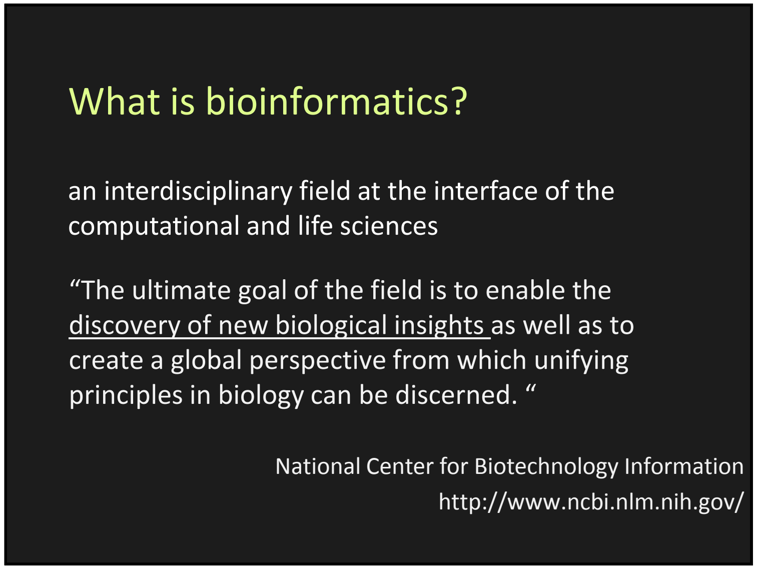
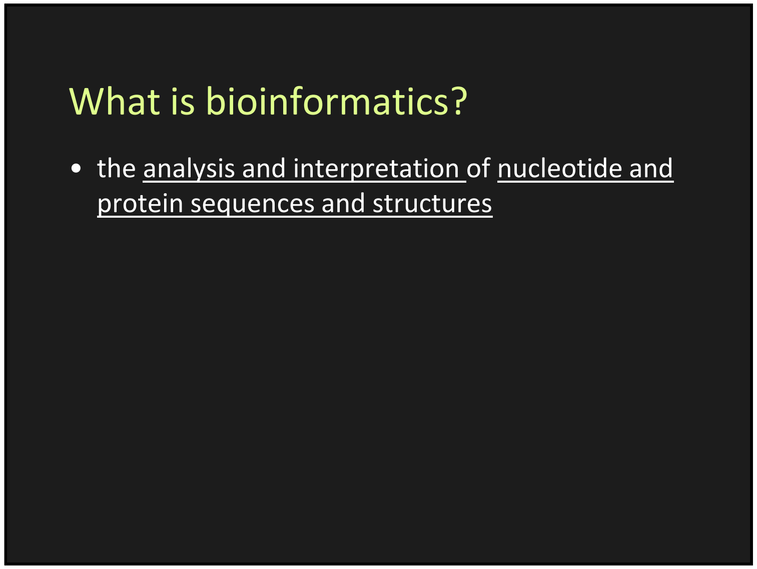


What is bioinformatics?

an interdisciplinary field at the interface of the computational and life sciences

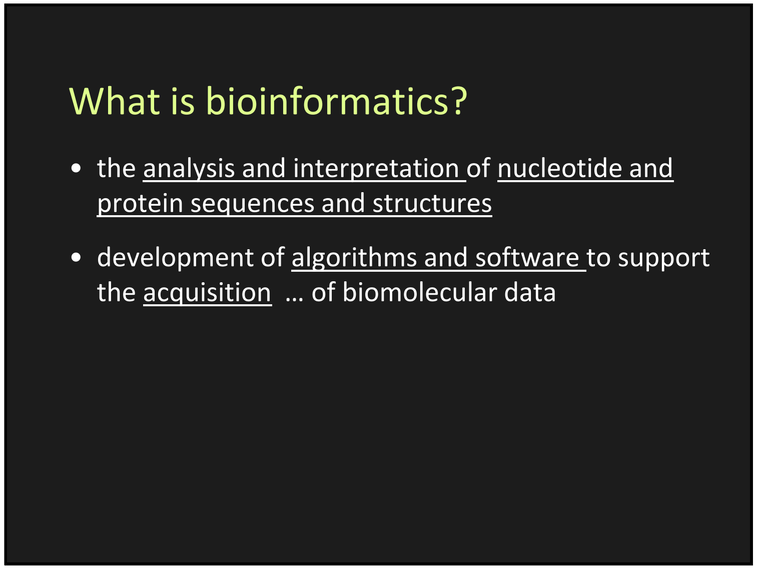
“The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. “

National Center for Biotechnology Information
http://www.ncbi.nlm.nih.gov/



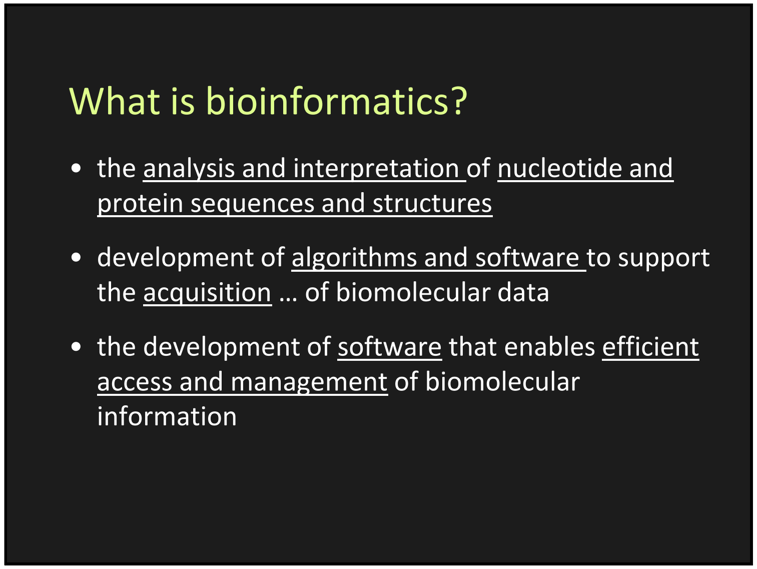
What is bioinformatics?

- the analysis and interpretation of nucleotide and protein sequences and structures



What is bioinformatics?

- the analysis and interpretation of nucleotide and protein sequences and structures
- development of algorithms and software to support the acquisition … of biomolecular data



What is bioinformatics?

- the analysis and interpretation of nucleotide and protein sequences and structures
- development of algorithms and software to support the acquisition … of biomolecular data
- the development of software that enables efficient access and management of biomolecular information

Bioinformatics stems from parallel revolutions in biology and computing

At the beginning of World War II (1939-1944):

- The shared program computer had not yet been invented, and there were no programming languages, databases, or computer networks.
- The relationship between genes and proteins, the molecular basis of genes, the structure of DNA and the genetic code were all unknown.

The Origins of Computational Biology

Amino acid sequencing

Sanger sequences i

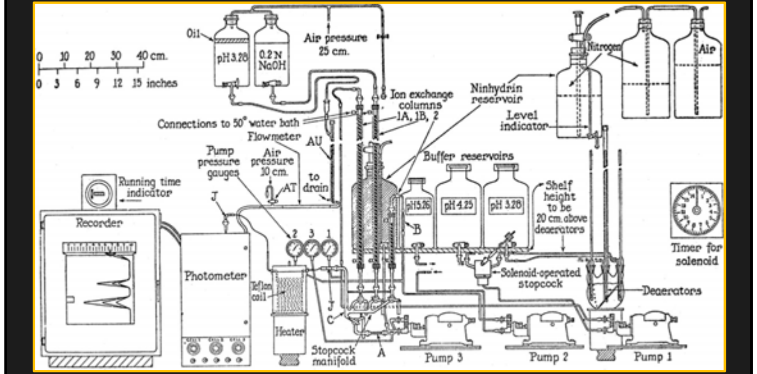
Stein, Moore, Spackn
automatic amino a
analyzer

Turing designs a stored
rogram computer

ac: 1st stored
rogram computer

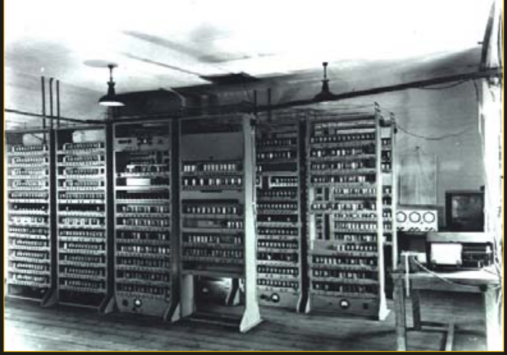
Trends in Biochem. Sci, 99

Automatic recording apparatus used in the chromatographic analysis of mixtures of amino acids



Stein, Moore, Spackman, 1958

EDSAC: The first stored-program computer.



The Origins of Computational Biology

Amino acid sequencing

Sanger sequences insulin.

Stein, Moore, Spackman: automatic amino acid analyzer

1950

Turing designs a stored program computer

Edsac: 1st stored program computer

Grace Murray Hopper co-invents COBOL and finds the first bug

Rear Admiral Grace Hopper over-seeing her team of programmers, Philadelphia Inquirer, 1957

Grace Murray Hopper finds the first bug

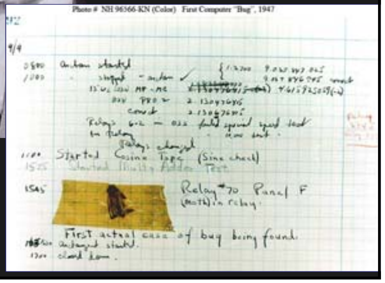


Photo # NH 96566-KN (Color) First Computer "Bug", 1947

9/9

0800 Arctan started
1000 " stopped - arctan ✓ {1.2700 9.037 847 025
9.037 846 995 conect
13"uc (032) MP-MC ~~(033)~~ 4.615925059(-2)
(033) PRO 2 2.130476415
conect 2.130676415
Relays 6-2 in 033 failed special speed test
In Relay " " 10.000 test.
Relays changed
1100 Started Cosine Tape (Sine check)
1525 Started Mult+ Adder Test.
1545 Relay #70 Panel F
(moth) in relay.
First actual case of bug being found.
1630 Antangent started.
1700 closed down.

Grace Murray Hopper

The Origins of Computational Biology

Amino acid sequencing

Sanger sequences insulin.

Stein, Moore, Spackman: automatic amino acid analyzer

1950

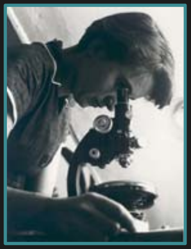
Turing designs a stored program computer

Edsac: 1st stored program computer

Grace Murray Hopper co-invents COBOL and finds the first bug

Discovery of DNA structure, Watson, Crick, Franklin

The helical structure of DNA



Rosalind Franklin

James Watson and Francis Crick

The Origins of Computational Biology

Determination of the genetic code

1960

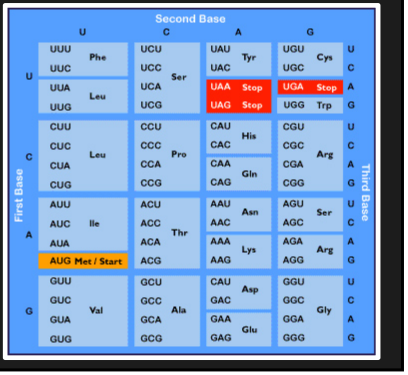
Fortran, Basic, LISP

1970

The genetic code

The genetic code is a degenerate, non-overlapping, triplet code. Crick, Barnett, Brenner, and Watts-Tobin, 1961

Determination of the genetic code



	Second Base				
First Base	U	C	A	G	Third Base
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC	UCC	UAC	UGC	C
	UUA Leu	UCA	UAA Stop	UGA Stop	A
	UUG	UCG	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA Gln	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA	AAA Lys	AGA Arg	A
	AUG Met / Start	ACG	AAG	AGG	G
G	GUU Val	GCU Ala	CAU Asp	GGU Gly	U
	GUC	GCC	GAC	GGC	C
	GUA	GCA	GAA Glu	GGA	A
	GUG	GCG	GAG	GGG	G

The Origins of Computational Biology

Determination of the genetic code

1960

Fortran, Basic, LISP

On going protein sequencing, Dayhoff publishes the Protein Atlas

1970

Atlas of Protein Sequence & Structure 1965 - 1978

Margaret Dayhoff

PhD in Chemistry, 47

Watson Computing Lab Fellow 47 - 48

The Origins of Computational Biology

Determination of the genetic code

Fortran, Basic, LISP

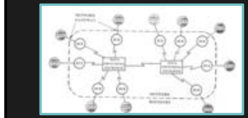
1960

On going protein sequencing, Dayhoff publishes the Protein Atlas

Arpanet

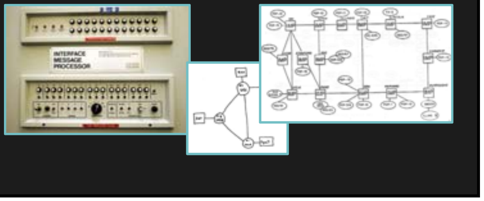
1970

The ARPAnet is established, precursor to the Internet.



Packet switching is invented, which supports flexible, multi-node network

Interface Message Processor (IMP) network is constructed, linking 4, and then 15 nodes



The Origins of Computational Biology

Determination of the genetic code

Fortran, Basic, LISP

1960

On going protein sequencing, Dayhoff publishes the Protein Atlas

Arpanet

The Protein Data Bank is established.

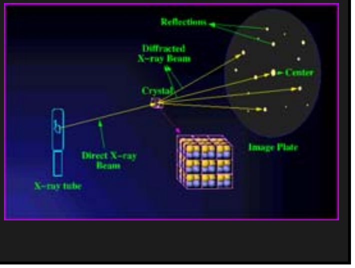
1970

Protein Data Bank (PDB)

Growing collection of X-ray diffraction protein structure data

Development of molecular graphics display for 3D visualization

SEARCH program supports remote access



The Origins of C

ational Biology

Sanger-Coulson sequen

Maxam-Gilbert sequenc

Gilbert, Sanger win Nob
Prize

TCP/IP

Internet

First royal email

USENET newsgroups

The Origins of Computational Biology

1970

Sanger-Coulson sequencing

Maxam-Gilbert sequencing

TCP/IP

Internet

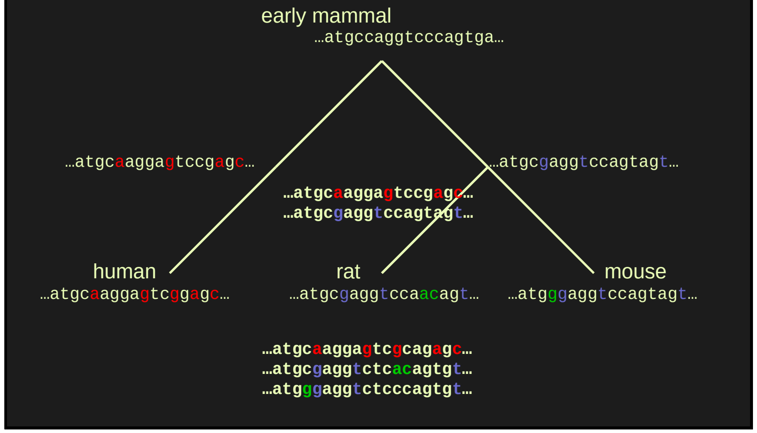
Gilbert, Sanger win Nobel
Prize

1980

First royal email

USENET newsgroups

Beginnings of molecular evolution



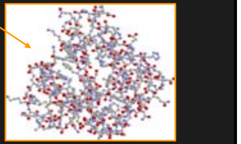
Sequence similarity → structural similarity

Structure?

…vkltpegtr_wgghpldekflske…

…vhltpettrgwgghmldekeiske…

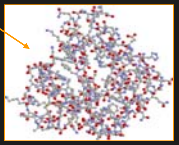
Estimate protein structure from a related protein with known structure and similar sequence.



Sequence similarity → structural similarity

Structure?

…vkltpegtr_wgghpldekflske…
…vhltpettrgwgghmldekeiske…

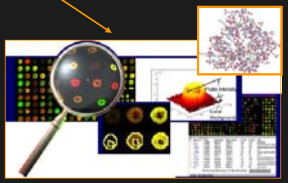


Estimate protein structure from a related protein with known structure and similar sequence.

Sequence similarity → functional similarity

?

...atgcgaggattctcgaagacagcga…
...atgcaaggagtc_cgttgacagagc…



The Origins of Computational Biology

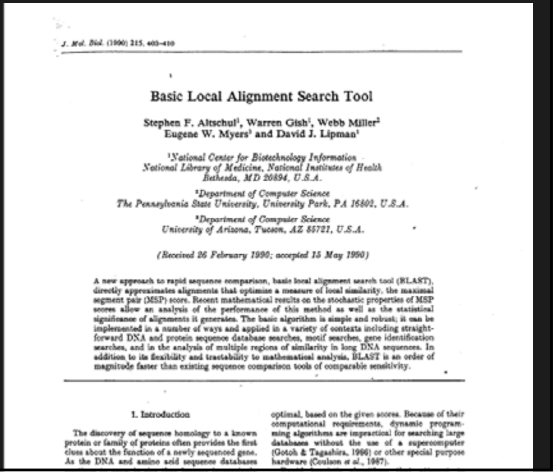
1990

Congress establishes Genbank

Human Genome Project begins (DOE)

Basic local alignment search tool (BLAST)

BLAST



J. Mol. Biol. (1990) 215, 403–410

Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

[1]National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, U.S.A.

[2]Department of Computer Science
The Pennsylvania State University, University Park, PA 16802, U.S.A.

[3]Department of Computer Science
University of Arizona, Tucson, AZ 85721, U.S.A.

(Received 26 February 1990; accepted 15 May 1990)

A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than existing sequence comparison tools of comparable sensitivity.

1. Introduction

The discovery of sequence homology to a known protein or family of proteins often provides the first clues about the function of a newly sequenced gene. As the DNA and amino acid sequence databases

optimal, based on the given scores. Because of their computational requirements, dynamic programming algorithms are impractical for searching large databases without the use of a supercomputer (Gotoh & Tagashira, 1986) or other special purpose hardware (Coulson et al., 1987).

The Origins of Computational Biology

Congress establishes Genbank

Human Genome Project begins (DOE)

Basic local alignment search tool (BLAST)

1990

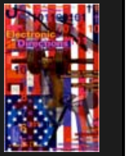
Information superhighway

World Wide Web,

NCSA Mosaic

The internet we know and love

Al Gore Creates Bill to Fund "Information Superhighway"



Tim Berners-Lee proposes a design for information sharing that becomes the World Wide Web

The first web browser

The Origins of Computational Biology

Congress establishes Genbank

Human Genome Project begins (DOE)

Basic local alignment search tool (BLAST)

GenBank goes online.

1990

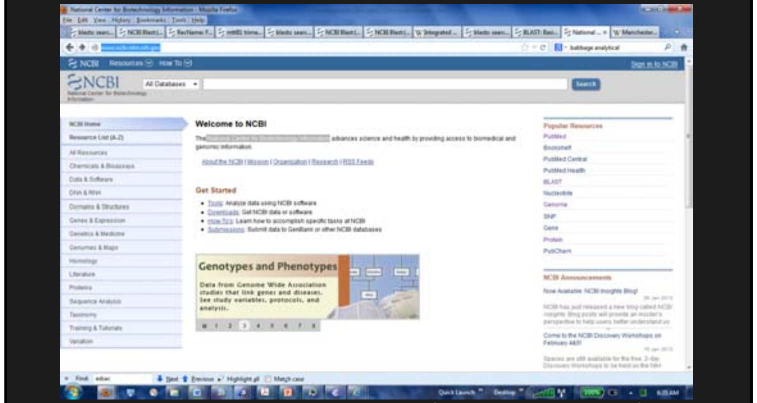
Information superhighway

World Wide Web,

NCSA Mosaic

Pizza Hut goes on line

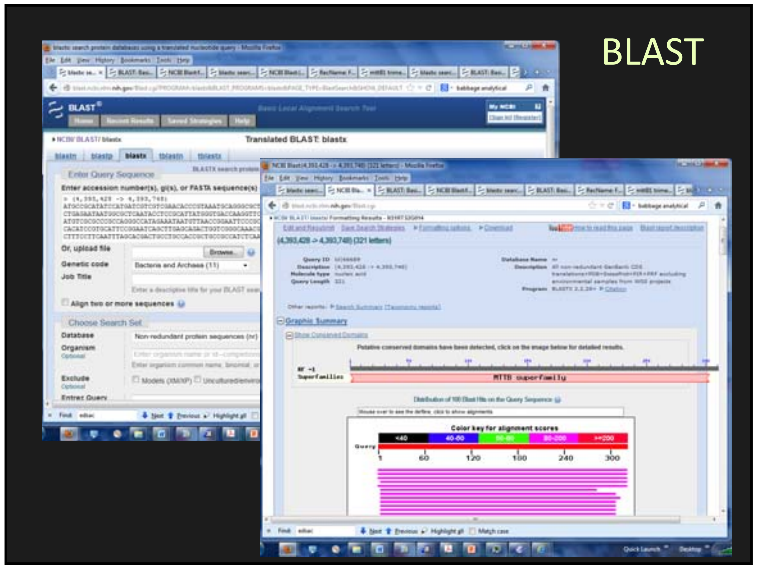
Genbank



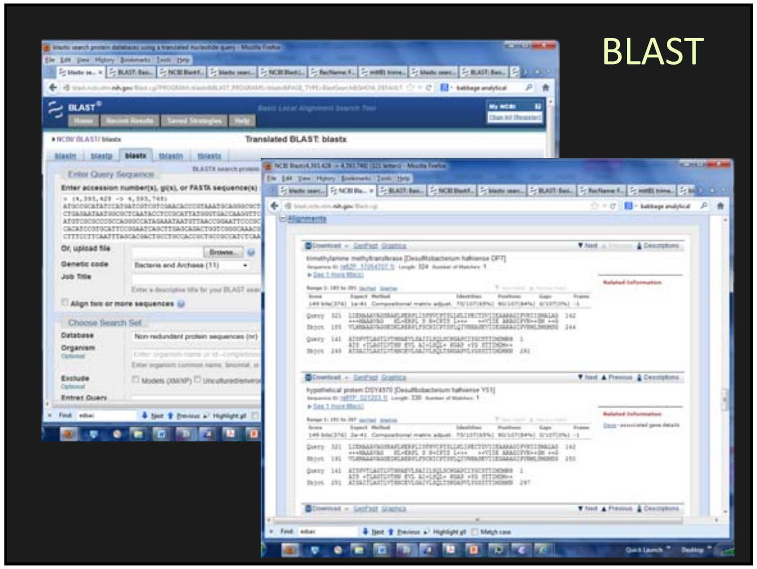
National Center for Biotechnology Information

http://www.ncbi.nlm.nih.gov/

BLAST



BLAST



The Origins of Computational Biology

Google

*H. Influenzae: 1*st whole genome sequence **1995**

Yeast – 1st eukaryote

1998

C. elegans - 1st animal

IBM's Blue Gene architecture for protein modeling

Fly genome **1999**

A.thaliana – 1st plant

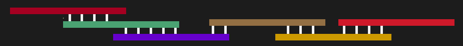
2000

Sequence Assembly

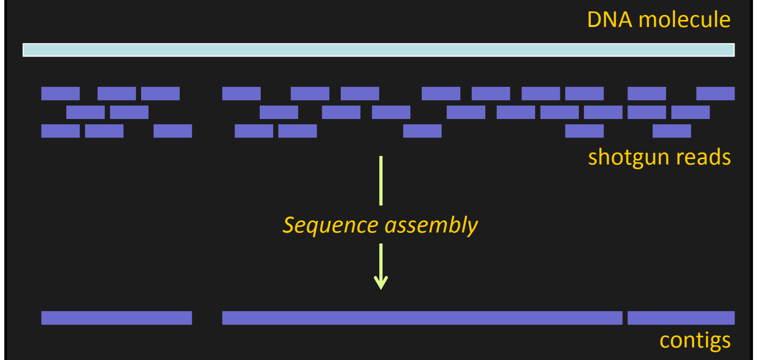
Limits of gel electrophoresis: ~ 500bp in one "read"

To sequence more than 500 bp:

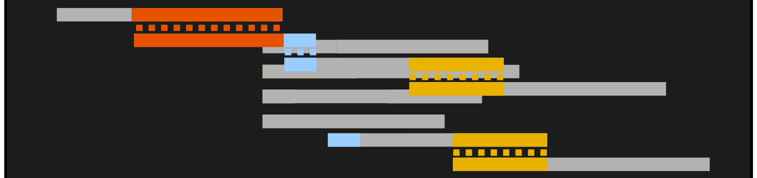
- Sequence 500bp fragments separately
- Combine *computationally* using sequence comparison



Shotgun sequencing



Sequence Assembly



[illegible] ence identity

The Origins of Computational Biology

2001

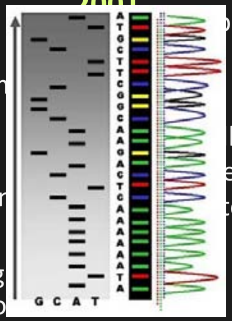
Draft human genom

ABI Capillary sequer

Chimp genome, alig
with human geno

pedia, Captcha

la Foundation:
en source, W3P
tent policy



How to interpret whole genome sequence

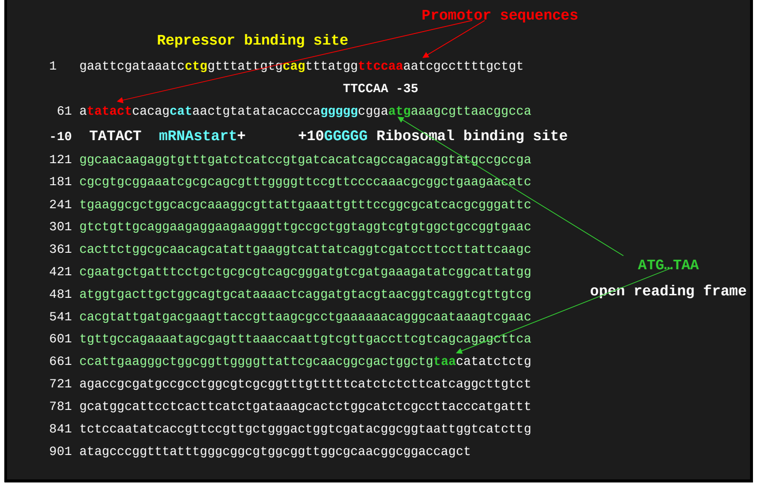
- Where are the genes?
- When and where are those genes expressed?
- What proteins do they encode?
- What do they do?
  - Molecular function?
  - Biological pathway or process?

How to interpret whole genome sequence

- ➢ Where are the genes?
- When and where are those genes expressed?
- What proteins do they encode?
- What do they do?
  - Molecular function?
  - Biological pathway or process?

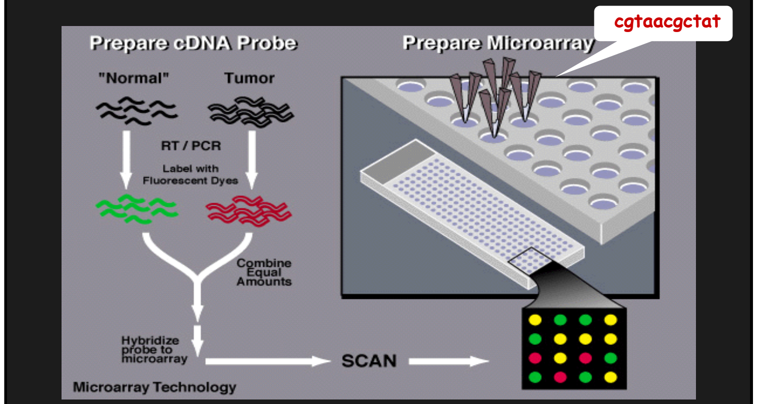
```
atatactcacagcataactgtatatacacccagggggcggaatgaaagcgttaacggcca
ggcaacaagaggtgtttgatctcatccgtgatcacatcagccagacaggtatgccgccga
cgcgtgcggaaatcgcgcagcgtttggggttccgttccccaaacgcggctgaagaacatc
tgaaggcgctggcacgcaaaggcgttattgaaattgtttccggcgcatcacgcgggattc
gtctgttgcaggaagaggaagaagggttgccgctggtaggtcgtgtggctgccggtgaac
cacttctggcgcaacagcatattgaaggtcattatcaggtcgatccttccttattcaagc
cgaatgctgatttcctgctgcgcgtcagcgggatgtcgatgaaagatatcggcattatgg
atggtgacttgctggcagtgcataaaactcaggatgtacgtaacggtcaggtcgttgtcg
cacgtattgatgacgaagttaccgttaagcgcctgaaaaaacagggcaataaagtcgaac
tgttgccagaaaatagcgagtttaaaccaattgtcgttgaccttcgtcagcagagcttca
cgcgtgcggaaatcgcgcagcgtttggggttccgttccccaaacgcggctgaagaacatc
tgaaggcgctggcacgcaaaggcgttattgaaattgtttccggcgcatcacgcgggattc
gtctgttgcaggaagaggaagaagggttgccgctggtaggtcgtgtggctgccggtgaac
ccattgaagggctggcggttggggttattcgcaacggcgactggctgtaacatatctctg
gaattcgataaatctctggtttattgtgcagtttatggttccaaaatcgccttttgctgt
agaccgcgatgccgcctggcgtcgcggtttgtttttcatctctcttcatcaggcttgtct
gcatggcattcctcacttcatctgataaagcactctggcatctcgccttacccatgattt
cgaatgctgatttcctgctgcgcgtcagcgggatgtcgatgaaagatatcggcattatgg
atggtgacttgctggcagtgcataaaactcaggatgtacgtaacggtcaggtcgttgtcg
cacgtattgatgacgaagttaccgttaagcgcctgaaaaaacagggcaataaagtcgaac
tgttgccagaaaatagcgagtttaaaccaattgtcgttgaccttcgtcagcagagcttca
ccattgaagggctggcggttggggttattcgcaacggcgactggctgtaacatatctctg
agaccgcgatgccgcctggcgtcgcggtttgtttttcatctctcttcatcaggcttgtct
gcatggcattcctcacttcatctgataaagcactctggcatctcgccttacccatgattt
tctccaatatcaccgttccgttgctgggactggtcgatacggcggtaattggtcatcttg
```

DNA PATTERNS IN THE *E.coli* lexA GENE



```
1   gaattcgataaatcctggtttattgtgcagtttatggttccaaaatcgccttttgctgt
                                              TTCCAA -35
 61 atatactcacagcataactgtatatacacccagggggcggaatgaaagcgttaacggcca
-10  TATACT  mRNAstart+      +10GGGGG Ribosomal binding site
121 ggcaacaagaggtgtttgatctcatccgtgatcacatcagccagacaggtatgccgccga
181 cgcgtgcggaaatcgcgcagcgtttggggttccgttccccaaacgcggctgaagaacatc
241 tgaaggcgctggcacgcaaaggcgttattgaaattgtttccggcgcatcacgcgggattc
301 gtctgttgcaggaagaggaagaagggttgccgctggtaggtcgtgtggctgccggtgaac
361 cacttctggcgcaacagcatattgaaggtcattatcaggtcgatccttccttattcaagc
421 cgaatgctgatttcctgctgcgcgtcagcgggatgtcgatgaaagatatcggcattatgg
481 atggtgacttgctggcagtgcataaaactcaggatgtacgtaacggtcaggtcgttgtcg
541 cacgtattgatgacgaagttaccgttaagcgcctgaaaaaacagggcaataaagtcgaac
601 tgttgccagaaaatagcgagtttaaaccaattgtcgttgaccttcgtcagcagagcttca
661 ccattgaagggctggcggttggggttattcgcaacggcgactggctgtaacatatctctg
721 agaccgcgatgccgcctggcgtcgcggtttgtttttcatctctcttcatcaggcttgtct
781 gcatggcattcctcacttcatctgataaagcactctggcatctcgccttacccatgattt
841 tctccaatatcaccgttccgttgctgggactggtcgatacggcggtaattggtcatcttg
901 atagcccggtttatttgggcggcgtggcggttggcgcaacggcggaccagct
```

DNA microarrays



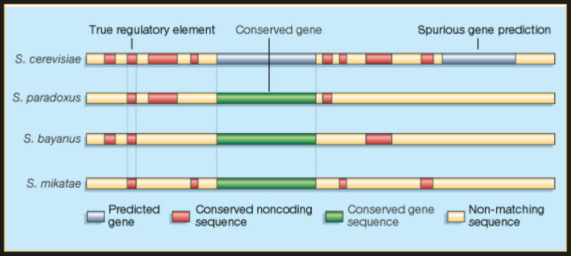
The Origins of Computational Biology

5 more yeast genomes **2004**

2005

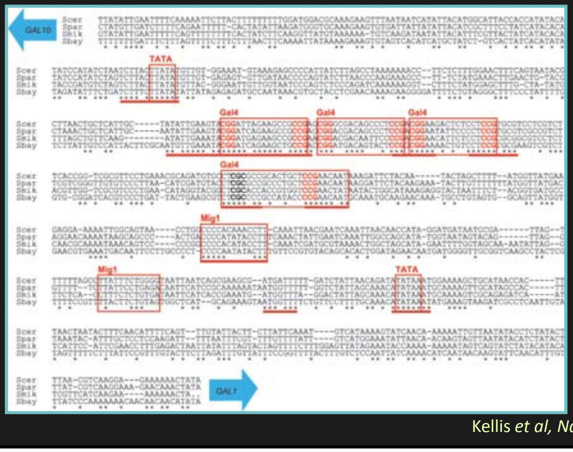
2006

Align genomes to confirm gene predictions and identify regulatory regions

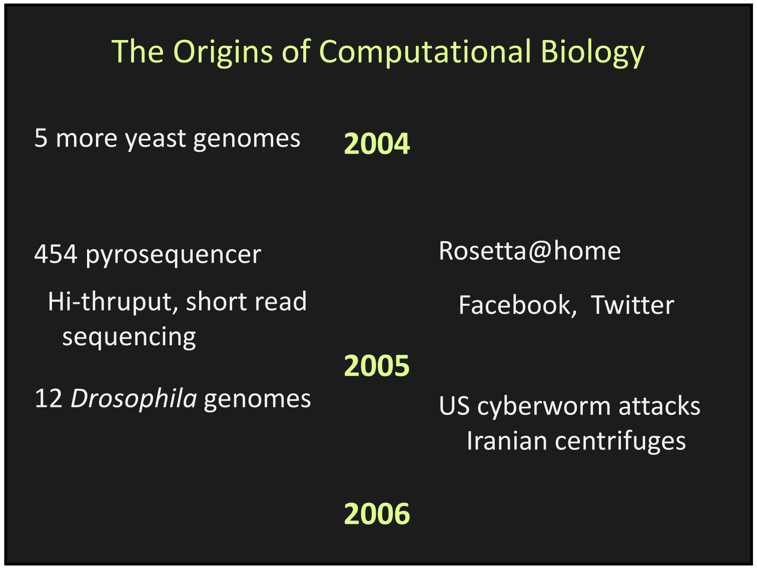


Salzberg, Nature, 2003

Global alignment of upstream sequences to identify regulatory regions



Kellis *et al*, *Nature*, 03



The Origins of Computational Biology

5 more yeast genomes **2004**

454 pyrosequencer

Hi-thruput, short read sequencing

Rosetta@home

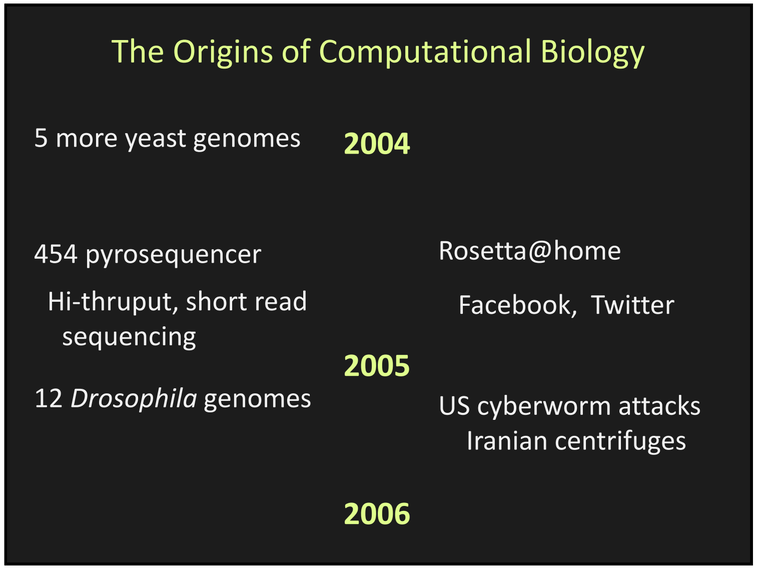
Facebook, Twitter

2005

12 *Drosophila* genomes

US cyberworm attacks Iranian centrifuges

2006



The Origins of Computational Biology

5 more yeast genomes **2004**

454 pyrosequencer

Hi-thruput, short read sequencing

Rosetta@home

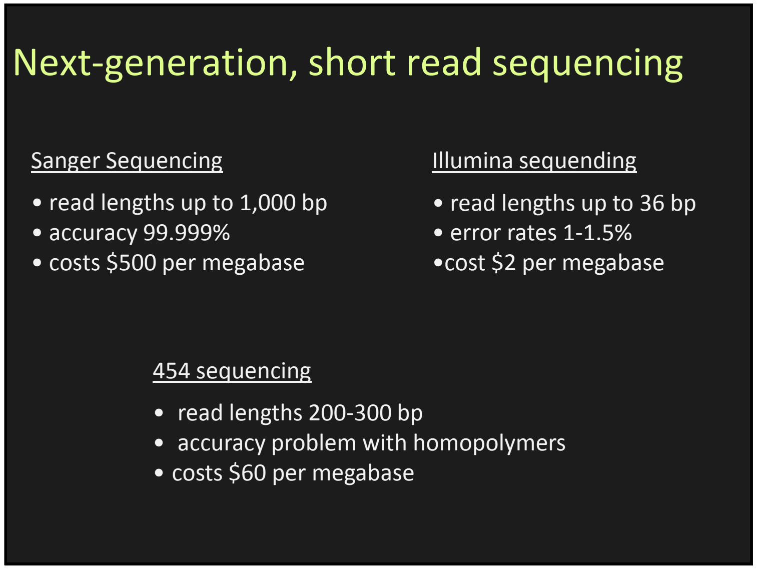
Facebook, Twitter

2005

12 *Drosophila* genomes

US cyberworm attacks Iranian centrifuges

2006



Next-generation, short read sequencing

Sanger Sequencing

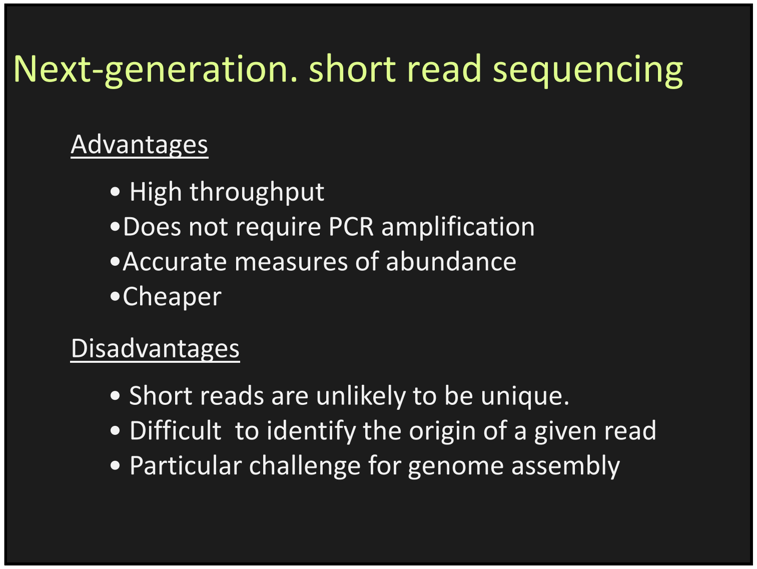
- read lengths up to 1,000 bp
- accuracy 99.999%
- costs $500 per megabase

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
- Cheaper

Disadvantages

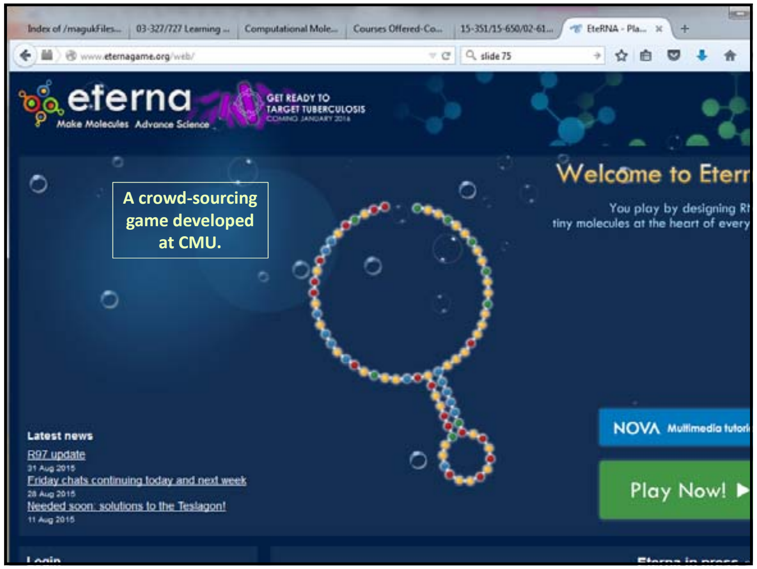
- Short reads are unlikely to be unique.
- Difficult to identify the origin of a given read
- Particular challenge for genome assembly

Some next generation sequencing applications

- Bacterial genomes
- Sample diversity in a bacterial population (e.g., your throat when you have strep)
- Transcription: more accurate and quantitative compared with microarrays
- Medical diagnostics: sequence short genomic regions to identify mutations associated with disease

The Origins of Computational Biology

1000 Genomes project Human microbiome project.	**2007**	Apple iPhone Estonia: First national elections via Internet
First tumor/normal genome published Draft Neanderthal genome	**2008**	Foldit: Crowd-sourced protein folding game
	2009	



Welcome to Eterr

You play by designing R
tiny molecules at the heart of every

Latest news

R97 update
31 Aug 2015

Friday chats continuing today and next week
28 Aug 2015

Needed soon: solutions to the Teslagon!
11 Aug 2015

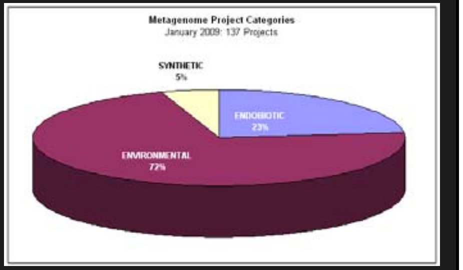
A crowd-sourcing game developed at CMU.

Metagenomics

- Sample communities of microbial organisms directly from their natural environments, bypassing the need for isolation and lab cultivation of individual species.
- Result: a collection of DNA fragments that characterize the organismal and functional diversity of the envirment

Metagenomics

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



What makes us human?

- Human metabolic features- combo of human and microbial traits
- Microbiota- microrganisms that live inside and on humans
- Microbiome- the genomes of the microbial symbionts

The Origins of Computational Biology

2010

Chocolate (*Theobroma cacao*) genome

Social networking topples regime in Egypt

2011

3rd Generation sequencing: Pac Bio, Ion Torrent

Crystal structure …solved by protein folding game players, Nature Structural Biology

2012

What is bioinformatics?

Development of algorithms and software to support the acquisition and interpretation of biomolecular data

- Acquisition: Microarray design, Sequence assembly
- Interpretation: Sequence comparison, clustering of microarrays, gene finding, phylogenetic profiling, …

What is bioinformatics?

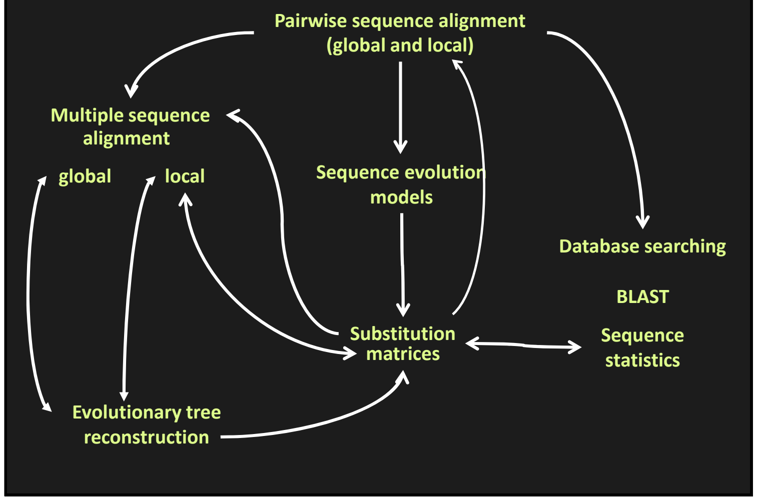
Development of software that enables efficient access and management of biomolecular information

- Genbank
- BLAST
- Protein visualization

What is bioinformatics?

The analysis and interpretation of nucleotide and protein sequences and structures

Application of all of these methods to address specific questions about specific biological systems


Pairwise sequence alignment
(global and local)
Multiple sequence alignment
global
local
Sequence evolution models
Database searching
BLAST
Substitution matrices
Sequence statistics
Evolutionary tree reconstruction