Why study computational genomics?

- DNA sequencing has changed biology
- Not everything is in the genome — but an awful lot is
- Latest revolution: comparative genomics — structure, function, evolution
- Large complex data sets require statistics and computation
- Interesting and exciting things happen at boundaries between disciplines

![Image of overlapping circles representing Natural Sciences, Engineering, and Mathematics]

Course Information

- Lectures: Tues/Thurs, 11:40-12:55, Warren 231
- Recitations: Thurs, 1:25-2:15, Bradfield 105
- Credit Hours: 4 (S/U or letter)
- Professor: Adam Siepel, 101 Biotech
- Office Hours: Tues/Thurs, 4:30-5:30
- TA: Chuan Gao, 169 Biotech
- Course web page: [http://compgen.bscb.cornell.edu/btry484](http://compgen.bscb.cornell.edu/btry484)
- Note: BTRY 484 and 684 are concurrent
Goals for this Course

- Learn nitty-gritty of probabilistic sequence analysis
- Survey phylogeny reconstruction, biological networks, gene expression analysis, and other topics
- Learn fundamental theory, also see how it is applied
- Along the way, learn some stats, computer science, molecular evolution, genomics
- Learn about modeling: be rigorous and mathematical, but also be flexible and practical
- Disclaimer: first time taught

Topics Covered

- Part I: Fundamentals (molecular biology, statistics, algorithms)
- Part II: Modeling and Inference for Sequences and Trees (motif models, HMMs, phylogenetic models)
- Part III: Advanced Modeling and Inference (pair HMMs, profile HMMs, Gibbs sampling, SCFGs for RNA)
- Part IV: Important Heuristics (BLAST, multiple alignment, distance-based tree reconstruction)
- Part V: Beyond Sequences (gene expression, networks)
- Part VI: Applied Comparative Genomics (conserved elements, phylogenetic gene prediction, lineage-specific selection)

What we will not cover

- Population genetics, association mapping, quantitative genetics, etc. (Bustamante, Mezey, Clark, Aquadro, Durrett)
- Protein folding (Elber)
- Genome rearrangements (Durrett)
- Physical mapping
- Sequence assembly

Main Textbook

- Biological sequence analysis
  - Probabilistic models of protein and nucleic acids

Some Other Useful Books

Prerequisites

- No formal prerequisites, but...
- This course is likely to be hard if you haven’t had a statistical methods class, intermediate programming/algorithms, or basic molecular biology/genetics
- Recommended: strong background in at least two out of three areas
Grading

- Components and breakdown:
  - Homeworks: 10% x 4 = 40%
  - Project: 30%
  - Midterm: 20%
  - Participation: 10%
- Two weeks per homework. Start early, use recitations
- Midterm designed to keep you honest, not kill you
- Participation is important.

Class Project

- Opportunity to apply new concepts and skills
- Should be substantial! Projects may grow into papers, Ph.D.s
- Different expectations for 484 and 684
- Proposal due mid-semester
- Example projects: motif identification with conservation, motif finder, phyloHMMs, pseudogenes
- For 484: literature review, implement published idea
- Projects due finals week

Collaboration

- Feel free to discuss homeworks, but you must turn in your own work and acknowledge all outside sources
- Make sure you understand everything you turn in
- Only individual projects are allowed

A Whirlwind Tour of Molecular Genetics

Stuff Floating in Cells

The Cell Cycle
The Structure of DNA

Credit: Watson et al., Molecular Biology of the Gene, CSHL Press, 2004
Genes are Intermittent Segments of Chromosomes

The Central Dogma

Protein Structure

Transcription

Table 3-5: The Genetic Code

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>RNA 1st position</th>
<th>RNA 2nd position</th>
<th>Protein</th>
<th>RNA 3rd position</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5' U</td>
<td>5' U</td>
<td>5' A</td>
<td>5' U</td>
</tr>
<tr>
<td>U</td>
<td>5' A</td>
<td>5' A</td>
<td>5' C</td>
<td>5' A</td>
</tr>
<tr>
<td>C</td>
<td>5' G</td>
<td>5' G</td>
<td>5' G</td>
<td>5' A</td>
</tr>
<tr>
<td>G</td>
<td>5' C</td>
<td>5' C</td>
<td>5' U</td>
<td>5' C</td>
</tr>
</tbody>
</table>

Figure 3-7: Four levels of protein structure.

Primary

Secondary

Tertiary

Quaternary
Not All Genes Encode Proteins

mRNA Splicing

Alternative Splicing

Initiation of Transcription in Eukaryotes

More Complicated Initiation
Modern View of Transcriptional Regulation

Post-transcriptional Regulation

Chromatin: Packaging of DNA

Processed Pseudogenes
Almost Half of the Human Genome is Transposons

Mutations, Large and Small

More on Mutations

- Other mutations:
  - Translocations
  - Transpositions
  - Duplications
  - Nucleotide substitutions
- Micro vs. macro mutations (e.g., indels)
- Silent vs. replacement mutations (e.g., subst.)
- Mutations occur in individuals and may or may not become fixed in populations
- We will generally view them as edit operations on genomes for whole species
- Modeling these edit operations is central

A Phylogeny
Species Trees, Gene Trees

Rooted and Unrooted Trees

Reconstructing a Tree from Sequences

Alignment of Sequences

Multiple Alignment
See you next week!

- Please review *Molecular Biology of the Gene* or similar, as needed
- Readings for Tues: DEKM ch 1, pp 299-301, 305-313; DTW ch 2-3 (optional)
- Readings for Thurs: DEKM ch 2; JP ch 2&5 or CLRS ch 15 (optional)
- First homework assigned Fri, Sep 1., due Fri, Sep 15.
- See class web page for full schedule