of data)
- There are rules
- Pattern finding
- It’s all about data
- Ability to visualize
- Simulations, temporal relationships
- Guess + verify (generate hypotheses for testing)
- Propose mechanisms / theory to explain observations
- Networks / combinations of variables
- Efficiency (reduce experimental space to cover)
- Informatics infrastructure (ability to combine datasets)
- Correlations, higher-order relationships
- Cycle from hypothesis generation to testing condensed
- Life itself is digital. Understand cellular instruction set
The components of genomes and gene regulation

Goal: A systems-level understanding of genomes and gene regulation:
- The genome: Map reads, align genes/genomes, assembly strategies
- The genes: Protein-coding exons, introns, non-coding RNA, RNA folding
- The control regions: Promoters, enhancers, insulators, chromatin states
- The actual words: Regulatory motifs, high-resolution accessibility maps
- The regulators: Transcription factors, chromatin modifiers, nucleosomes
- The dynamics: Changing maps between cell types, across development
- The networks: regulator→enhancer→target, ChIP-seq, correlated activity
- The grammars: TF/motif/mark combinations, predictive models
- Human variation: Human diversity, population genomics, linkage maps
- Evolution: Phylogenetics, phylogenomics, coalescent, human ancestry
- GWAS/QTLs: Genome variation ♫ organismal/molecular phenotypes
- Disease: Personal (epi)genomics, pharmacogenomics, synthetic biology

Challenges in Computational Biology

II. Life itself is digital

Quick introduction to molecular biology and information transfer within the cell
DNA: The molecule of heredity
- Self-complementarity sets molecular basis of heredity
  - Knowing one strand, creates a template for the other
  - "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, 1953

DNA: the four bases
<table>
<thead>
<tr>
<th>Purine</th>
<th>Pyrimidine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weak</td>
<td>Strong</td>
</tr>
<tr>
<td>Amino</td>
<td>Keto</td>
</tr>
</tbody>
</table>

DNA: chemical details
- Bases hidden on the inside
- Phosphate backbone outside
- Weak hydrogen bonds hold the two strands together
- This allows low-energy opening and re-closing of two strands
  - Anti-parallel strands
  - Extension 5'-3' triphosphate coming from newly added nucleotide

The only pairings are:
- A with T
- C with G

Module 1: Comparative genomics
- Foundations vs. frontiers
  - Foundations: Classical computational methods / biological topics
  - Frontiers: Latest developments, open questions, research areas
  - Duality for each: basic problems / fundamental techniques
- Sequence alignment:
  - Local/global alignment: infer nucleotide-level evolutionary events
  - Database search: scan for regions that may have common ancestry
- Comparative genomics
  - Detect evidence of purifying selection / function: 5% of human
  - Detect specific types of purifying selection: evolutionary signatures
Comparing Genomes

Chromosomes inside the cell
- Eukaryote cell
- Prokaryote cell

DNA packaging
- Why packaging
  - DNA is very long
  - Cell is very small
- Compression
  - Chromosome is 50,000 times shorter than extended DNA
- Using the DNA
  - Before a piece of DNA is used for anything, this compact structure must open locally
- Now emerging:
  - Role of accessibility
  - State in chromatin itself
  - Role of 3D interactions

Diverse epigenetic modifications
- 100+ different histone modifications
- Histone protein → H3/H4/H2A/H2B
- AA residue → Lysine4(K4)/K36...
- Chemical modification → Met/Pho/Ubi
- Number → Me-Me-Me(me3)
- Shorthand: H3K4me3, H2BK5ac
- In addition:
  - DNA modifications
  - Methyl-C in CpG / Methyl-Adenosine
  - Nucleosome positioning
  - DNA accessibility
- The constant struggle of gene regulation
  - TF/histone/nucleo/GFs/Chrom compete

"Central dogma" of Molecular Biology

DNA → makes → RNA → makes → Protein
Module II: Modeling genes and gene expression

- **Computational Foundations**
  - Hidden Markov Models (HMMs): Central tool in CS
  - Decoding, evaluation, parsing, likelihood, scoring
  - Unsupervised Learning: Expectation Maximization
  - Supervised learning: generative/discriminative models

- **Biological frontiers:**
  - PS2: Modeling conservation, GC content, CpG islands
  - L6/L7: Genome annotation and parsing
  - L8: Gene expression analysis: cluster genes/conditions
  - L9: Regulatory motif discovery: EM, gibbs sampling, info

Genes control the making of cell parts

- The gene is a fundamental unit of inheritance
  - Each DNA molecule \( \Rightarrow \) 10,000+ genes
  - 1 gene \( \Rightarrow \) 1 functional element (one “part” of cell machinery)
  - Every time a “part” is made, the corresponding gene is:
    - Copied into mRNA, transported, used as blueprint to make protein

- RNA is a temporary copy
  - The medium for transporting genetic information from the DNA information repository to the protein-making machinery is an RNA molecule
  - The more parts are needed, the more copies are made
  - Each mRNA only lasts a limited time before degradation

---

**mRNA: The messenger**

- Information changes medium
  - single strand vs. double strand
  - ribose vs. deoxyribose sugar
  - A T T A C G G T A C C G T
  - U A A U G C C A U G G C A
  - Compatible base-pairing in hybrid

**From DNA to RNA: Transcription**

- In Eukaryotes, not every part of a gene is coding
  - Functional exons interrupted by non-translated introns
  - During pre-mRNA maturation, introns are spliced out
  - In humans, primary transcript can be 10^6 bp long

**From pre-mRNA to mRNA: Splicing**

- Single Strand allows complex structure
  - Self-complementary regions form helical stems
  - Three-dimensional structure allows functionality of RNA

**RNA can be functional**

- Four types of RNA
  - mRNA: messenger of genetic information
  - tRNA: codon-to-amino acid specificity
  - rRNA: core of the ribosome
  - snRNA: splicing reactions

- To be continued...
  - We’ll learn more in a dedicated lecture on RNA world
  - Once upon a time, before DNA and protein, RNA did all
RNA structure: 2ndary and 3rdary

Splicing machinery made of RNA

Module II: Modeling genes and gene expression

- Computational Foundations
  - Hidden Markov Models (HMMs): Central tool in CS
  - Decoding, evaluation, parsing, likelihood, scoring
  - Unsupervised Learning: Expectation Maximization
  - Supervised learning: generative/discriminative models
- Biological frontiers:
  - PS2: Modeling conservation, GC content, CpG islands
  - L6/L7: Genome annotation and parsing
  - L8: Gene expression analysis: cluster genes/conditions
  - L9: Regulatory motif discovery: EM, gibbs sampling, info

Natural 1st step: group similar rows/columns

Clustering

- Similar cell types
- Similarly-behaving groups of genes

If labels are known: find more of same type

Classification

- Classify diseases
- Classify genes in different pathways

Find features that distinguish known classes
Find additional members of existing gene classes
Predict function of uncharacterized genes

"Central dogma" of Molecular Biology

DNA makes RNA makes Protein
Proteins carry out the cell’s chemistry

- More complex polymer
  - Nucleic Acids have 4 building blocks
  - Proteins have 20. Greater versatility
  - Each amino acid has specific properties
- Sequence → Structure → Function
  - The amino acid sequence determines the three-dimensional fold of protein
  - The protein’s function largely depends on the features of the 3D structure
- Proteins play diverse roles
  - Catalysis, binding, cell structure, signaling, transport, metabolism

Protein building blocks

- Amino Acids

From RNA to protein: Translation

The Genetic Code

- Substitutions that preserve AA properties tolerated in coding exons
- Leads to specific evolutionary signatures associated with protein-coding genes
- The code itself could be rediscovered simply based on observed substitution patterns

Structure of genetic code ≜ evolutionary signatures

These specify different rates of codon substitution, which in turn lead to different probabilities of any given alignment:
Summary: The Central Dogma

DNA makes RNA makes Protein

III. From Building Blocks to Networks

Introduction to regulatory and systems biology and control of information flow

Challenges in Computational Biology

1. Gene Finding
2. Sequence alignment
3. Database lookup
4. Genome Assembly
5. Regulatory motif discovery
6. Comparative Genomics
7. Evolutionary Theory
8. Gene expression analysis
9. Cluster discovery
10. Gibbs sampling
11. Protein network analysis
12. Metabolic modeling
13. Emerging network properties

Cellular dynamics and regulation

How cells move through this Central Dogma

Animal/Human gene regulation:
One genome ⇔ Many cell types

Eukaryotic Gene Regulation
Diverse roles for regulatory non-coding RNAs

- **Small RNA pathways (18-21 nt)**
  - microRNAs:
    - Repress genes by targeting their 3’UTRs by complementarity
    - Double-stranded RNA is then recognized and degraded
    - Recently found to also target promoter regions in rare cases
  - piwiRNAs
    - Target and repress transposable elements in germline
- **Long non-coding RNAs (1000s nt, many exons)**
  - Scaffolds for protein/TF binding
  - Scaffolds for 3D structure of RNA

Regulation of Gene Expression

- **Upstream of genes are promoter regions**
- **Contain promoter sequences or motifs**
- **Transcription factors (TFs) bind to motifs**
- **TFs recruit RNA polymerase**
- **Gene transcription**

Regulatory Interactions

- **Gene Activation**
- **Gene Repression**
- **Combinatorial Regulation**

Computational Motif Prediction

**How do we find new transcription factor binding sites?**

- Probabilistic model of promoters
  - Expectation maximization
  - Gibbs Sampling
- Comparative sequence analysis
  - Evaluate motif conservation across several related species

Network components reveal functional modules

- Feed-forward loops in developmental patterning
- Cooperation of master reg. & downstream reg.

Emerging properties of regulatory networks

- **Hierarchical levels of regulatory control**
  - Small number of backward-pointing edges
- **Specific / distinct feedback by microRNAs at each level**
  - Two classes of TFs: miRNA regulators and miR-regulated

Zeilinger et al., Genes & Development 2007
From Systems Biology to Synthetic Biology

- Components with known properties
- Assemble based on engineering goals / principles
- Implement within engineered cells and organisms
- Study behavior & adjust as needed

module V: Evolution/phylogeny/populations

- Phylogenetics /Phylogenomics
  - Phylogenetics: Evolutionary models, Tree building, Phylo inference
  - Phylogenomics: gene/species trees, reconciliation, coalescent, pops
- Population genomics:
  - Learning population history from genetic data (David Reich)
  - Statistical genetics: disease mapping in populations (Mark Daly)
  - Measuring natural selection in human populations (Pardis Sabeti)
  - The missing heritability in genome-wide associations (Yaniv Erlich)
- And we’re done! Last pset Nov 21st, In-class quiz on Nov 22nd
  - No lab 4! Then entire focus shifts to projects, Thanksgiving, Frontiers

Extinctions part of life

Over-express a single microRNA leads to new wing

- Discovery of sense/anti-sense miRNAs
- Regulatory switch selects between two developmental programs
- By over-expressing one strand (miRNAs) the balance is tilted
- Wing program launched vs. haltere

Stark et al, Genes & Development 2007
Phylogenetics

**General Problem:**
Infer complete ancestry of a set of ‘objects’ based on knowledge of their ‘traits’

‘Objects’ can be: Species, Genes, Cell types, Diseases, Cancers, Languages, Faiths, Cars, Architectural Styles

‘Traits’ can be: Morphological, molecular, gene expression, TF binding, motifs, words…

**Historical record varies:** Fossils, imprints, timing of geological events, ‘living fossils’, sequencing of extinct species, paintings, stories.

**Today:** Phylogenies using only extant species data

→ gene trees (paralog / ortholog / homolog trees)