Intro to Intro

What is Computational Biology?

Using

- computation
- algorithms
- mathematics
- statistics
- visualization
- databases
- simulations

to help understand

- biology
- molecules
- reactions
- cells
- systems
- disease
- evolution

In Bio131, we will use algorithms to help understand biological sequences

The Quest to Sequence the Human Genome

http://mpnforum.com/genetics-crash-mpn-course/

The Quest to Sequence the Human Genome

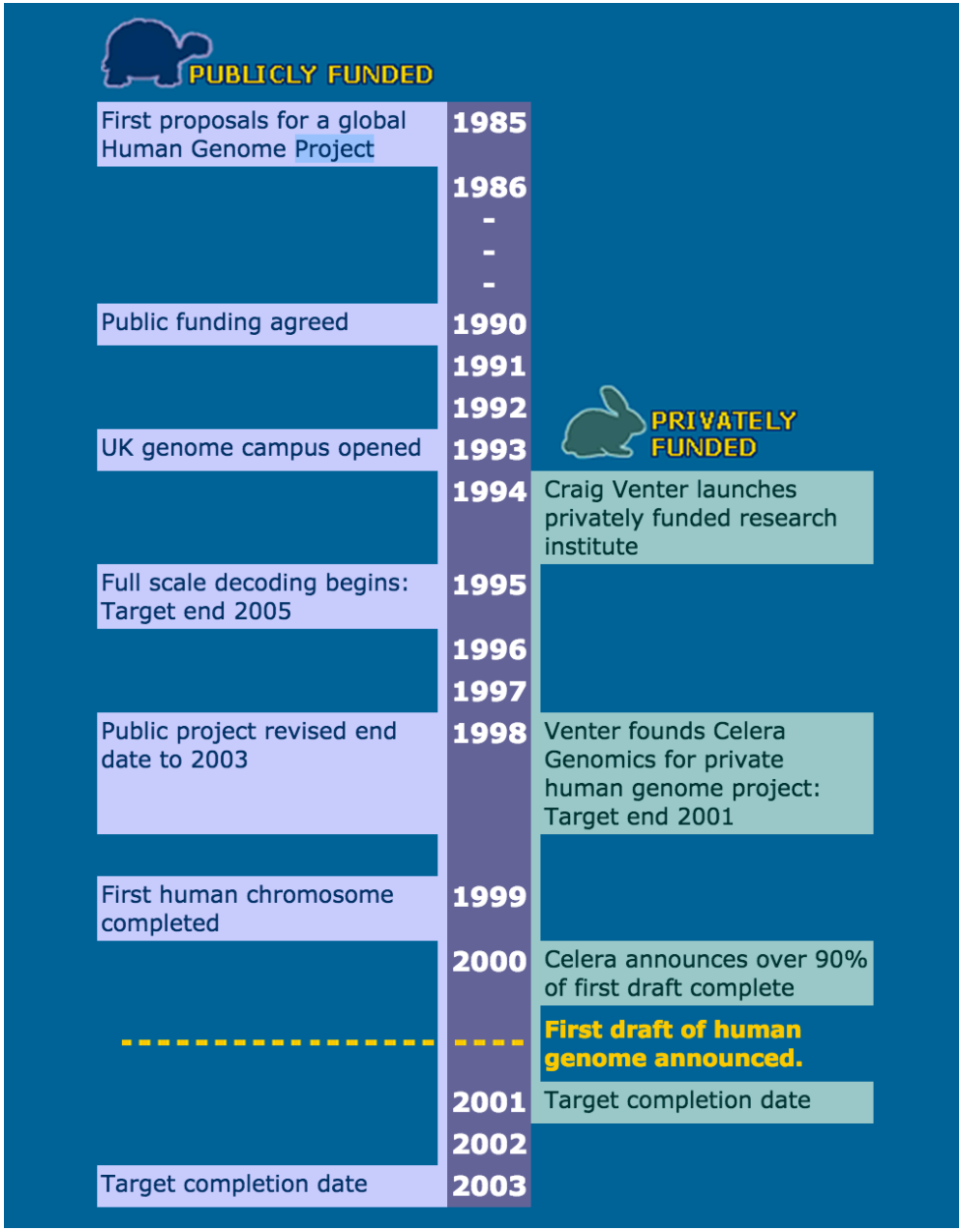
"Regardless of who wins this sprint, the next race--to make sense of the genome--will be a marathon with many runners…Just the first step, which researchers call annotation, could very well take many months." – Karow, 2000

ENCyclopedia Of DNA Elements (ENCODE) Project

2003 ENCODE Project Established

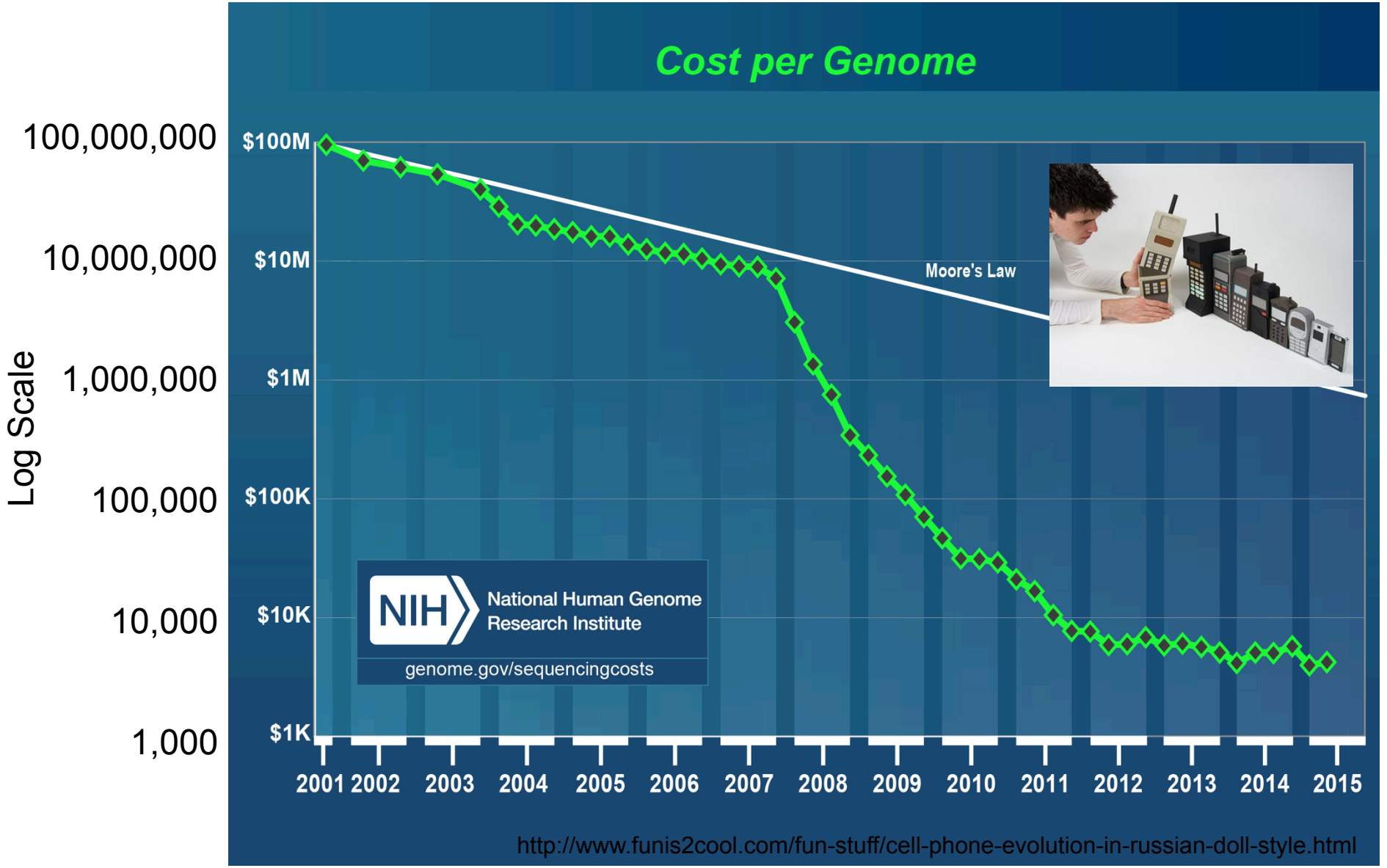
2007 ENCODE Pilot Completed

2012 ENCODE Project Completed



The Human Genome Race, Julia Karow, Scientific American, 2000.
http://www.genome.gov/10001763
http://news.bbc.co.uk/hi/english/static/in_depth/sci_tech/2000/human_genome/timeline/default.stm

A Deluge of Data



A Deluge of Data

adsp
Alzheimer's Disease Sequencing Project

1000 Plant Genomes

Autism Genome 10K

T1K
Fish-T1K
Transcriptomes of 1000 Fishes

GENOME

International Cancer Genome Consortium

HMP
NIH HUMAN MICROBIOME PROJECT

NIH
THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

Example: Teleost Fishes

Wikipedia

Example: Teleost Fishes

SCIENTIFIC DATA

OPEN

Data Descriptor: Whole genome sequencing data and *de novo* draft assemblies for 66 teleost species

Martin Malmstrøm[1], Michael Matschiner[1], Ole K. Tørresen[1], Kjetill S. Jakobsen[1] & Sissel Jentoft[1,2]

Received: 14 September 2016
Accepted: 07 December 2016
Published: 17 January 2017

Teleost fishes comprise more than half of all vertebrate species, yet genomic data are only available for 0.2% of their diversity. Here, we present whole genome sequencing data for 66 new species of teleosts, vastly expanding the availability of genomic data for this important vertebrate group. We report on *de novo* assemblies based on low-coverage (9–39×) sequencing and present detailed methodology for all analyses. To facilitate further utilization of this data set, we present statistical analyses of the gene space completeness and verify the expected phylogenetic position of the sequenced genomes in a large mitogenomic context. We further present a nuclear marker set used for phylogenetic inference and evaluate each gene tree in relation to the species tree to test for homogeneity in the phylogenetic signal. Collectively, these analyses illustrate the robustness of this highly diverse data set and enable extensive reuse of the selected phylogenetic markers and the genomic data in general. This data set covers all major teleost lineages and provides unprecedented opportunities for comparative studies of teleosts.

Example: The Canadian Beaver

De Novo Genome and Transcriptome Assembly of the Canadian Beaver (*Castor canadensis*)

Si Lok,*,†,1 Tara A. Paton,*,† Zhuozhi Wang,*,† Gaganjot Kaur,
Wilson W. L. Sung,*,† Joseph Whitney,*,† Janet A. Buchanan,*
Beverly Apresto,*,† Nan Chen,*,† Matthew Coole,*,† Travis J.
Sanjeev Pullenayegum,*,† Kozue Samler,*,† Arun Shipstone,*
Sergio L. Pereira,*,† Pirooz Rostami,*,† Carol Ann Ryan,*,† Ar
Yogi Sundaravadanam,§ Jared T. Simpson,§,** Burton K. Lim
Christopher J. Dutton,‡‡ Kevin C. R. Kerr,‡‡ Maria Franke,‡‡
and Stephen W. Scherer§§,***,1

*The Centre for Applied Genomics and †Program in Genetics and Genc to,
Ontario M5G 0A4, Canada, §§McLaughlin Centre, Faculty of Medicine, U
‡Donnelly Centre for Cellular and Biomolecular Research, University of Toro for
Cancer Research, MaRS Centre, Toronto, Ontario M5G 0A3, Canada, **De to,
Ontario M5S 3G4, Canada, ††Royal Ontario Museum, Toronto, Ontario M5S 2C6, Canada, ‡‡Toronto Zoo, Ontario M1B 5K7,
Canada, and ***Department of Molecular Genetics, Faculty of Medicine, University of Toronto, Ontario M5S 1A8, Canada

G3 Genes | Genomes | Genetics Volume 7 February 2017 | Genome Assembly and Annotation of Canadian Beaver (*C. canadensis*)

Example: An Ancient Genome [Spring 2016]

Neolithic woman from Ballynahatty, Ireland (~5,000 years ago)
- DNA suggests Near Eastern origin

Bronze-age individuals from similar region (~4,000 years ago)
- DNA suggests European origin

https://www.washingtonpost.com/news/speaking-of-science/wp/2015/12/29/ancient-irish-genome-reveals-a-massive-migration-from-the-east/

Example: The Octopus Genome [Fall 2015]

Compared to Human:

- Genome Size: 2.7Gb vs. 3.0Gb
- # of Protein Coding Genes: 33K vs. 25K
- # of Genes Related to Neuron Regulation: 168 vs. 70
- # of Neurons: 500M vs. ~20B (mouse has 4M)

http://www.nature.com/news/octopus-genome-holds-clues-to-uncanny-intelligence-1.18177

Overview of Topics

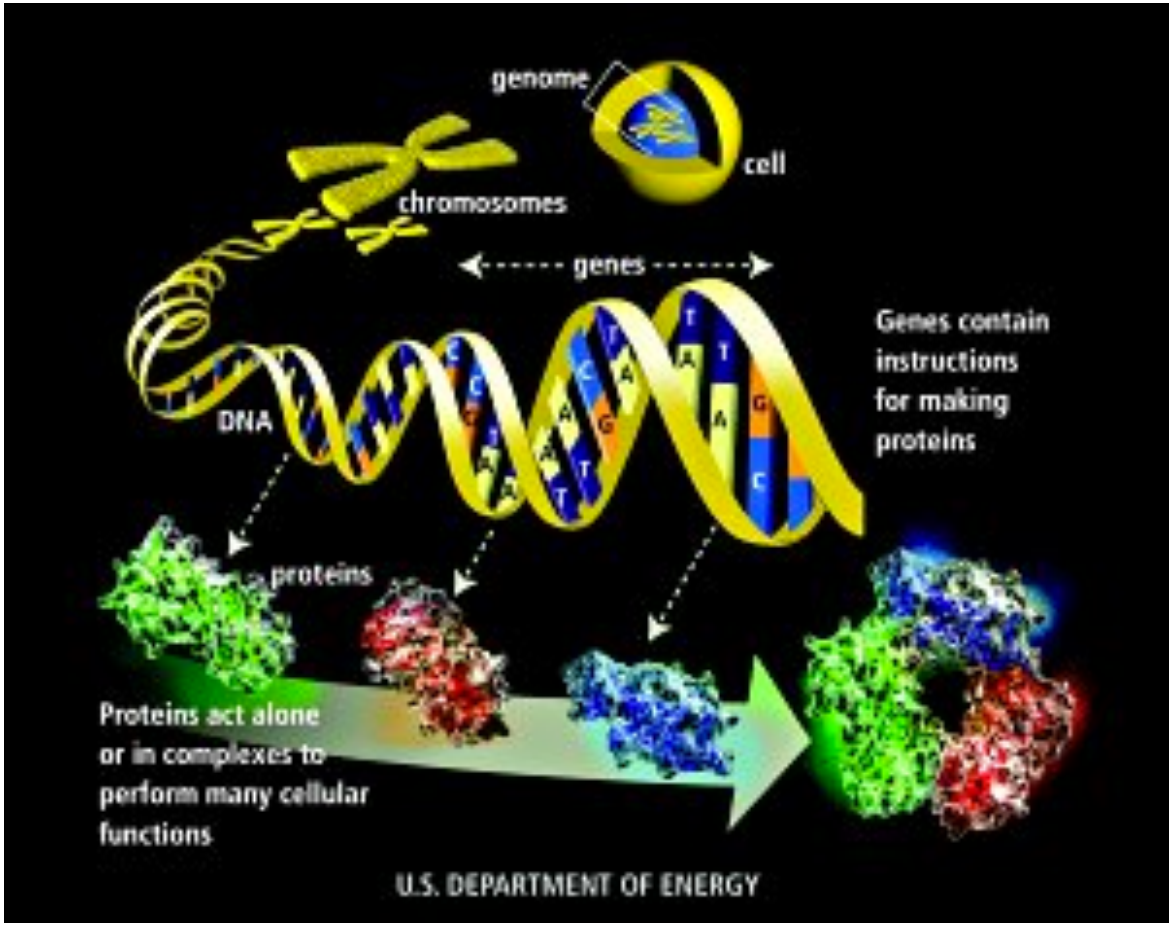
- **Topic 1**: How do cells "read" the genome?
- **Topic 2**: Where do DNA-binding proteins bind?
- **Topic 3**: How do we compare biological sequences?
- **Topic 4**: How do we assemble genomes?
- **Topic 5**: How do we compare entire genomes?

All of these topics involve biological sequence analysis!

Topic 1: How do Cells "Read" the Genome?

3.72×10^{13} cells =
37,200,000,000,000
= 37.2 **trillion** cells

Each **cell** has the same genome with 3 **billion** bases
= 3,000,000,000



http://www.ncbi.nlm.nih.gov/pubmed/23829164

http://science.kqed.org/quest/2007/07/09/a-wonderful-mess/

Topic 1: Program the Central Dogma

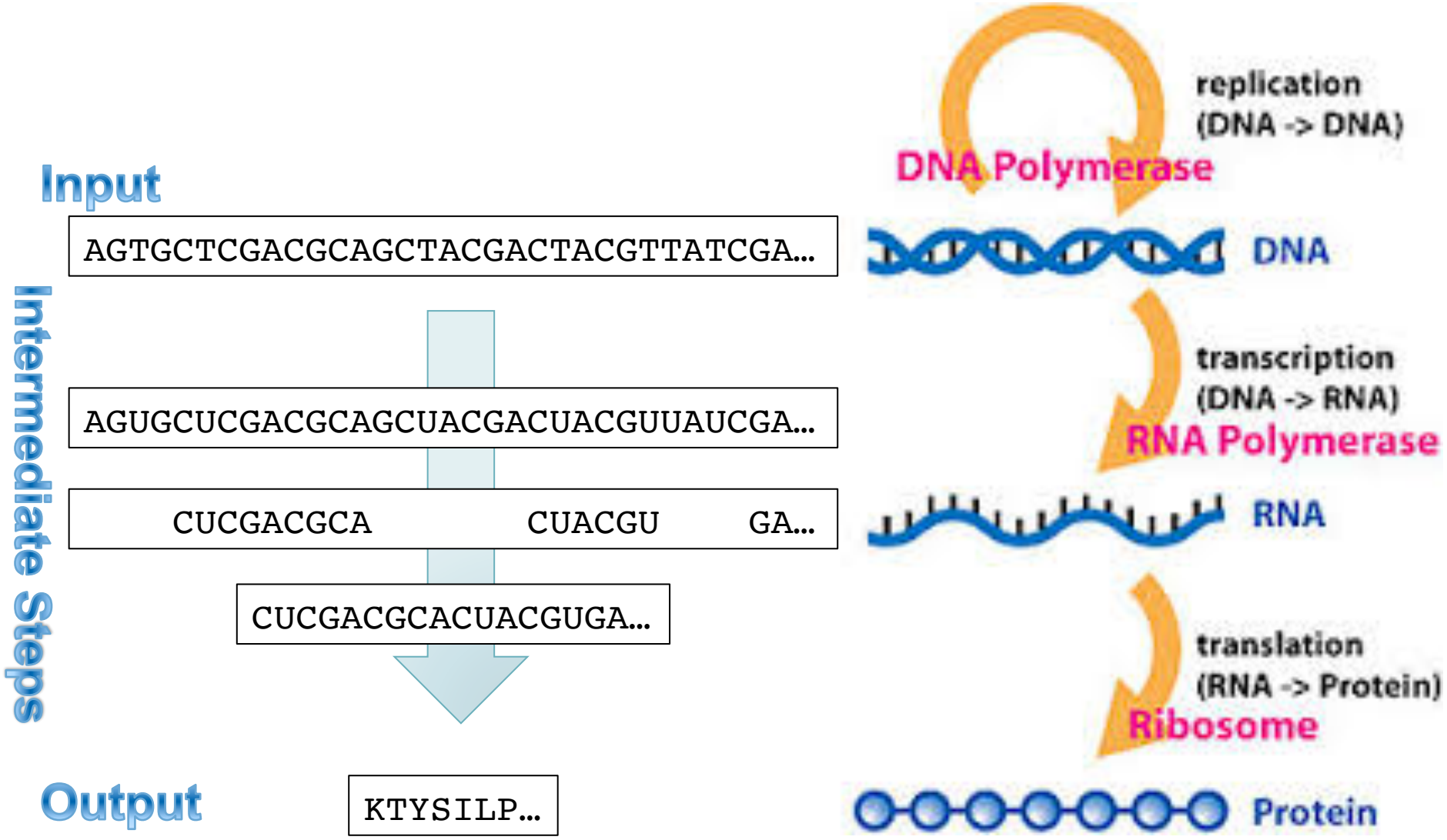


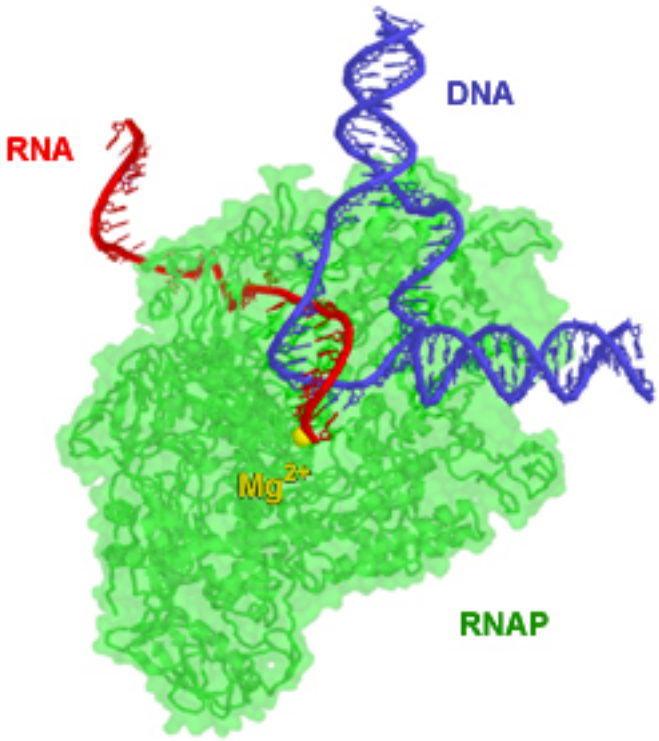
Image from Wikipedia

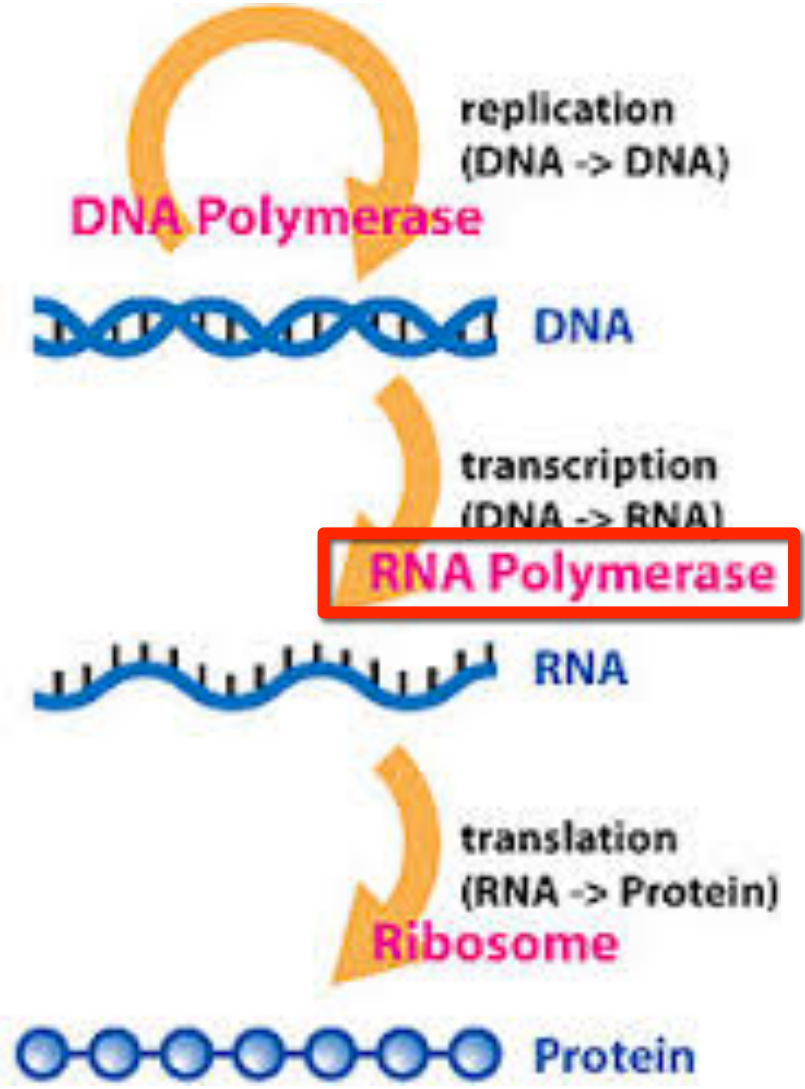
By the end of…

Topic 1:

Write a Python program to convert **any** gene to its protein sequence.

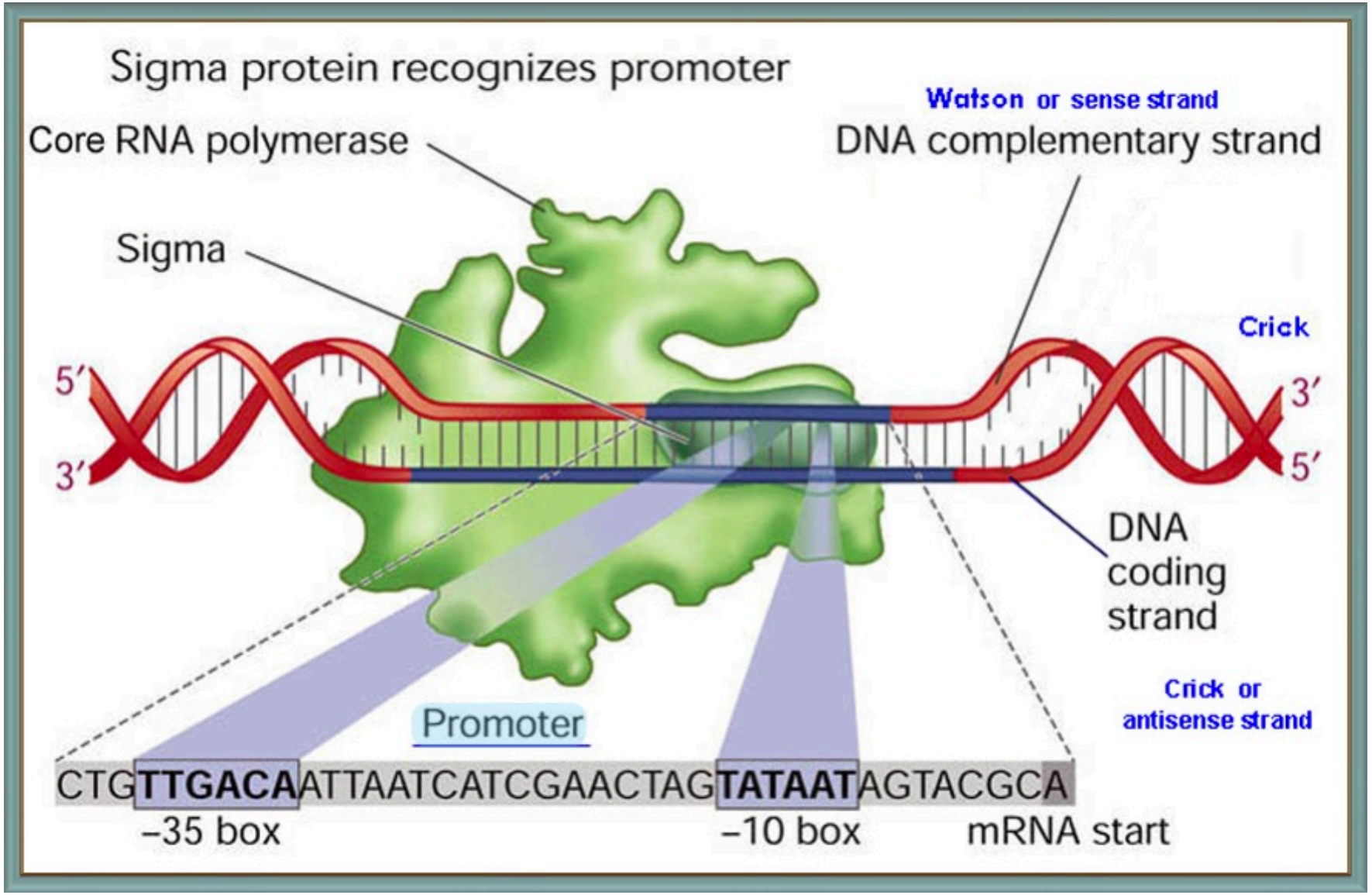
Topic 2: Where do DNA-binding Proteins Bind?





wikipedia; http://www.bio.miami.edu/~cmallery/150/gene/mol_gen.htm

Topic 2: Where do DNA-binding Proteins Bind?



wikipedia; http://www.bio.miami.edu/~cmallery/150/gene/mol_gen.htm

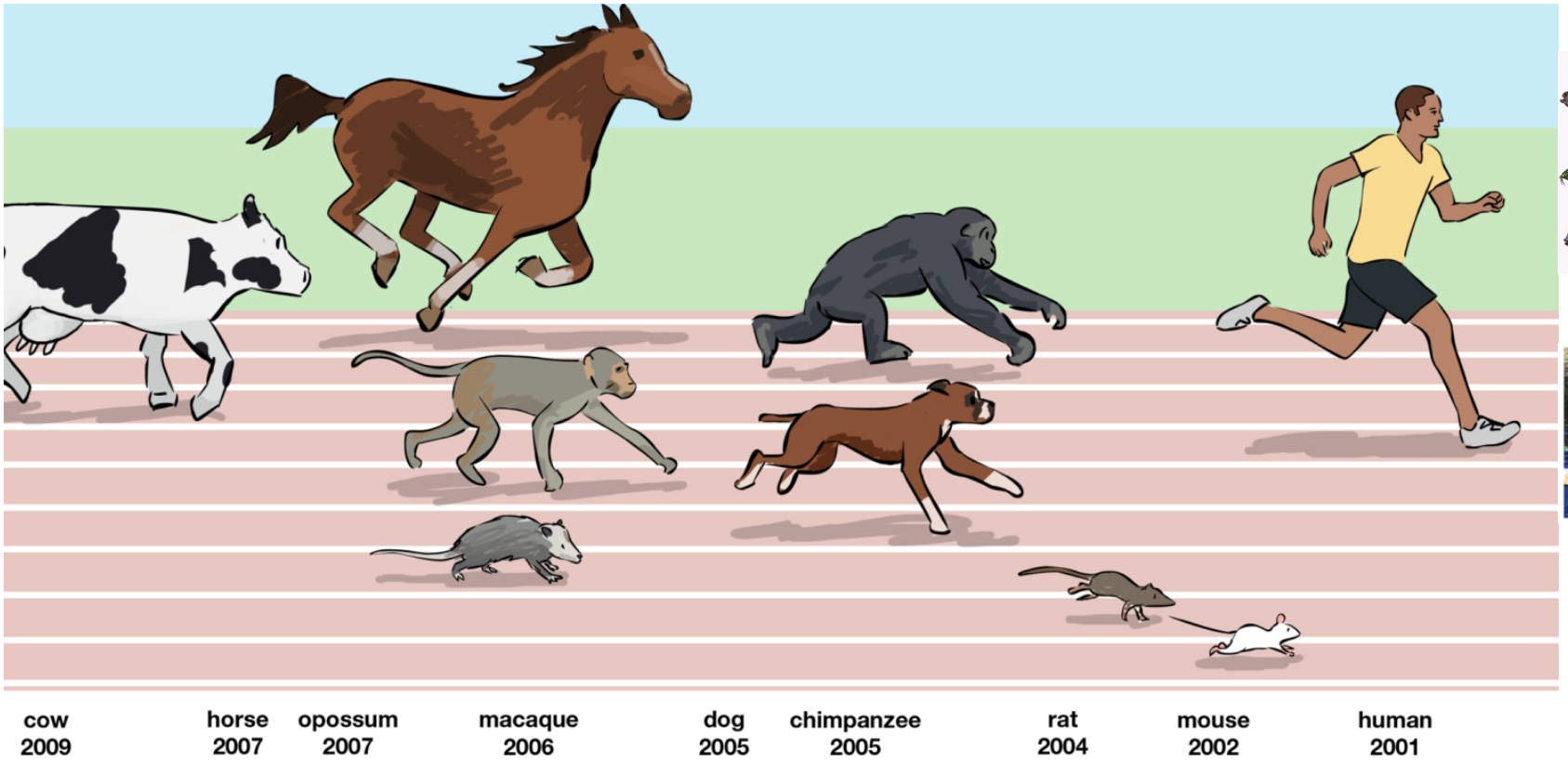
By the end of…

Topic 1: Write a Python program to convert **any** gene to its protein sequence.

Topic 2: Write a Python program to identify and summarize motifs from a collection of sequences.

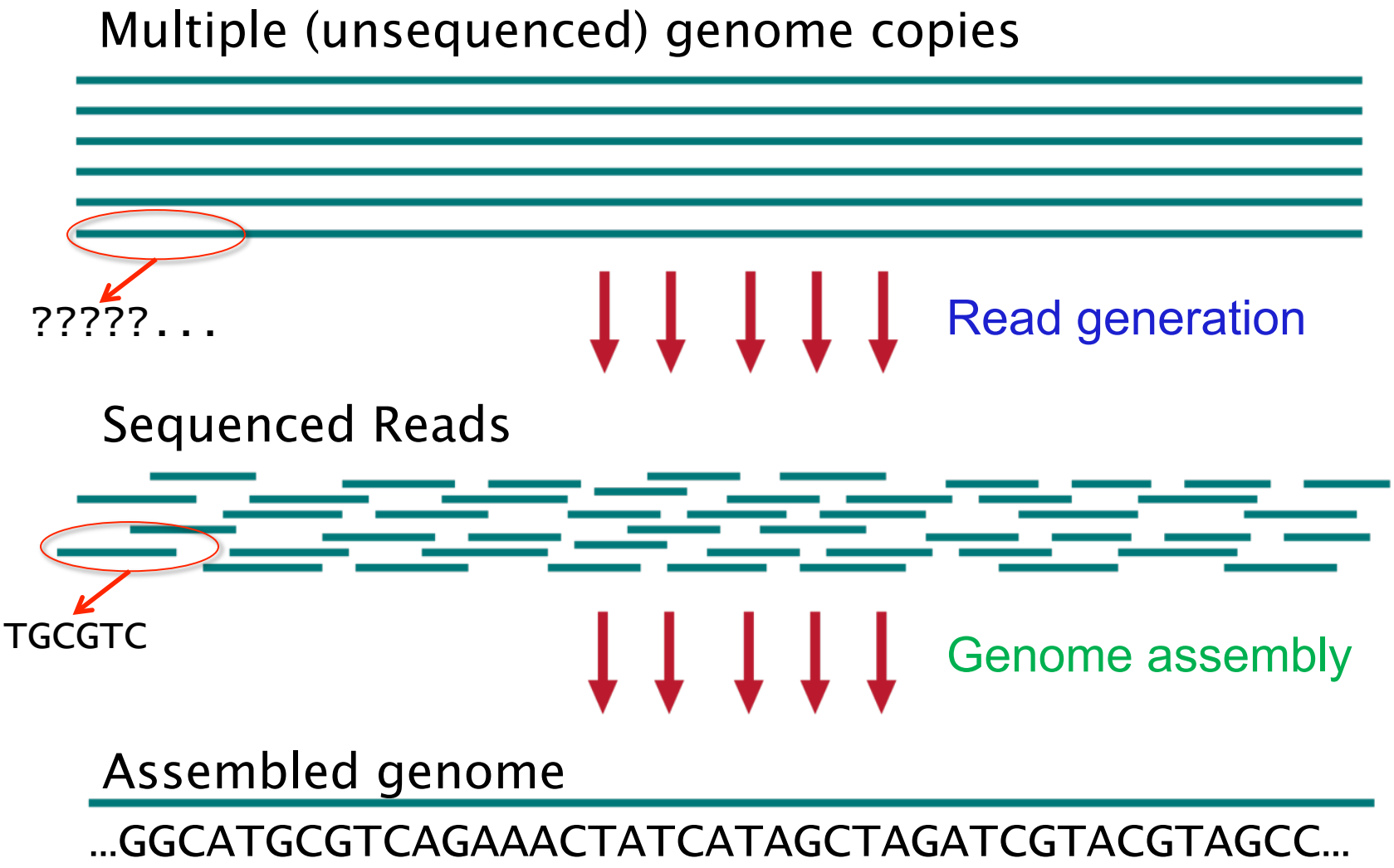
Topic 5: Assembling Genomes "From Scratch"

- How was the first human genome assembled?
- How are the genomes of other species assembled?



2017

Topic 5: Assembling Genomes "From Scratch"



By the end of…

Topic 1: Write a Python program to convert **any** gene to its protein sequence.

Topic 2: Write a Python program to identify and summarize motifs from a collection of sequences.

Topic 3: Write a Python program to simulate an assembler

Topic 4: How do we compare sequences?

(hypothetical)

Human Gene: ACTCGACTGAGAGGATTTCGAGCATGA

Mouse Gene: ACTCAACTGAGATTCGAGCTTCAATGA

```
ACTCGACTGAGAGGATTTCGAGCATGA
||||  |||||||        |    ||||
ACTCAACTGAGATTCGAGCTTCAATGA
```

11 substitutions

If we “shift” the sequences by adding gaps, can we improve the number of substitions?

Finding an Optimal Sequence Alignment

By the end of…

Topic 1: Write a Python program to convert **any** gene to its protein sequence.

Topic 2: Write a Python program to identify and summarize motifs from a collection of sequences.

Topic 3: Write a Python program to simulate an assembler

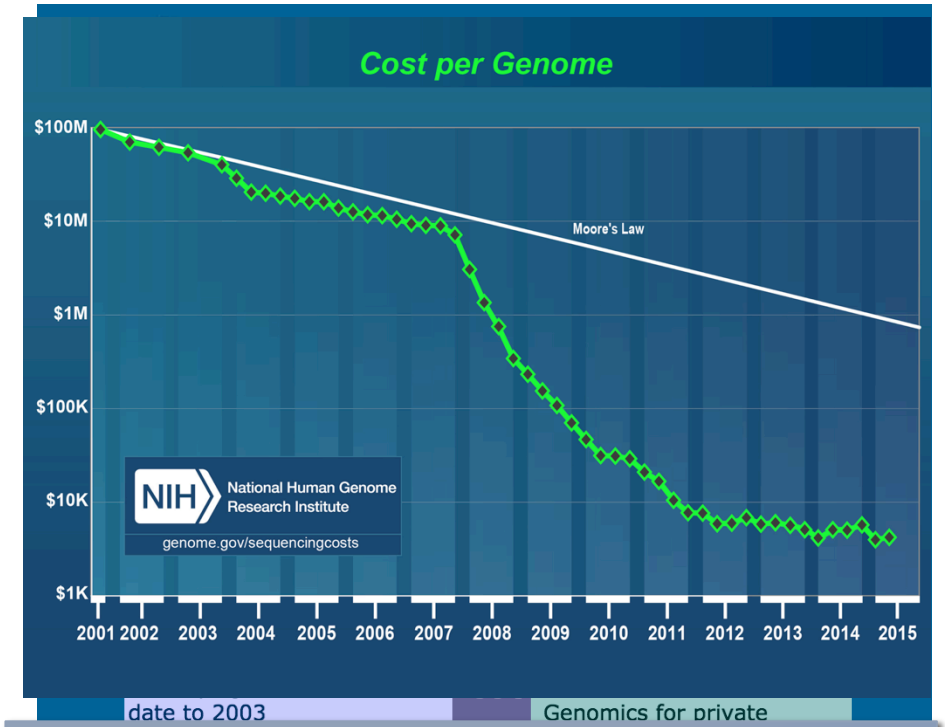
Topic 4: Write multiple Python programs to align pairs of DNA or protein sequences.

Topic 4: How do we compare genomes?

Human Genome Project
- $2.7 billion
- About 6 years

Celera Genomics
- $300 million
- About 3 years

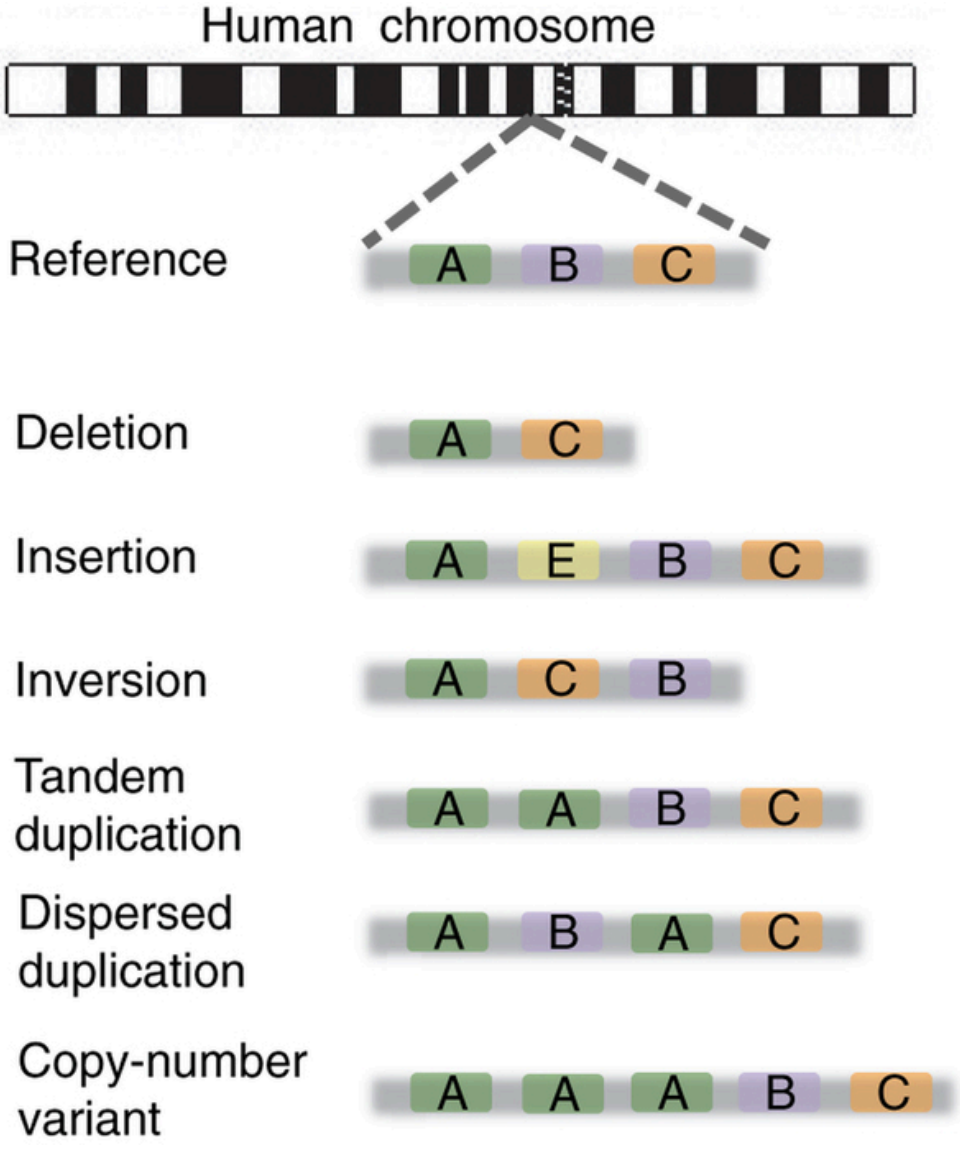
Today:
- $1,000 (announced 2015)
- $4,000 typically
- A week (I'm generous)



Compare the new genome to the already-sequenced *reference* genome

http://www.genome.gov/10001763
http://news.bbc.co.uk/hi/english/static/in_depth/sci_tech/2000/human_genome/timeline/default.stm

Topic 5: Genome Rearrangements



http://www.nature.com/nmeth/journal/v9/n2/full/nmeth.1858.html

By the end of…

Topic 1: Write a Python program to convert **any** gene to its protein sequence.

Topic 2: Write a Python program to identify and summarize motifs from a collection of sequences.

Topic 3: Write a Python program to simulate an assembler

Topic 4: Write multiple Python programs to align pairs of DNA or protein sequences.

Topic 5: Visualize large genomic rearrangements between two genomes.

By the end of…

Topic 1: Write a Python program to convert **any** gene to its protein sequence.

Topic 2: Write a Python program to identify and summarize motifs from a collection of sequences.

Topic 3: Write a Python program to simulate an assembler

Topic 4: Write multiple Python programs to align pairs of DNA or protein sequences.

Topic 5: Visualize large genomic rearrangements between two genomes.

Topic 6: Student-Selected Topics & Final Projects

About This Class

- Hybrid biology/computer science class
- We will learn Python!

- Course Components
  - 8 graded programming assignments (Collaboration Policy)
  - 2 in-class exams on biology/computational concepts
  - 2 code conferences
  - Labs handed in for participation points
  - Final project on a topic of your choice

- More information on the syllabus

About This Class

- Moodle & Website
  - https://moodle.reed.edu/course/view.php?id=1251
  - http://reed.edu/biology/courses/bio131/

- Textbook
  - *Bioinformatics Algorithms: an Active Learning Approach*
  - Edition 2, Volume 1
  - Pevzner & Compeau
  - www.bioinformaticsalgorithms.org/

- Additional problems will be assigned via ROSALIND
  - http://www.rosalind.info/

About This Class

- Respect different levels of experience
- Adhere to the Collaboration Policy
  - Clearly defines the Honor Principle as it relates to coursework
  - Properly "cite" people you worked with
- Everyone will have an oh sh!t moment

- On the flip side, real life is more important than class
  - If something is happening, take time to deal with it
  - Come to me with any issues, but be aware that I am an obligated reporter
- Save, Save, Save, Save, Save

http://on-select.com/p/f/frustrated-smiley/4/

People

Professor
Anna Ritz
aritz@reed.edu, B200B
Office Hours
W/Th 11am-12pm
or by appointment

Tutors/TAs
Rose Driscoll (Mon Lab; Thurs DoJo)
Mina Marden (Tues Lab)
Elaine Kushkowski (Tues DoJo)

Other Resources are Available – talk to Anna

HW1 Out (Moodle)

- Excel assignment
  - Excel is available for free on the CIS software page
- Due Monday Jan 30 before class

- Part 1: Warmup with Excel Functions
- Part 2: Sequence Analysis (Wed. lecture)

Enter numbers/strings in green cells
Enter formulas (starting with "=") in blue cells.
Enter responses/text in red cells

Upcoming Schedule

Week 1	
Mon 1/23	Lecture HW1 Out
Lab1: Python Setup & Command Line Tools	
Wed 1/25	Lecture
Fri 1/27	Lecture
Week 2	
Mon 1/30	Lecture HW1 Due HW2 Out
Lab2: Python Pattern Maker	
Wed 2/1	Lecture
Fri 2/3	Lecture