Computational Genomics and Molecular Biology

Instructor: Dannie Durand
TAs: Philip Davidson, Han Lai
Fall 2011

Course overview

Course web page:

http://www.cs.cmu.edu/~durand/03-711

Reading lists:

http://www.cs.cmu.edu/~durand/03-711/reading.html

Materials available on Electronic Reserves:

http://www.library.cmu.edu/

Recommended text books
(No required textbook)

Course overview

Syllabus:

http://www.cs.cmu.edu/~durand/03-711/syllabus.html

- Reading assignments will be posted on this page.
  - Some material is password protected;
  - Id: compbio, passwd: genomics
- You can download homework, solution sets, and class notes from this page.

Course overview

Materials available on Electronic Reserves:

http://www.library.cmu.edu/

Electronic reserves:

http://www.library.cmu.edu/

Click on:
Course Reserves

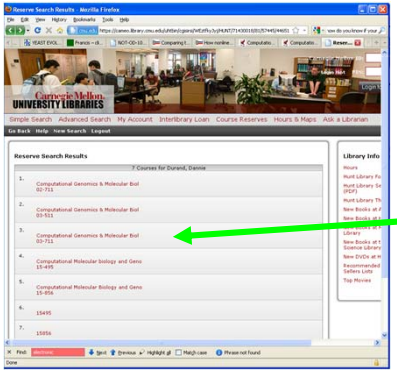
Select: Look up items on reserve by instructor

Type: "Durand"

Select: 03-711

Electronic reserves:

http://www.library.cmu.edu/



Click on:
Course Reserves

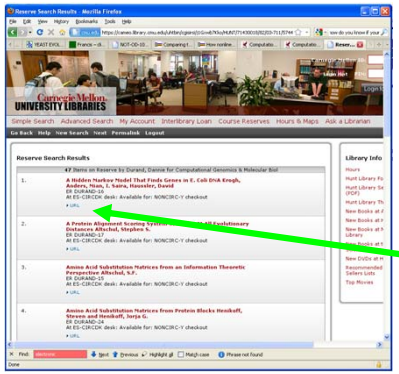
Select: Look up items on reserve by instructor

Type: "Durand"

Select: 03-711

Electronic reserves:

http://www.library.cmu.edu/



Click on:
Course Reserves

Select: Look up items on reserve by instructor

Type: "Durand"

Select: 03-711

Click on URL
to download PDF

Course overview

Coursework and policies:

http://www.cs.cmu.edu/~durand/03-711/policies.html

How to do well in this course

- Come to class
- Take notes
- Come to office hours
- Preparing for exams
  - Homework is more focused on working problems
  - Exams are more focused on concepts
  - Study from your notes as well as your homework

I speak quickly.

My handwriting is terrible.

Please interrupt and ask questions.

Outline

- Origins of computational molecular biology
- An overview of computational molecular biology
- Functional and computational Genomics.

The Origins of Computational Biology

Sanger: peptide sequencing by partition chromatography

Edman: stepwise protein degradation

Sanger sequences insulin,

Discovery of DNA structure

Stein, Moore, Spackman: automatic amino acid analyzer

Myoglobin

Ribonuclease

Lysozyme

1950

1960

Turing, v. Neumann: specs for stored program computer

First transistor

Edsac: 1st stored program computer

Grace Murray Hopper: First compiler

Fortran

First integrated circuit

Basic

The Origins of Computational Biology

Sanger: peptide sequencing by partition chromatography

Edman: stepwise protein degradation

Sanger sequences insulin,

Discovery of DNA structure

Stein, Moore, Spackman: automatic amino acid analyzer

Myoglobin

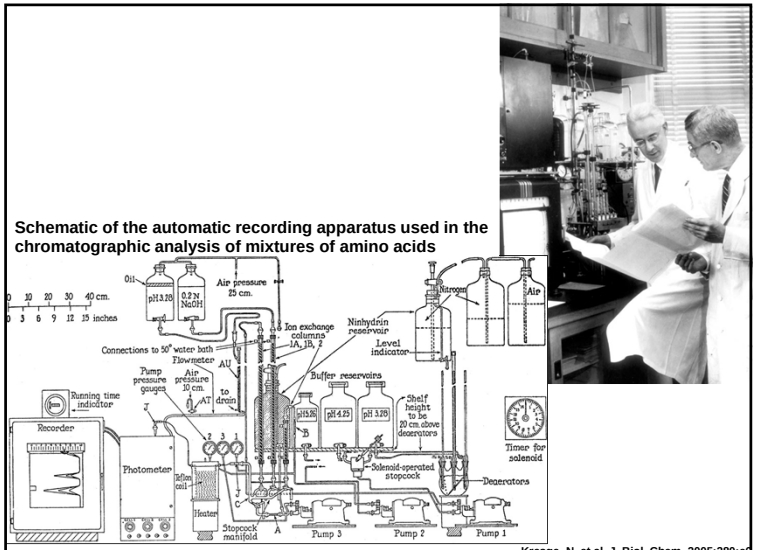
Ribonuclease

Lysozyme

Turing, v. Neumann: specs for stored program computer

Trends in Biochem. Sci, 99

Basic



Schematic of the automatic recording apparatus used in the chromatographic analysis of mixtures of amino acids

Kresge, N. et al. J. Biol. Chem. 2005;280:e6

The Origins of Computational Biology

Sanger: peptide sequencing by partition chromatography

Edman: stepwise protein degradation

Sanger sequences insulin,

Discovery of DNA structure

Stein, Moore, Spackman: automatic amino acid analyzer

Myoglobin

Ribonuclease

Lysozyme

1950

1960

Turing, v. Neumann: specs for stored program computer

ENIAC

First transistor

EDSAC: 1st stored program computer

Grace Murray Hopper: First compiler

Fortran

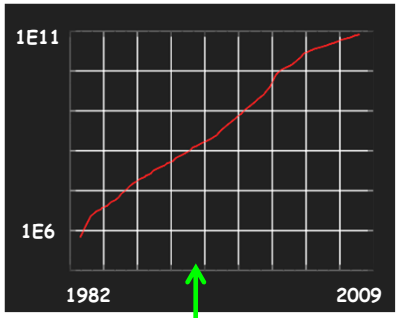
First integrated circuit

Basic

The Origins of Computational Biology

		ARPANET
	1970	
Sanger-Coulson sequencing Maxam-Gilbert sequencing		TCP/IP Internet
Gilbert, Sanger win Nobel Prize	**1980**	
		First royal email USENET newsgroups
Congress establishes Genbank Human Genome Project begins GenBank goes online.	**1990**	World Wide Web, Gopher NCSA Mosaic Pizza Hut goes on line
First whole genome sequence	**1995**	

Genbank doubles every 18 months



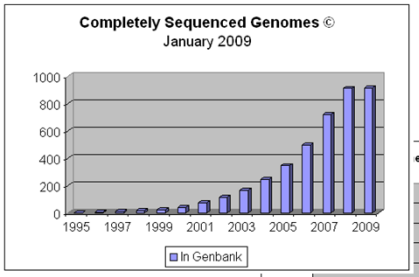
First whole genome sequence

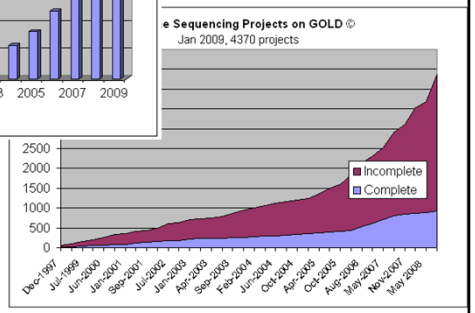
Whole Genome Sequencing Highlights (A Eukarya-centric View)

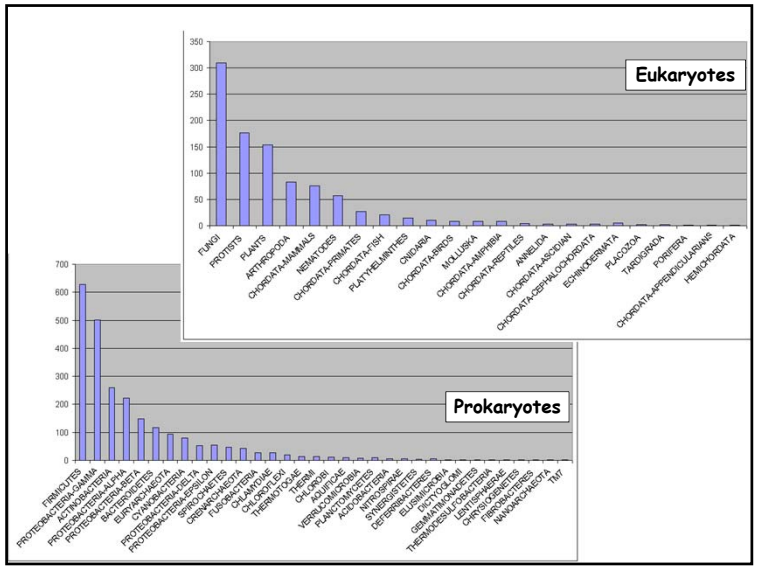
1995 *H. influenzae* – 1st whole genome sequence
1997 Yeast – 1st eukaryotic sequence
1998 *Caenorhabditis elegans* – 1st multicellular organism
2000 Fly, *Arabidopsis thaliana* – 1st plant
2001 Human
2002 Mouse, *Ciona intestinalis,*
2003 *Caenorhabditis briggsae, Neurospora Crassa*
2004 Five more yeasts, silkworm, rat, *C. merolae,* tetraodon
2005 Dictyostelium, zebrafish, chimpanzee
2007 Twelve drosophila genomes
…

Whole Genome Sequencing

www.genomesonline.org







Next-generation short read sequencing

Sanger Sequencing

- read lengths up to 1,000 bp
- accuracy 99.999%
- costs $0.50 per kilobase

Illumina sequending

- read lengths up to 36 bp
- error rates 1-1.5%
- cost $2 per megabase

454 sequencing

- read lengths 200-300 bp
- accuracy problem with homopolymers
- costs $60 per megabase

Next-generation short read sequencing

Advantages

- High throughput
- Does not require PCR amplification
- Accurate measures of abundance

Disadvantages

- Short reads are unlikely to be unique.
- Difficult to identify the origin of a given read
- Particular challenge for genome assembly

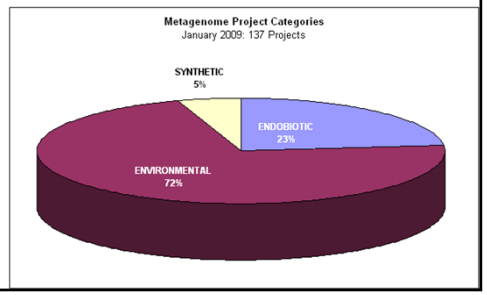
Next-generation short read sequencing

Applications

- Sequencing bacterial genomes
- Resequencing
  - identify mutations associated with drug resistance;
  - identify mutations in cancerous tissues.
  - genetic diagnostics
- Functional genomics

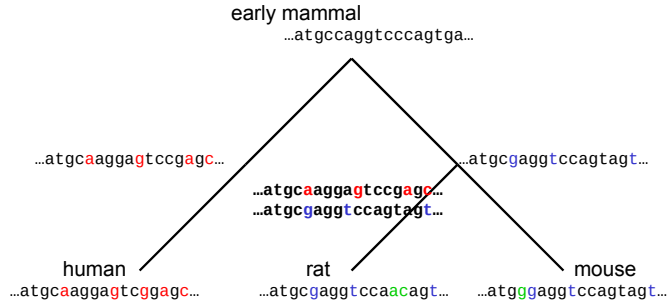
Metagenomics

- *Production-scale plant fermenter*
- *Fungal communities from the Arctic*
- *Singapore indoor air filters*
- *Yellowstone Obsidian Hot Spring*
- *Many fecal microbiomes*
- *Fossil microbiome*

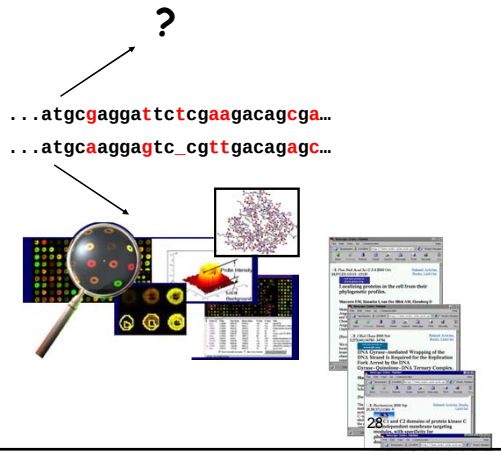


Why sequence data is so powerful:

Sequences are related!



Sequence similarity → functional similarity



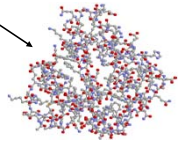
Sequence similarity → structural similarity

Structure?

…vkltpegtr_wgghpldekflske…

Estimate protein structure from a related protein with known structure and similar sequence.

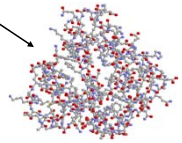
…vhltpettrgwgghmldekeiske…



Sequence similarity → structural similarity

Structure?

…vkltpegtr_wgghpldekflske…
…vhltpettrgwgghmldekeiske…



Estimate protein structure from a related protein with known structure and similar sequence.

Sequence Comparison

…atgcaaggagtcccagagcctgagctgactacgt…
…atgcgaggtctcccagtgtctgaactgactaagt…

global pairwise alignment

…atgcaagtcgtcccagtgccagaactcccctacgt…
…accagtggtctccgagtggctgaactgactaaca…

local pairwise alignment

tfisll v..frrda.h ksevahrfkd lgeenfkalv…
tfisll v..frrea.h kseiahrfnd vgeehfiglv…
tfisll v..frrdt.y kseiahrfkd lgeqyfkglv…
tlisfi lqrfardaeh kseiahrynd lkeetfkava…

global multiple alignment

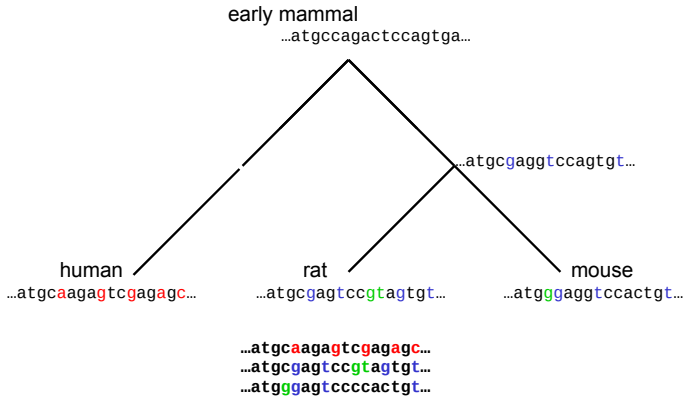
…mkwvtfisll flfssaysrg v..frrda.h ksev
…mkwvtfisll flfssaysrg v..frrea.h kseia
…~~wvtfisll flfssaysrg v..frrdt.y kseia
…mkwvtlisfi flfssatsrn lqrfardaeh kseia

local multiple alignment

Applications

- Database searching
- RNA structure prediction
- Evolutionary tree reconstruction
- Gene finding
- Sequence assembly….

Reconstructing Evolutionary History

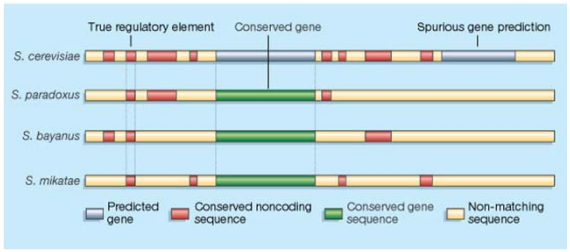


DNA PATTERNS IN THE *E.coli* lexA GENE

Global alignment of upstream sequences

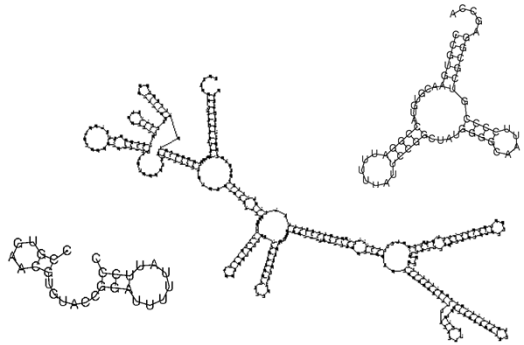
Kellis *et al*, *Nature*, 03

Sequence conservation



Salzberg, Nature, 2003

RNA Secondary Structure

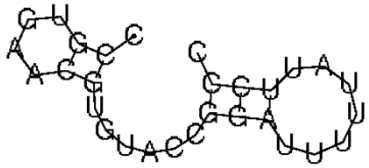


RNA Secondary Structure

CCGUGUACGUGUAACGGAUCUUUAUUCCC . . .
CCGUGAACGUAUACUGCAGUUUUAGUGCG . . .
GGCUCACGCUGUUCCGGAUUAUGAUUCCC
CGCAGAAGCUCCACGCGUUUUCUGUACGA

RNA Secondary Structure

CCGUGUACGUGUAACGGAUCUUUAUUCCC . . .
CCGUGAACGUAUACUGCAGUUUUAGUGCG . . .
GGCUCACGCUGUUCCGGAUUAUGAUUCCC
CGCAGAAGCUCCACGCGUUUUCUGUACGA

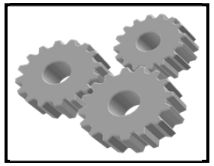


The Fantasy

Whole genome sequence

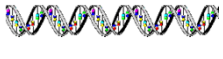
GAAATAAACAACCAGGCAGCAGTTATTAACACGGGAACATGGCGGCCGCAGCCTGGGCTCCCGCGGCGGCGGCGG…

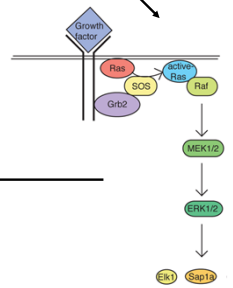
Cell Simulator Compiler

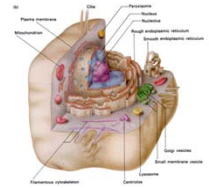


Cell Function Simulator

From genes to cells







Functional and computational genomics

- mRNA expression
- Splice variants
- Protein structure
- Protein expression
- Sub-cellular localization
- Protein-protein interactions
- Protein-DNA interactions

Design and interpretation of all of these assays requires sequence analysis.