Computational Biology: Genomes, Networks, Evolution

MIT 6.047 / 6.878
HSPH IMI.231
HST.507

Prof. Manolis Kellis
TA: Matt Edwards

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
  – Matt Edwards
    (MIT CSAIL, Computational Genomics)
  – Matt’s research:
    Evolution of gene regulation in yeast

Course Information

• Lectures
  – TR 9:30 – 10:00, Room 4-159

• Recitations:
  – On Friday: Time to be determined (fill up survey!)
  – Recitations at MIT (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 Students”

(Optional) Books for the Course

Durbin, Eddy, Krogh, Mitchison
Availability: BU Coop, MIT Coop, amazon.com (~$40-60)
All three books on reserve at the MIT and BU Engineering libraries

Duda, Hart, Stork

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 four modules**

- Each module corresponds to an active area of research
  - Comparative genomics: alignment and feature discovery
  - Genome annotation: coding and non-coding genes
  - Gene regulation & networks: expression, motifs, epigenomics
  - Evolution & phylogenomics: species, populations, selection
- First half of each module: the foundations
  - Dynamic programming, string matching, hashing, HMMs, EM, Gibbs Sampling, Clustering, Classification, Feature selection, SVMs, CRFs, CF Grammars, tree building, gene/species trees, evolutionary theory, disease mapping
- Second half of each module: the frontiers
  - Evolutionary signatures, RNAseq, microRNAs, PhyloSCFGs, Bayesian network inference and analysis, Epigenomics, Recent human selection and ancestry, metabolic modeling, bacterial genomics, lincRNAs and chromatin regulation

**Four modules ↔ four psets ↔ four research areas**

- 4-5 lectures each  ↔ 4-5 problem each. Not a coincidence

---

**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

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
- First draft of scribe notes due 2 days after lecture
  - Unless it’s already there from previous year
- Final draft of scribe notes due 6 days after lecture
  - Your grade depends on the improvement from previous year

Final Project

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

Final Project at a Glance

Final Project

- Project description: Describe your previous research, areas of interest in computational biology, type of project that best fits your interests. Put these in a profile that helps your colleagues know you and find potential partners.
- Due Date: Feb 1
- Project goals: Specific project goals, division of work, milestones, deliverables, challenges, aims, algorithms, in form of PH proposal.
- Project proposal due Mar 15.
- Final project description due Apr 15.
- Project progress report. Continue making progress on proposed project. Adjust milestones as needed. Write outline of final project report.
- Due Date: Apr 30.
- Final project report. Finalize methods, milestones, results, figures, finalize in conference paper format. In report, comment on lessons learned and overall project experience.
- Final project presentation: 10 min conference talk on your final project. May 15.

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:
  - Guest lectures by: David Reich, John Rinn, James Galagan, Mark Daly, Pardis Sabeti, Eric Alm
  - 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:
  - Talks by: Mike Lin, Stefan Washietl, Pouya Kheradpour, Sushmita Roy, Jason Ernst, Matt Rasmussen
  - On: coding genes, ncRNAs, regulatory motifs, networks, epigenomics, phylogenomics (again on each module)
  - Mentorship sessions with entire MIT CompBio group

- Also: collaborators, datasets, learn active research directions, frontiers, living, breathing changing field

Course Grading

- Grading:
  | Problem sets 35% | Final Project 40% | Midterm 20% | Scrut% |
- 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 23). 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?

- Lots of data (* lots of data)
- There are rules
- Pattern finding
- It’s all about data
- Ability to visualize
- Simulations
- 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
- Life itself is digital. Understand cellular instruction set

Putting it all together

The term at a glance

Why Computational Biology: Last year’s answers
Challenges in Computational Biology

1. Genome Assembly
2. Regulatory motif Discovery
3. Gene Finding
4. Comparative Genomics
5. Evolutionary Theory
6. Database lookup
7. Gene expression analysis
8. Protein network analysis
9. Metabolic modelling
10. Emerging network properties

“Central dogma” of Molecular Biology

DNA makes RNA

RNA makes Protein

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: chemical details
• Bases Hidden on the inside
  Phosphodiester bonds hold the two strands together
• This allows low-energy opening and re-closing of two strands
  • Anti-parallel strands
  • Extensions 5'-3' tri-phosphate coming from newly added nucleotide
  The only pairings are:
    • A with T
    • C with G

DNA: the four bases

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

Diversity of epigenetic modifications

Genes control the making of cell parts

- The gene is a fundamental unit of inheritance
  - Each DNA molecule ⊆ 10,000+ genes
  - 1 gene ⊆ 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

“Central dogma” of Molecular Biology

mRNA: The messenger

- Information changes medium
  - single strand vs. double strand
  - ribose vs. deoxyribose sugar
  - Compatible base-pairing in hybrid

DNA wrapped around histone proteins

Image source: http://nihroadmap.nih.gov/epigenomics/

Image source: http://nihroadmap.nih.gov/epigenomics/
**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**

- Alternative splicing can yield different exon subsets for the same gene, and hence different protein products.

**RNA can be functional**

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

**“Central dogma” of Molecular Biology**

- DNA makes RNA.
- 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 structure

- Alpha-beta horseshoe
  - Common motif for DNA-binding proteins that often play a regulatory role as mRNA level transcription factors

- Beta-barrel
  - Some antiparallel β-sheet domains are better described as β-barrels rather than β-sandwiches, for example streptavidin and porin. Note that some structures are intermediate between the extreme barrel and sandwich arrangements.

Beta-turn helix

- Helix-turn-helix
  - Some antiparallel β-sheet domains are better described as β-barrels rather than β-sandwiches, for example streptavidin and porin. Note that some structures are intermediate between the extreme barrel and sandwich arrangements.

Protein building blocks

- Amino Acids

From RNA to protein: Translation

- Ribosome

The Genetic Code

- Use evolutionary and compositional properties to computationally discover protein-coding genes

Structure of genetic code vs evolutionary signatures

- 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
- These specify different rates of codon substitution, which in turn lead to different probabilities of any given alignment:

\[ P_{Q_{c}} = \frac{2}{3} \]
\[ P_{Q_{t}} = \frac{1}{3} \]
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. Genome Assembly
2. Sequence alignment
3. Database lookup
4. Comparative Genomics
5. Regulatory motif discovery
6. Evolutionary Theory
7. Gene expression analysis
8. RNA transcript cluster discovery
9. Gibbs sampling
10. Protein network analysis
11. Metabolic modeling
12. Emerging network properties

Cellular dynamics and regulation

How cells move through this Central Dogma

DNA makes RNA makes Protein

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

Eukaryotic Gene Regulation

Gene regulatory sequences
Gene regulatory proteins
Transcription factors
RNA polymerase
Protein synthesis
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
  - snoRNAs
  - 21U-RNAs

- Long non-coding RNAs (1000s nt, many exons)
  - Scaffolds for protein/TF binding
  - Scaffolds for 3D structure of RNA

The components of genomes and gene regulation

Goal: A systems-level understanding of genomes and gene regulation:
- The regulators: TFs, GFs, miRNAs, their specificities
- The regions: enhancers, promoters, insulators
- The targets: individual regulatory motif instances
- The grammars: combinations predictive of tissue-specific activity
  - The parts list = Building blocks of gene regulation

Our tools: Comparative genomics & large-scale experimental datasets.
- Evolutionary signatures for promoter/enhancer/3'UTR motif annotation
- Chromatin signatures for integrating histone modification datasets
- Sequence signatures associated with TF binding, chromatin, dynamics
  - Infer regulatory networks, their temporal and spatial dynamics
  - Integrate diverse datasets

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?

- Gene regulated by same TF
  - 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

Network modeling of metabolic pathways

- Met: Energy complex molecular
- Nodes: metabolites, edges: proteins
- Product of one = substrate of next
- Genome metabolic map
- Expression reaction rates
- Predict: steady-state, time-course, mutant phenotypes, gene regulation

From Regulatory Networks to Systems Biology

- Predictive models
- Large-scale behavior (systems-level)
- Feedback and stochastic processes
- Emergent behavior can be unexpected

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
  
Jay Keasling

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 (miRNAas) the balance is tilted
- Wing program launched vs. haltere

Challenges in Computational Biology

- Genome Assembly
- Regulatory motif discovery
- Gene Finding
- Comparative Genomics
- Evolutionary Theory
- Gene expression analysis
- Protein network analysis
- Metabolic modeling
- Emerging network properties

Stark et al. Genes & Development 2007
The term at a glance