Genome 373: Genomic Informatics

Professors
Elhanan Borenstein and Jim Thomas
Genome 373

• This course is intended to introduce students to the breadth of problems and methods in computational analysis of genomes, arguably the single most important new area in biological research.

• The specific subjects will include:
  • Sequence alignment
  • Sequencing and next generation sequencing
  • Gene prediction
  • Molecular evolution
  • Evolutionary relationships and phylogeny
  • Clustering, classification, enrichment analysis
Outline

• Course logistics
• Introduction to Bioinformatics
• Introduction to Python
Instructors

• **Elhanan Borenstein**: Weeks 1-5
• **Jim Thomas**: Weeks 6-10
• Office hours: Monday 11:20-12:00

• **Rachel Diederich** (TA) will teach additional topics including programming and problem solving skills.

• Material covered in section is required, and will be on the exams.
Webpage

• Web site:
  
  http://elbo.gs.washington.edu/courses/GS_373_13_sp/

• Page has links to
  – Lecture notes
  – Handouts
  – Homework assignments
  – Many useful resources on:
    • Bioinformatics
    • Python
Programming

• Note: Historically, this course required prior programming experience.

• The first couple of weeks in class and the first few weeks of section will focus on learning to program in Python.

• If you do not have any programming experience, that’s ok, but ... you will need to work hard to catch up.
Grading

• 50% homework
• 20% midterm exam (in class)
• 30% final exam, Mon, June 10
Homework

• Posted online each Wednesday and due the following Wednesday.
• Homework is a mix of written problems and programming.
• Homework assignments are to be submitted by email!
• Programming assignments should be implemented in Python. For other languages, please ask Rachel.
• More on home assignment submission in the quiz section.
Textbooks

References:
Electronic access to journals is generally free from on-campus computers. For off-campus access, follow the "[offcampus]" links or look at the library "proxy server" instructions.
3. How dictionaries work (aka hash tables or hash maps)

Python Resources:
General
- Free (legal) PDF text book
- Python quick reference guide
- Much more extensive python quick reference guide
- Python.org
- Python Documentation
  - Tutorial
  - Howtos
Regular Expressions
- Intro
- Howto
- Another Howto
Library Reference
"RegExPal" (For Javascript rather than Python, but similar and quite handy. Try it!)

Biopython
- Biopython.org
- Tutorial and Cookbook

Python Books
- Python for Software Design: How to Think Like a Computer Scientist by Allen B. Downey. (Includes early drafts of our text book, cheaper than the published version, but less polished...)
- Learning Python by Mark Lutz, O'Reilly (Very comprehensive. Much is accessible to beginners.)
- Dive Into Python 3 by Mark Pilgrim. (Another online book. Based on Python 3, so some differences, and more advanced, but also free.)

Bioinformatics Books
- Inferring Phylogenies, Joseph Felsenstein, Sinauer, 2004. (Excellent reference on this topic)
- Bioinformatics: Sequence and Genome Analysis, David W. Mount, Cold Spring Harbor Laboratory Press.
- Python for Bioinformatics, Sebastian Bassi, CRC Press, 2010. (A little too advanced as a programming book for beginners, but fine now that you're experienced.)
- Python for Bioinformatics, Jason Kinser, Jones and Bartlett, 2009. (Ditto.)
Background survey

Please write on the index card your:

1. Name and email
2. Major
3. Primary background (biology, computation, other)
4. Programming experience (how much, what language)
5. Registered/not-registered/waiting-list
Why Bioinformatics?
Find the binding sequence: caattatatgttaaa
Find the binding sequence: caattatgttaaa
Moore’s law

Computer processing power doubles every ~2 years.

Microprocessor Transistor Counts 1971-2011 & Moore’s Law

dotted line - 2 year doubling
Sequencing cost decreasing much faster than computing cost

>2-fold drop per year

? - changing so fast hard to be specific
• Viruses ~3-1200 Kb
• Bacteria ~1-5 Mb
• Archaea ~1-5 Mb
• Fungi ~10-50 Mb
• Animals ~100-5,000 Mb
• Plants ~100-10,000 Mb

As of 2011 done or nearly done
• > 2,000 viruses
• > 1,000 bacteria and archaea
• Hundreds of fungi
• Dozens of protists
• Dozens of nematodes and insects
• 6 fish, 1 reptile, 4 birds, 1 amphibian
• About 10 plants
• About 40 mammals (+multiple individuals)
• Microbial communities (e.g., human microbiome)
A computational bottleneck
Find the binding sequence: **caattatgtttaaa**

... allowing for one mutation and one insertion

cggtatgtttaaa  catttatgtttaaa
caattt-atgtttaaa  cagttat-gttttaaa
cagtttatgtttaaa  caatttatgtttaaa
cagtttatgtttaaa-a  caatttatgttttaaa
cagtttatgttttaat  caaatttatgtttaaa
caat-tatgtttaaa  caatttatgtttttaaa
caat-tatgtttttaa  c-atttatgtttttaa
caat-tatgtttttaga
How well can the string

**GAATTCAGTTA**

match the string

**GGATTCGA?**

*(what is the best *alignment* between the two strings?)*
Informatics Challenges: Examples

• Sequence comparison:
  – Find the best match (alignment) of a given sequence in a large dataset of sequences
  – Find the best alignment of two sequences
  – Find the best alignment of multiple sequences

• Motif and gene finding

• Relationship between sequences
  – Phylogeny

• Clustering and classification

• Many many many more …