

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>

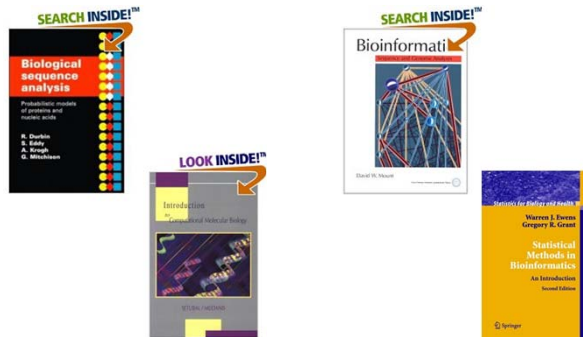
Reading lists:

<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.

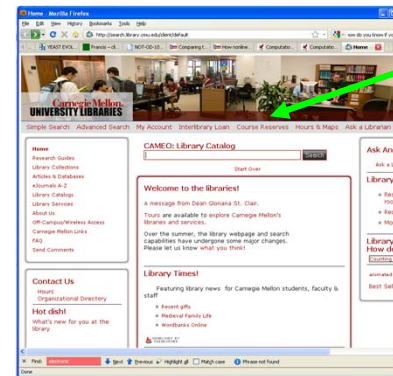
Course overview

Materials available on Electronic Reserves:

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

Electronic reserves:

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



Click on:

Course Reserves

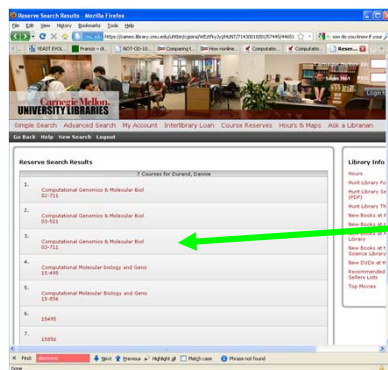
Select: Look up items on reserve by instructor

Type: "Durand"

Select: 03-711

Electronic reserves:

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



Click on:
Course Reserves

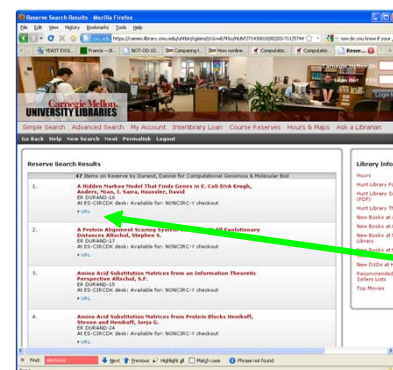
Select: Look up items on reserve by instructor

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

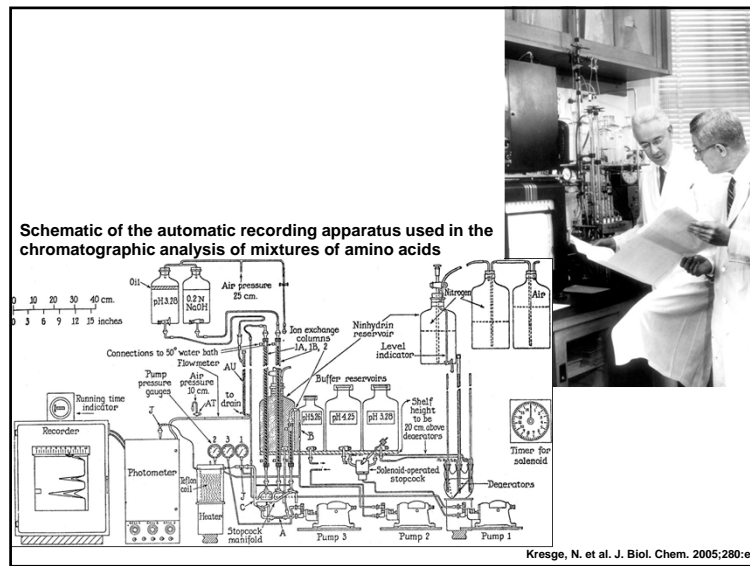
| | | |
|--|------|---|
| Sanger: peptide sequencing by partition chromatography | | Turing, v. Neumann: specs for stored program computer |
| Edman: stepwise protein degradation | | First transistor |
| Sanger sequences insulin, Discovery of DNA structure | 1950 | Edsac: 1 st stored program computer |
| | | Grace Murray Hopper: First compiler |
| Stein, Moore, Spackman: automatic amino acid analyzer | | Fortran |
| Myoglobin | 1960 | First integrated circuit |
| Ribonuclease | | |
| Lysozyme | | Basic |

The Origins of Computational Biology

| | | |
|--|--|---|
| Sanger: peptide sequencing by partition chromatography | | Turing, v. Neumann: specs for stored program computer |
| Edman: stepwise protein degradation | | First transistor |
| Sanger sequences insulin, Discovery of DNA structure | | Edsac: 1 st stored program computer |
| | | Grace Murray Hopper: First compiler |
| Stein, Moore, Spackman: automatic amino acid analyzer | | Fortran |
| Myoglobin | | First integrated circuit |
| Ribonuclease | | |
| Lysozyme | | Basic |



Trends in Biochem. Sci, 99



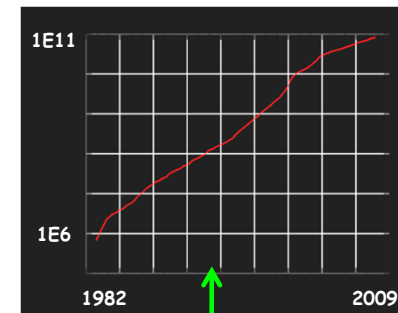
The Origins of Computational Biology

| | | |
|--|------|---|
| Sanger: peptide sequencing by partition chromatography | | Turing, v. Neumann: specs for stored program computer |
| Edman: stepwise protein degradation | | ENIAC |
| Sanger sequences insulin, Discovery of DNA structure | 1950 | First transistor |
| | | EDSAC: 1 st stored program computer |
| | | Grace Murray Hopper: First compiler |
| Stein, Moore, Spackman: automatic amino acid analyzer | | Fortran |
| Myoglobin | | First integrated circuit |
| Ribonuclease | 1960 | |
| Lysozyme | | Basic |

The Origins of Computational Biology

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

Genbank doubles every 18 months

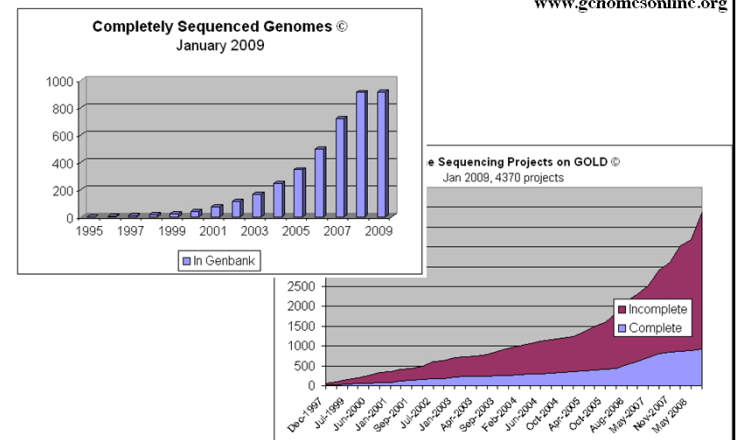


First whole genome sequence

Whole Genome Sequencing Highlights (A Eukarya-centric View)

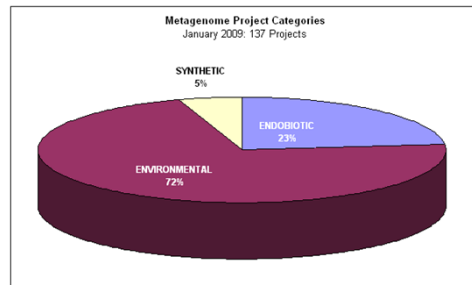
- 1995 *H. influenzae* – 1st whole genome sequence
- 1997 Yeast – 1st eukaryotic sequence
- 1998 *Caenorhabditis elegans* – 1st multicellular organism
- 2000 Fly, *Arabidopsis thaliana* – 1st plant
- 2001 Human
- 2002 Mouse, *Ciona intestinalis*,
- 2003 *Caenorhabditis briggsae*, *Neurospora Crassa*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

Whole Genome Sequencing



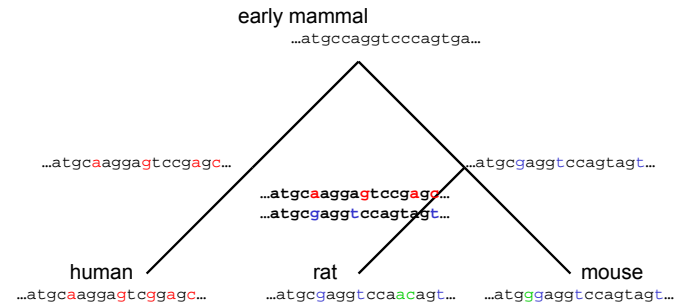
Metagenomics

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

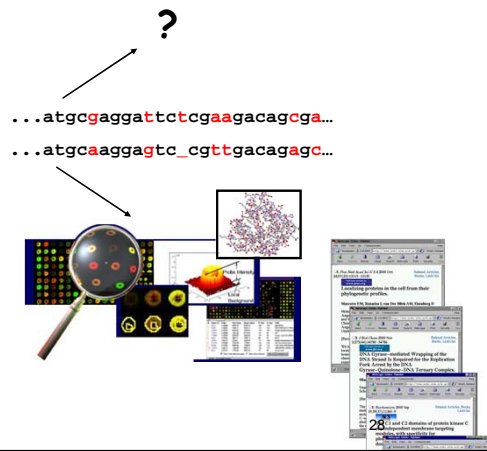


Why sequence data is so powerful:

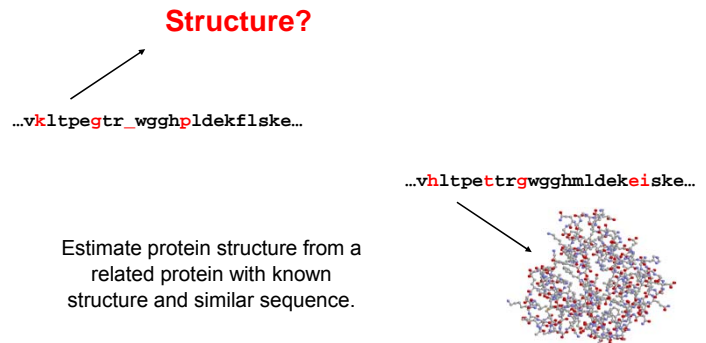
Sequences are related!



Sequence similarity → functional similarity



Sequence similarity → structural similarity

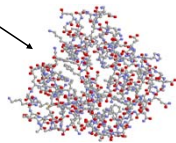


Sequence similarity \rightarrow structural similarity

Structure?

...v<ltpe<tr_wgghpldekflske...
...v<ltpet<trwgghmldekeiske...

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



Sequence Comparison

...atgcaaggagttcccagagcctgagctgactacgt...
...atgcgaggctctccagtgctgaactgactaagt...
global pairwise alignment

local pairwise alignment

```
tfsill v..frrda.h ksevahrfkd lgeenfkalv...
tfsill v..frrea.h kseiahrfnd vgeehfiglv...
tfsill v..frrdt.y kseiahrfkd lgeqyfgklv...
tlisfi lqrfardaeh kseiahrynd lkeetfkava...
```

```
...mkwvtfisll flfssaysrg v..frrda.h kseva
...mkwvtfisll flfssaysrg v..frrda.h kseia
...mkwvtfisll flfssaysrg v..frrdt.y kseia
...mkwvtfisfi flfssatsrn lqrfardaeh kseia
```

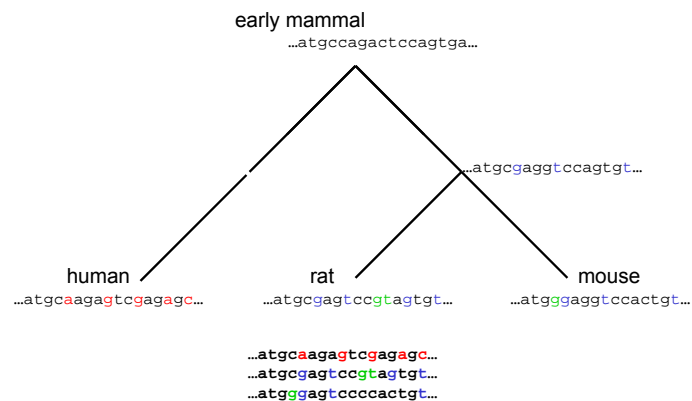
global multiple alignment

local multiple alignment

Applications

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

Reconstructing Evolutionary History



atatactcacagcat aactgttatataccacggggggcggaat gaaagcgttaacggcca
ggcaacaagaggtgttgatctcatcgtgtacatcacgcagacaggtatgccgcga
ctgctgcgggaaatcgcgagcgtttggggttcggttccccaaacgcggctgaagaacatc
tgaaggcgctggcgcgcaaggcgcttattgaaattgtttcggcgcatcacgcgggattc
gtctgttcggcaagaaggaagaagggttcgcgctggtaggtctgttggtgcgcgtgaac
cactcttggcgcaacagcatattgaaggtcattatcaggtcgatctctcttattcaagc
cgaaatgctgatttctcgtcgcgcgtcagcgggagtcgatgaaagatatcggcatattgg
atggtgacttctggtcagtcataaaaatcaggatgtaagtaacggtcaggtcgtttctgc
cagctattgatgacgaagtaccgttaagcgctcaaaaaacagcggaataaagtcgaac
tgttgcagaaaaatagcgagtttaaaccaattgtcgttgacctcgtgcagagagctca
cgctgcgggaaatcgcgagcgtttggggttcggttccccaaacgcggctgaagaacatc
tgaaggcgctggcgcgcaaggcgcttattgaaattgtttcggcgcatcacgcgggattc
gtctgttcggcaagaaggaagaagggttcgcgctggtaggtctgttggtgcgcgtgaac
ccaattgaaggctggcggttgggggtatttcgcaacgcgcgactgggtctgaacatatctctg
gaattcgataaaatctcgtggtttattgtgcagtttatggttccaaaatcgctcttttctgt
agaccgcgatacgcgcgtggcgctgcgggtttgtttttcatctctcttcacaggttgtct
cgatggcatctctcattcatctgataaagcactctggcattcgcttaccatgattt
gcaatgctgatttctcgtcgcgcgtcagcgggagtcgatgaaagatatcggcatattgg
atggtgacttctggtcagtcataaaaatcaggatgtaagtaacggtcaggtcgtttctgc
cagctattgatgacgaagtaccgttaagcgctcaaaaaacagcggaataaagtcgaac
tgttgcagaaaaatagcgagtttaaaccaattgtcgttgacctcgtgcagagagctca
ccaattgaaggctggcggttgggggtatttcgcaacgcgcgactggctgaacatatctctg
agaccgcgatacgcgcgtggcgctgcgggtttgtttttcatctctcttcacaggttgtct
cgatggcatctctcattcatctgataaagcactctggcattcgcttaccatgattt
tctccaatatcacggttcggtttcgtctgggactgtgtcgatcagcgcggaattgggtcatttg

DNA PATTERNS IN THE *E.coli* *lexA* GENE

Repressor binding site

Promotor sequences

TTCCAA -35

TATACT mRNAstart+ +10GGGGG Ribosomal binding site

ATG...TAA

open reading frame

```
1  gaattcgataaaatcctggtttattgtgcagtttaatggttccaaataatgcgctttttgctg
61  atatactcacagcataactgtataacacccagggggcgaatgaagcgcttaacggcca
120 TATACT mRNAstart+ +10GGGGG Ribosomal binding site
121 ggcaacaagagggtgtttgatctcatccgtgatcacatcagccagacaggtattccgcgca
181 cgcgtgcggaaatcgcgcagcgtttggggtttcgcttccccaaacgcgctgaagaatc
241 tgaaggcgctggcagcgaagcgcttattgaaattgtttccgcgcacatcagcgggatc
301 gtcctgttcaggaagaaggaagagggttgcgcgtgtaggtcgcttgctgcgcgtgaac
361 caactctggcgcaacagcatattgaaggtcattatcaggtcgatccttcttattcaacg
421 cgaatgctgatttctcgtgcgcgtcagcgggatgctcgatgaagatattcgccattatgg
481 atggtagcttctggcagctgcataaaactcaggatgtacgtaacggctcaggtcgtttg
541 cagctattgatgacgaagtaccgtttaagcgctgaaaaaacagggccaataaagtcgaac
601 tgttgccagaaaatagcaggtttaaaccaattgtcgttgaccttcgcagcagagctta
661 ccattgaagggctggcggttggggttattgcgaacggcagctgggtgtaacatatctctg
721 agaccgcgatgcgcgctggcgtgcgggtttgttttcaactctcttcaatcaggtctgt
781 gcattggcattctcactcaatctgataaagcactctggcatctgccttaccatgattt
841 tctccaatatcacggttcgctgctgggactggtcgatcagcgggtaattggctcatctg
901 atagcccggtttatttggggcgctggcggttggcgcacaggcggaccagct
```

[illegible]

Sequence conservation

Sequence conservation

True regulatory element Conserved gene Spurious gene prediction

S. cerevisiae

S. paradoxus

S. bayanus

S. mikatae

Predicted gene Conserved noncoding sequence Conserved gene sequence Non-matching sequence

Salzberg, Nature, 2003

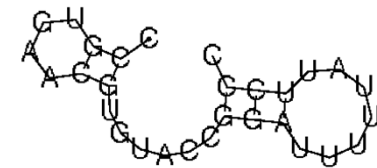
RNA Secondary Structure

RNA Secondary Structure

CCGUGUACGUGUAACGGAUCUUUAUUC...
 CCGUGAACGUAUACUGCAGUUUUAGUGCG...
 GGCUCACGCUGUCCGGAUUAUGAUUCCC
 CGCAGAAGCUCACGCGUUUUCUGUACGA

RNA Secondary Structure

CCGUGUACGUGUAACGGAUCUUUAUUC...
 CCGUGAACGUAUACUGCAGUUUUAGUGCG...
 GGCUCACGCUGUCCGGAUUAUGAUUCCC
 CGCAGAAGCUCACGCGUUUUCUGUACGA

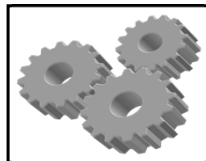


The Fantasy

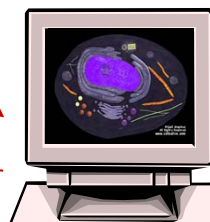
Whole genome sequence

GAAATAAACACCAGGCAGCAGTTATTAACACGGGAACATGGCGGCCGAGCCTGGGCTCCCGGGCGGCGGG...

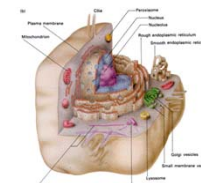
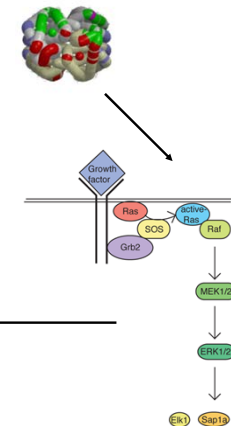
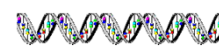
Cell Simulator Compiler



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.

