# Bioinformatics Algorithms
# COS-BIOL-530/630
# Lecture01

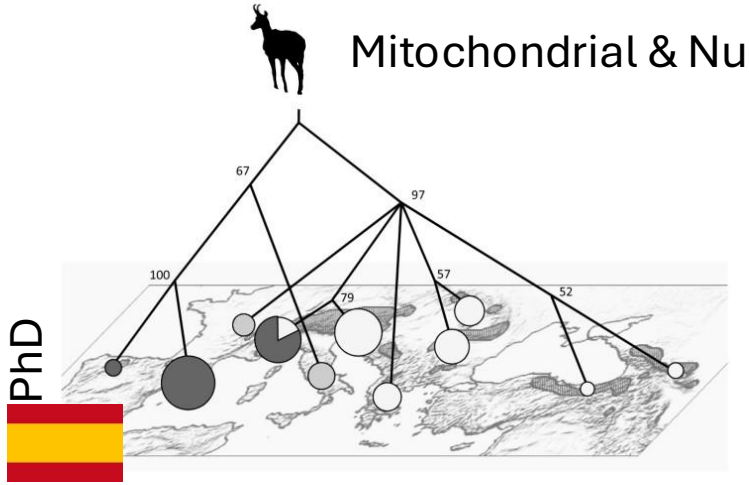
| Days & Times | Room | Meeting Dates |
|---|---|---|
| Tu 2:00PM - 3:50PM | Thomas Gosnell Hall (GOS)-2178 | 01/13/2025 - 04/28/2025 |
| Th 2:00PM - 3:50PM | Thomas Gosnell Hall (GOS)-2178 | 01/13/2025 - 04/28/2025 |

Instructor:
Fernando Rodriguez
email: frvsbi@rit.edu
Office: Orange Hall 1311

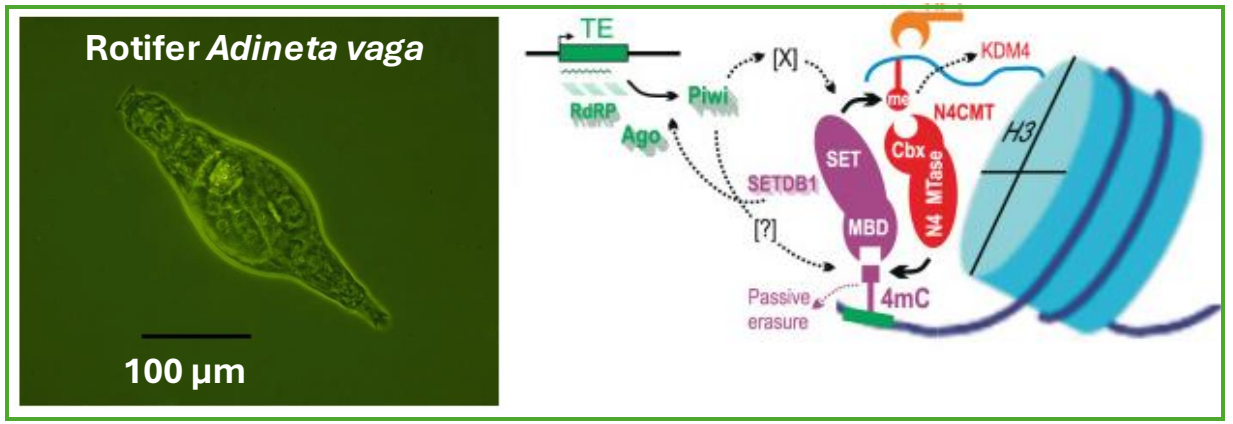
# Fernando Rodriguez

- I am a molecular biologist interested in genomics in eukaryotes

- I have studied genomic structural organization/regulation in different model organisms:
  - Cattle, chamois (mammals)
  - Fungi: Neurospora
  - Rotifers
  - Desert ants
  - Ostracods (crustaceans)








Epigenetic mechanism to repress transposons (TEs)

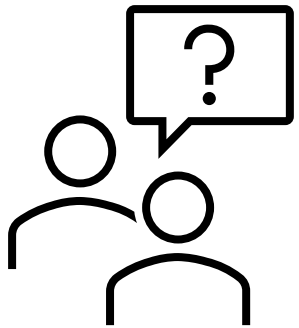
# Bioinformatics Algorithms
# Intro
# - Lecture01-

- What is Bioinformatics?
- Models and Algorithms
- Computational Biology Databases
- Syllabus and Schedule

**Announcements**
**Week1:**
Discussion 1
Activity 1

Quiz 1 opens on Friday. Due on Tuesday 2pm.

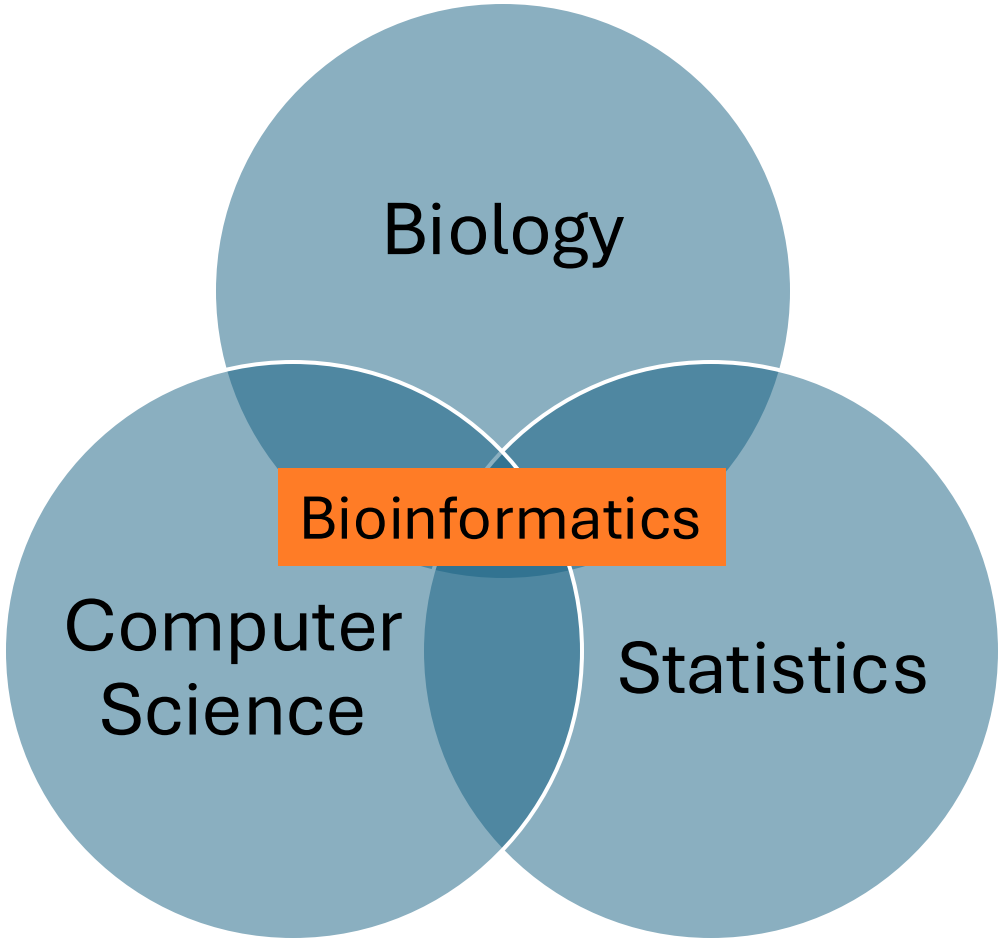
- Why Bioinformatics Algorithms?

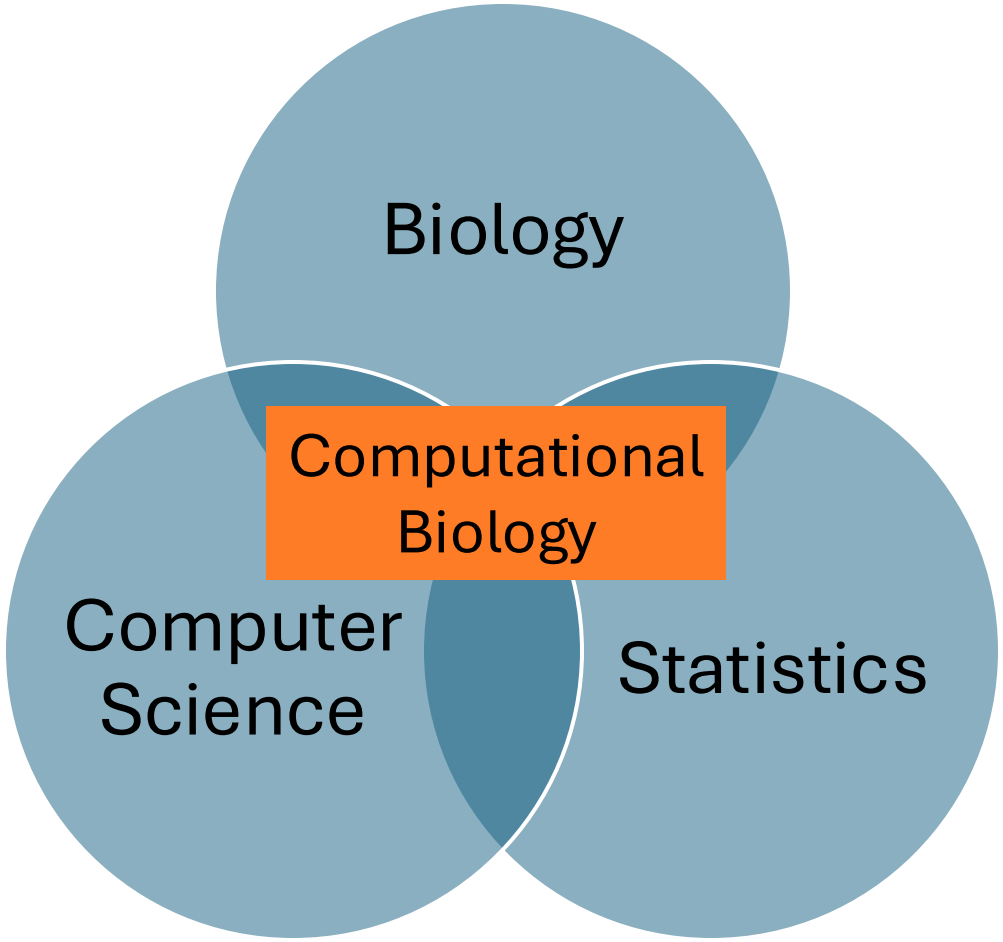



# What is Bioinformatics?

- The use of computer technology to
  - collect-store-analyze biological datasets

- Development of tools:
  - DNA/Protein sequence analysis
  - RNA/Protein structure (amino acid sequence -> 3D model)
  - Databases
    - Gene
    - Protein
    - Genome
    - Function


Biology
Bioinformatics
Computer Science
Statistics


Biology
Computational Biology
Computer Science
Statistics

# Computational Biology vs. Bioinformatics

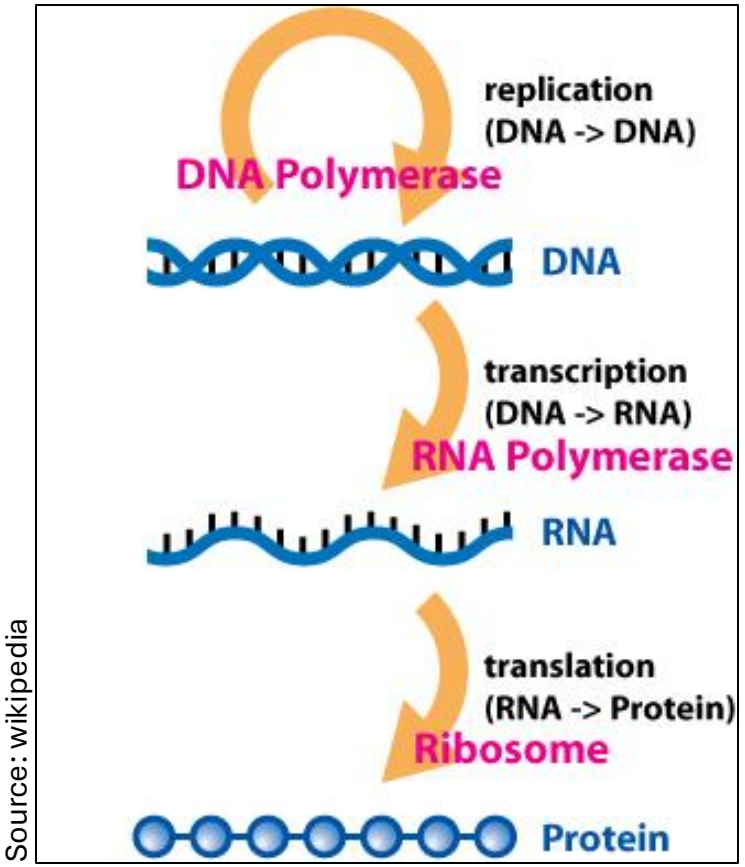
**Computational biology**

- The study of biology using computational techniques
  - Computer science
  - Statistics
  - Mathematics
- Involves the development and application of data-analytical and theoretical methods to the study of biology

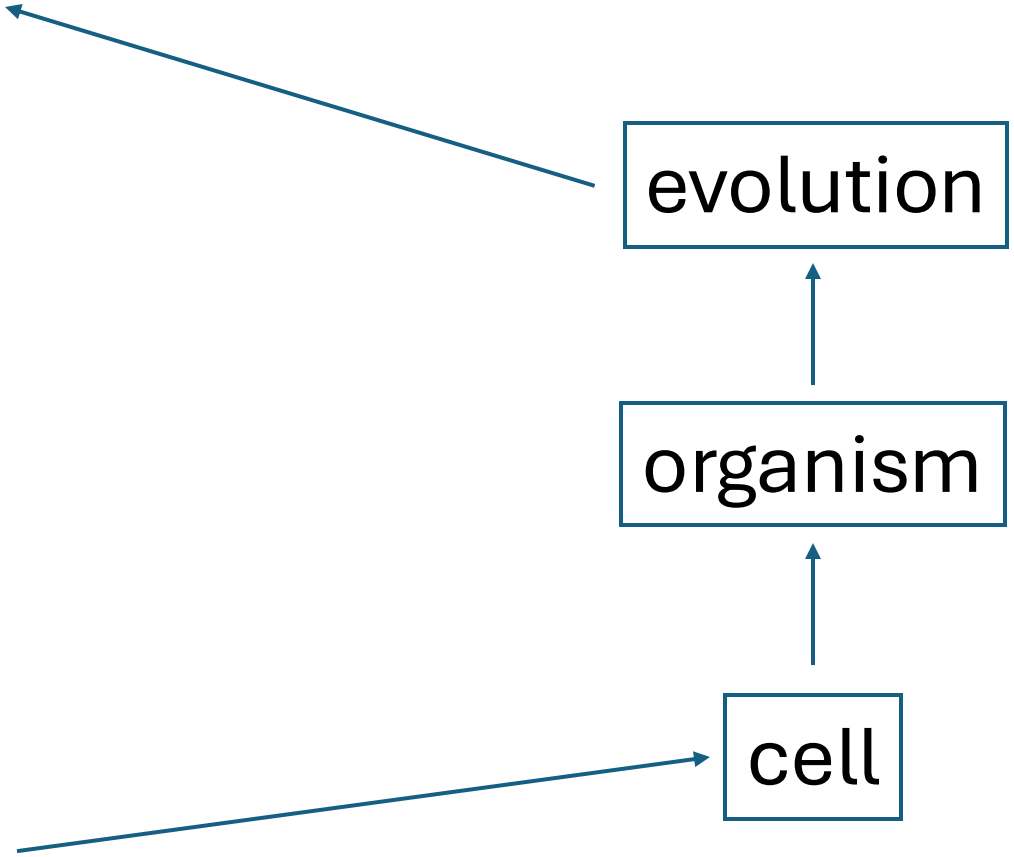
**Bioinformatics**

- The creation of **tools** that solve biological problems
  - Algorithms
  - Databases
- Involves the development of methods and software tools to understand biological data
- …But you need to provide some **rules**

# The central ~~rule~~ dogma of Molecular Biology



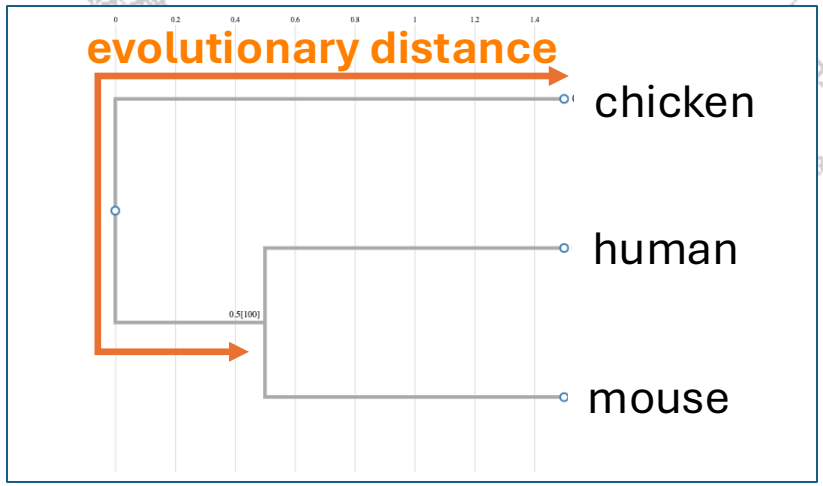

Source: wikipedia



# Computational biology

```
Mouse   MTNMRKTHPLFKIINHSFIDLPAPSNISSWWNFGSLLGVCLMVQIITGLFLAMHYTSDTM
Human   MTPMRKINPLMKLINHSFIDLPTPSNISAWWNFGSLLGACLILQITTGLFLAMHYSPDAS
Chicken NIRKSHPLLKMINNSLIDLPAPSNISAWWNFGSLLAVCLMTQILTGLLLAMHYTADTSLA
```

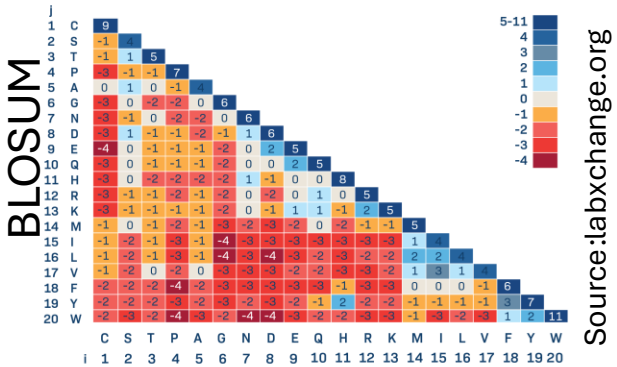
```
Mouse          MTNMRKTHPLFKIINHSFIDLPAPSNISSWWNFGSLLGVCLMVQIITGLFLAMHYTSDTM
Chicken        --NIRKSHPLLKMINNSLIDLPAPSNISAWWNFGSLLAVCLMTQILTGLLLAMHYTADTS
Human          MTPMRKINPLMKLINHSFIDLPTPSNISAWWNFGSLLGACLILQITTGLFLAMHYSPDAS
                  :** :**:*:**:*:****:*****:********..**: ** ***:*****:.*:
```




# Bioinformatics

Sequence alignment algorithm: Parameters
- -Global alignment
- -Local alignment
- -Gaps (-)

Scoring matrix




Phylogenetic trees: evolutionary relationships
- -Molecular (Central Dogma)
- -Heritable traits: morphology, behavioral…

# BIOL-530/630

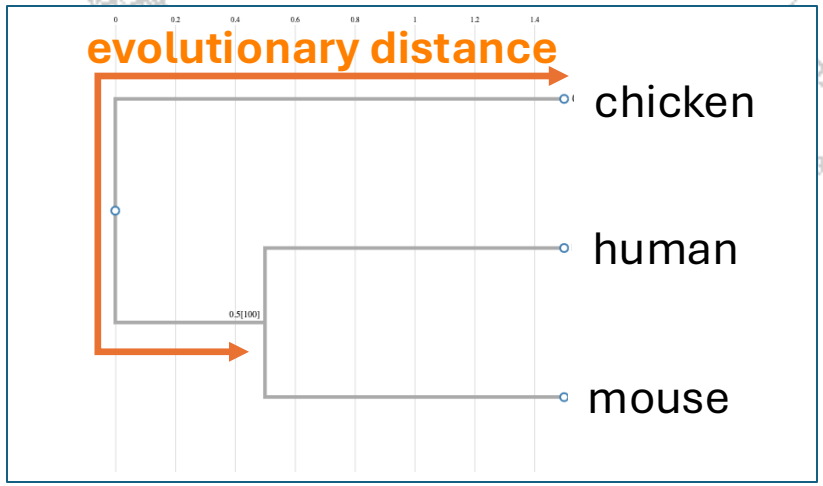
## Labs

### Computational biology

```
Mouse   MTNMRKTHPLFKIINHSFIDLPAPSNISSWWNFGSLLGVCLMVQIITGLFLAMHYTSDTM
Human   MTPMRKINPLMKLINHSFIDLPTPSNISAWWNFGSLLGACLILQITTGLFLAMHYSPDAS
Chicken NIRKSHPLLKMINNSLIDLPAPSNISAWWNFGSLLAVCLMTQILTGLLLAMHYTADTSLA
```

Cytochrome b
1140 bp

```
Mouse          MTNMRKTHPLFKIINHSFIDLPAPSNISSWWNFGSLLGVCLMVQIITGLFLAMHYTSDTM
Chicken        --NIRKSHPLLKMINNSLIDLPAPSNISAWWNFGSLLAVCLMTQILTGLLLAMHYTADTS
Human          MTPMRKINPLMKLINHSFIDLPTPSNISAWWNFGSLLGACLILQITTGLFLAMHYSPDAS
                  :** :**:*:**:*:****:*****:********..**: ** ***:*****:.*:
```

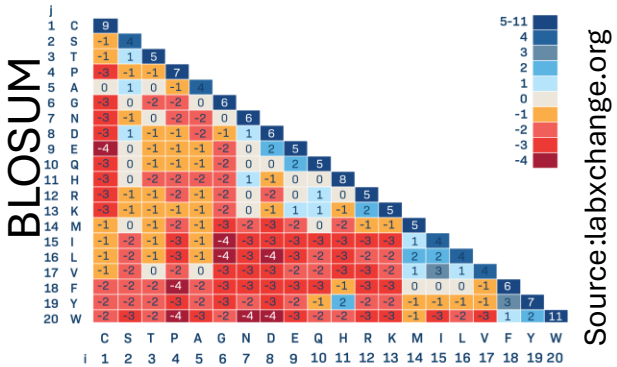



## Lectures

### Bioinformatics

Sequence alignment algorithm: Parameters
- -Global alignment
- -Local alignment
- -Gaps (-)

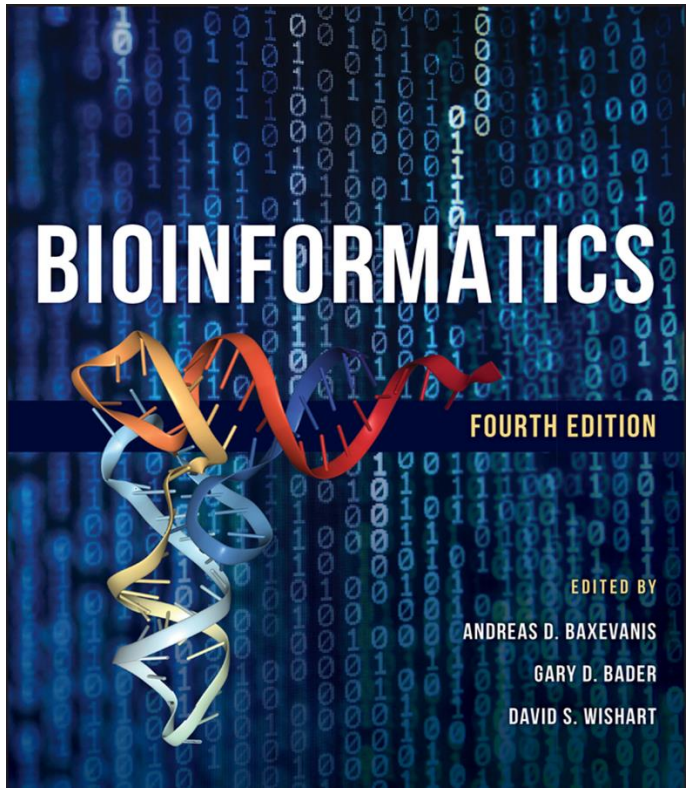
Scoring matrix




Phylogenetic trees: evolutionary relationships
- -Molecular (Central Dogma)
- -Heritable traits: morphology, behavioral…

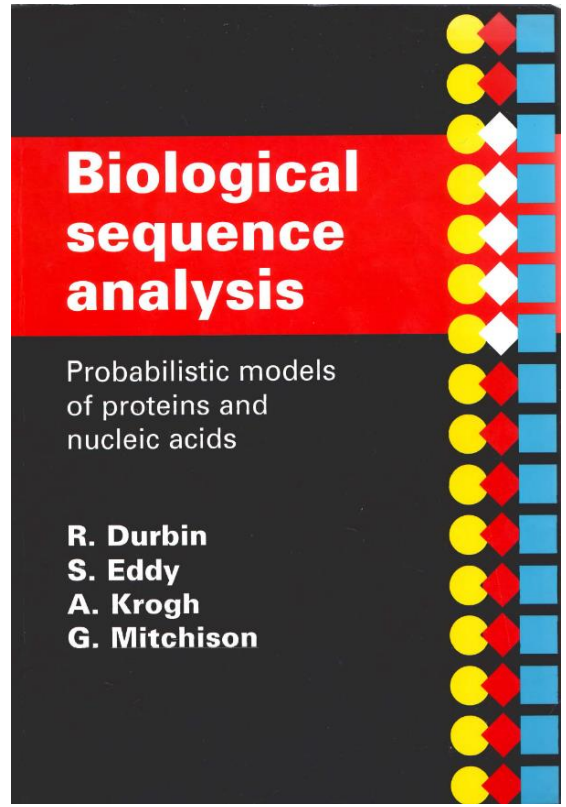
# BIOL-530/630

## Labs




‘Practical Guide”

## Lectures




Models and Algorithms

# Models and Algorithms used in Computational Biology

A **model** is a parametric explanation of the observations of interest.

**Probabilistic models/methods:**

- Maximum Likelihood
- Bayesian
- Machine Learning
- Markov Chain Models
- Hidden Markov Models

**"All models are wrong, but some models are useful"**

George Box

An **algorithm** is a set of instructions for solving a problem, e.g. , inferring the optimal value of a model's parameter.

**Algorithms/Methods:**

- Sequence (string): sort/search algorithms
- Optimization algorithms:
  - Linear programing
  - Dynamic programing
  - Greedy algorithms
  - Heuristic methods

**Correct *versus* incorrect Algorithms**

# Models and Algorithms

**Structural modeling** can be used to generate hypotheses about the structure (and therefore to imply things about the function) of macromolecules

**Experimental biology** plays an equally important role

An **algorithm** is a sequence of instructions that one must perform in order to solve a well-formulated problem.

You can write an algorithm with a pseudocode

And use a programming language to implement

**A number *n* is even if and only if n = 2k, where k is an integer**

**Example of a pseudocode:**

```
Even/odd number
1.INPUT n
2.remainder = n % 2
3.IF remainder is not
equal to 0
4.answer = odd
5.ELSE
6.answer = even
7.OUTPUT answer
```

# Models and Algorithms

**Structural modeling** can be used to generate hypotheses about the structure (and therefore to imply things about the function) of macromolecules

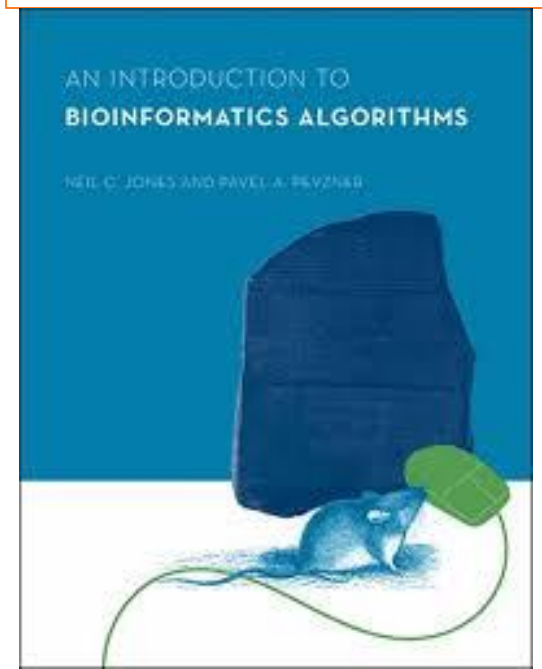
**Experimental biology** plays an equally important role

An **algorithm** is a sequence of instructions that one must perform in order to solve a well-formulated problem.

You can write an algorithm with a pseudocode

And use a programming language to implement

**A number *n* is even if and only if n = 2k, where k is an integer**




***Introduction to bioinformatics algorithms*, Jones and Pevzner**

"How to design algorithms that solve biological problems"

# Algorithms and Models in Bioinformatics

**Model of a gene**

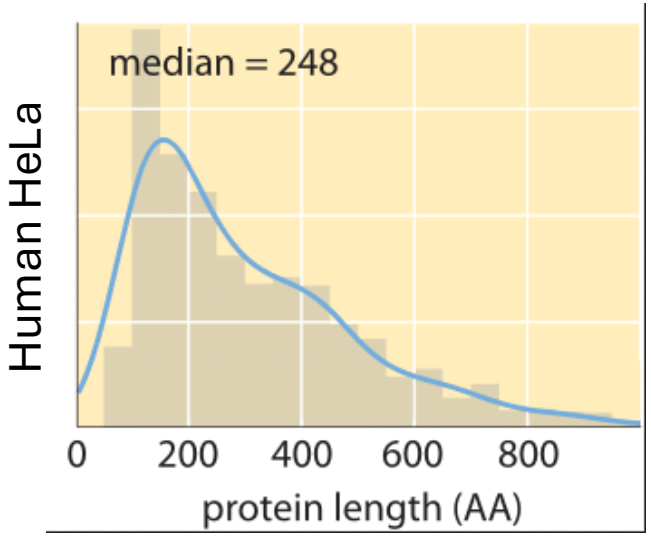
A gene is a sequence of nucleotides that encodes a protein sequence

- between _____residues in length
- A gene starts with _____
- And ends with _____

**Gene finding algorithm**

Search for and identify all sequences that:

- start with ____
- end with either _____
- between _____nucleotides in length




Source: https://book.bionumbers.org

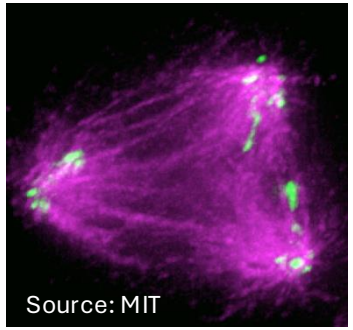
# Bioinformatics Applications

Levels

- Genes -> Phylogenetics | Gene prediction | Annotation
- Assembly/genome -> Comparative genomics
- Cell regulation -> Transcription and gene regulation

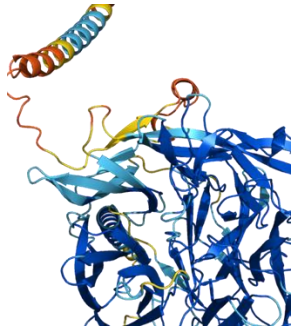
**Genomics**

**Cell Biology & Development**



Source: MIT

**Proteomics**



Source: Alpha Fold

**Pharmacology**

# Computational Biology Databases

- GenBank is the first database of nucleotide and protein sequences (Los Alamos National Laboratory 1982)
- NCBI was created as part of NLM in 1998
- Development of tools to allow access to sequence databases (BLAST 1990, ENTREZ 1992)


BLAST (Heuristic method)

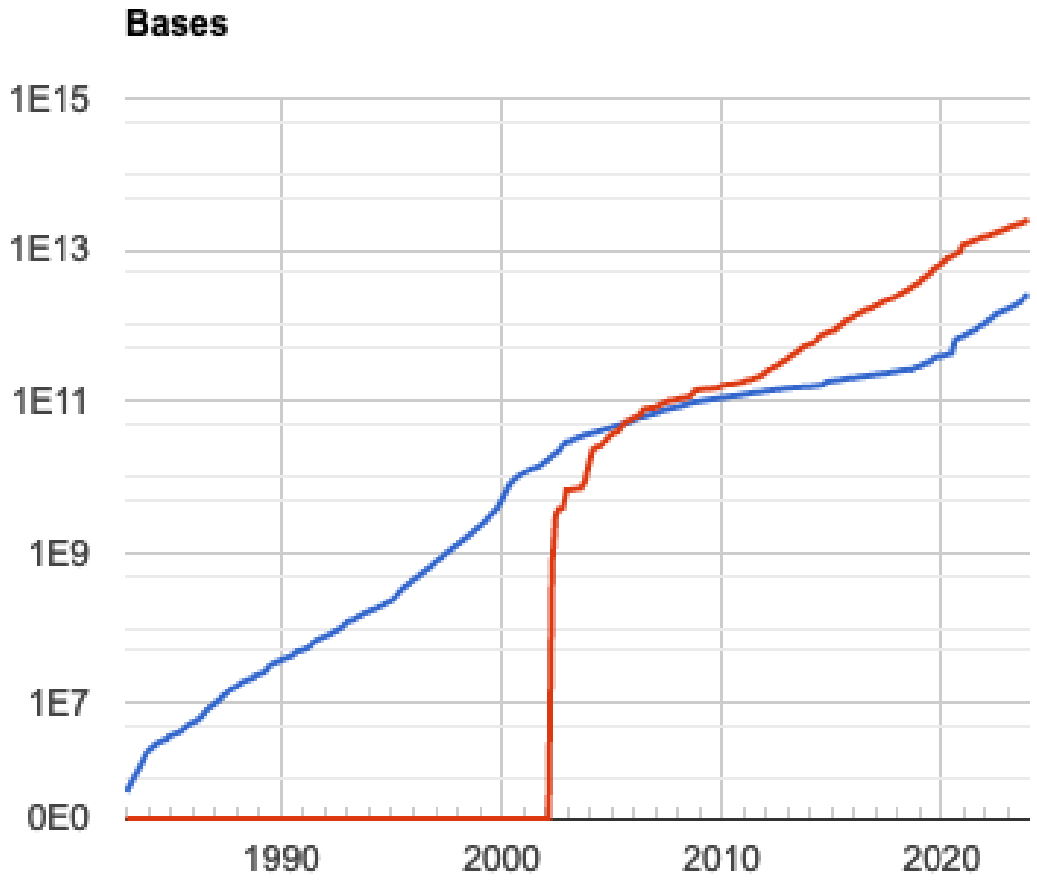
Mouse MTNMRKTHPLFKIINHSFIDLPAPSNISSWWN

Human MTPMRKINPLMKLINHSFIDLPTPSNISAWWN

Chicken NIRKSHPLLKMINNSLIDLPAPSNISAWWN

Seqs NIRKSHPLLKMINNSLIDLPAPSNISAWWN ??

# Genome analysis: Big data




First ever genome sequenced?

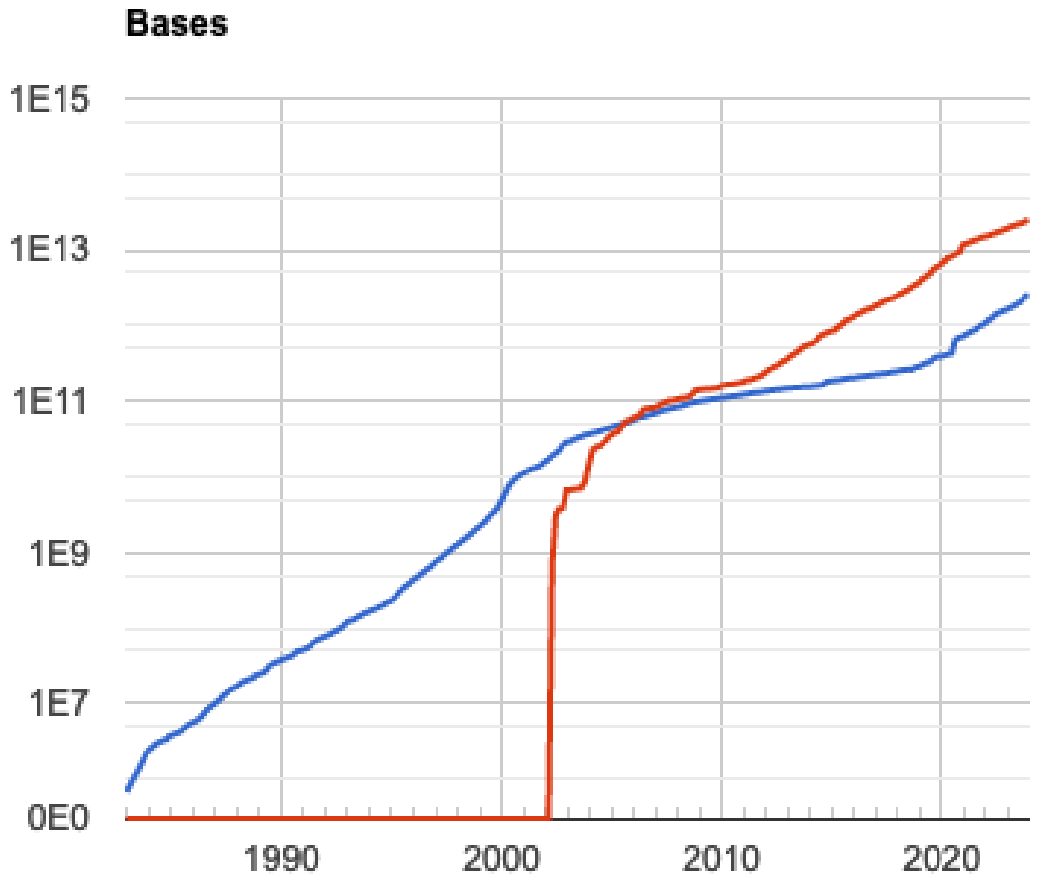
First cell genome sequenced?

First eukaryotic genome sequenced?

First metazoan genome sequenced?

https://www.ncbi.nlm.nih.gov/genbank/statistics/

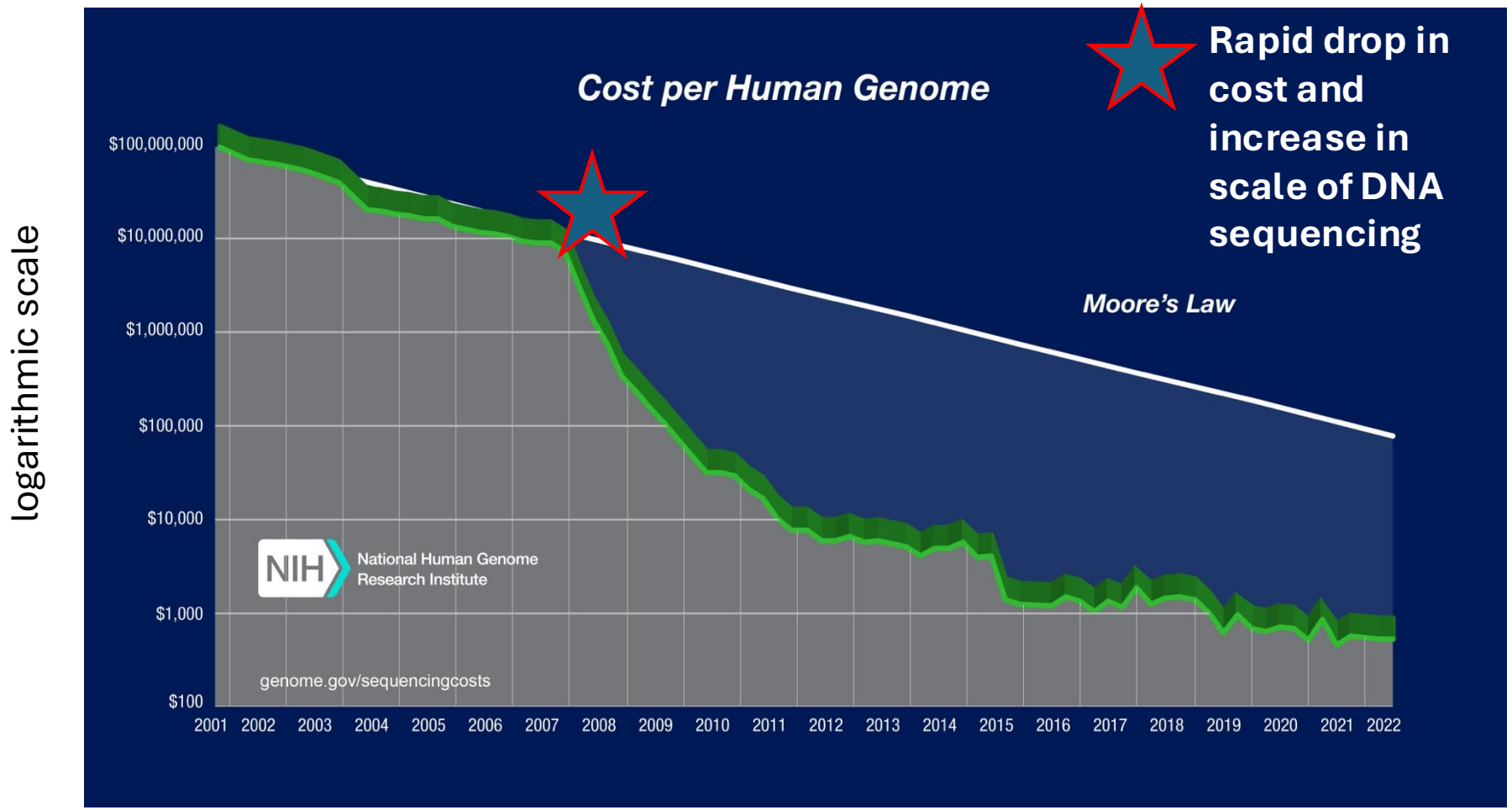
# Genome analysis: Big data




- Phage Phi-x174 (1977) - 5kb
- *Haemophilus influenzae* (1995) - 1.8Mb
- Budding yeast *Saccharomyces cerevisiae (*1996) - 12Mb
- The roundworm *C. elegans* (1998) - 100Mb
- *Drosophila melanogaster* (2000) - 180 Mb
- Human (2003) - 3.1Gb

https://www.ncbi.nlm.nih.gov/genbank/statistics/

# Genome analysis: Big data



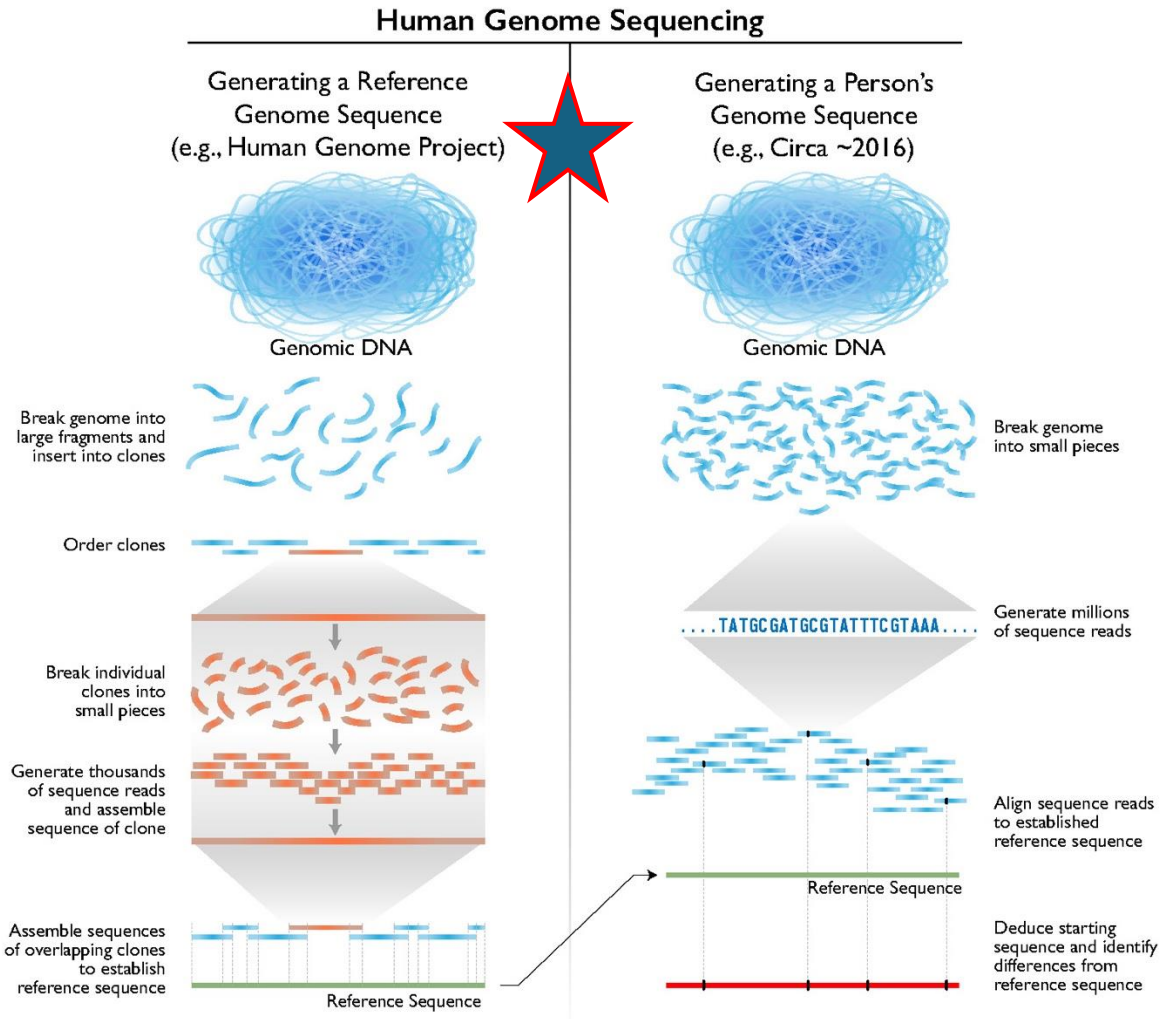

*Moore's Law* states that the number of components on a single chip *doubles every* two years at minimal cost

# Genome analysis: Big data

<u>the Human Genome Project</u>

- $3 billion project was formally founded in 1990
- Declared complete in April 2003
- Protein-coding genes?
  - Genome size 3.1 Gb
  - Gene average size 24 Kilobases

$$\frac{3100\ Mb}{0.024\ Mb}$$

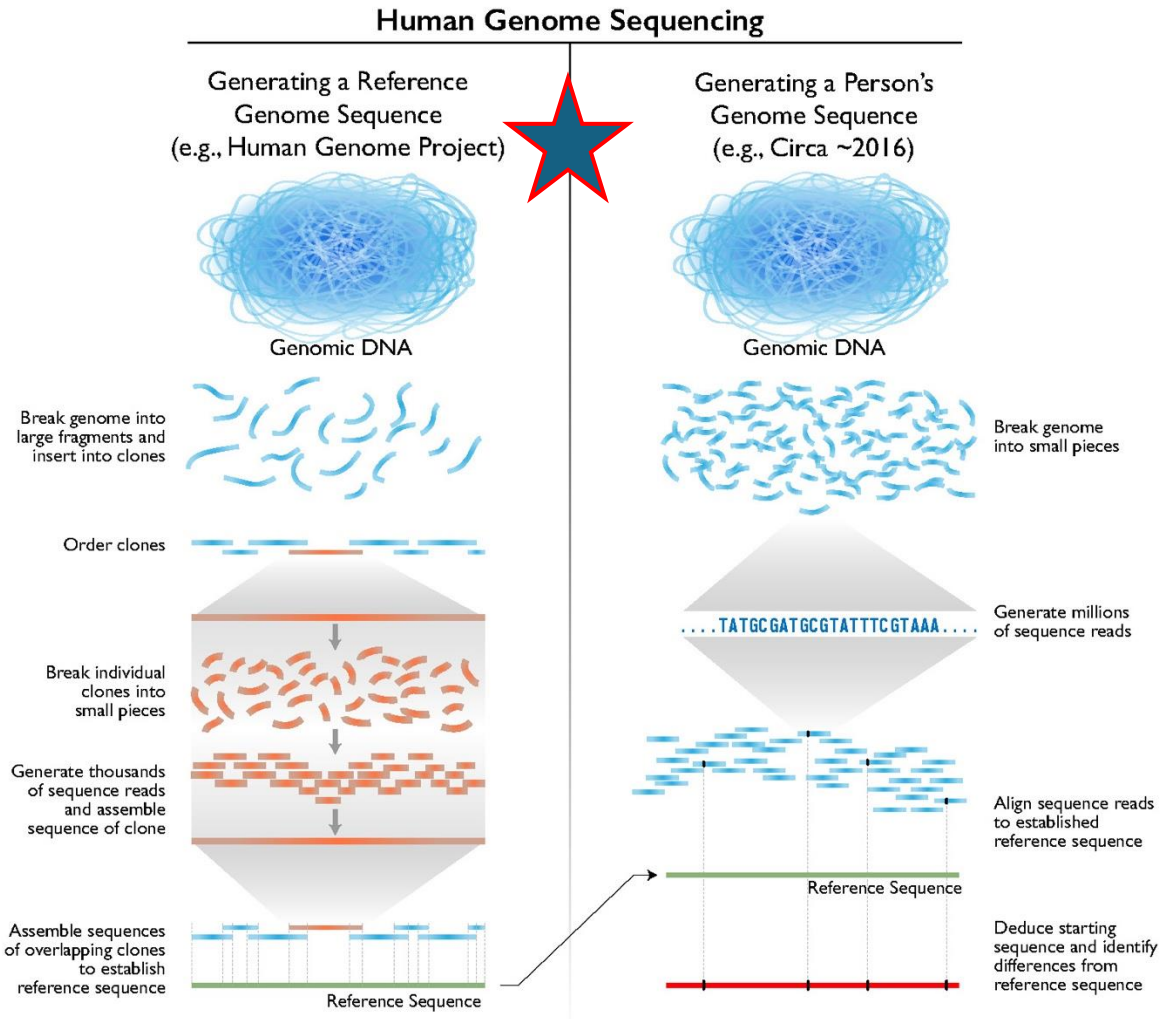



https://www.genome.gov/sequencingcosts

# Genome analysis: Big data

<u>the Human Genome Project</u>

- $3 billion project was formally founded in 1990
- Declared complete in April 2003
- Only ~20,000 protein-coding genes (instead of ~100,000)…**WHY?**

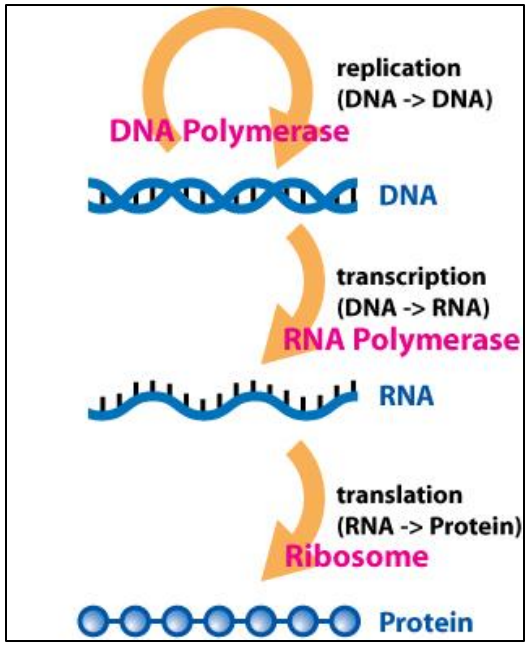



$$\frac{3100\ \frac{Mb}{}}{0.024\ Mb}$$

https://www.genome.gov/sequencingcosts

# Genome analysis: Big data…new rules?

central dogma




Exceptions to the central dogma

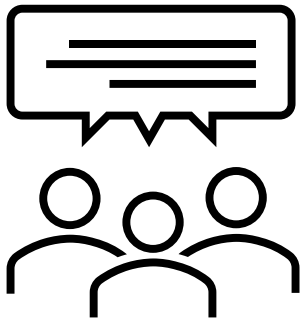
- Horizontal gene transfer
- Repetitive elements
- Transposons
- Alternative splicing: transcript variants
- Epigenetics
  - Histone modification
  - DNA methylation
  - Non-coding RNA

…we are going to need more **bioinformatic tools**!

# Course Objectives

- To gain a thorough understanding of the concepts of **bioinformatics**.
- Understand how probabilistic **models** work.
- Understand how to apply computational **algorithms** to biology.
- To demonstrate knowledge of the theory behind and **application** of methods involved in phylogenetic tree construction and other bioinformatics analyses.

# BIOL-530/630 Syllabus/Schedule Spring2025



# Schedule

Lectures - Tuesday

- Slides - myCourses
- Lecture assignments - Perusal
- Weekly Quiz - myCourses

Labs - Thursday

- Dry Lab workshop
- Copy lab materials in your folder
- Document your steps
- Discussions - myCourses
- Assignments - myCourses

| Day | Date | What are we doing in class today? |
|---|---|---|
| Tuesday | Jan 14 | Sequence scavenger hunt: asynchronous activity |
| Thursday | Jan 16 | Introductions, Syllabus review, Introduction to Bioinformatics Algorithms |
| Tuesday | Jan 21 | Sequence alignment<br>Pairwise/Multiple |
| Thursday | Jan 23 | Pairwise alignment<br>DotBlots<br>Alignment scores<br>Global alignment |
| Tuesday | Jan 28 | Sequence alignment and database search<br>Genome-level comparisons |
| Thursday | Jan 30 | Blast |
| Tuesday | Feb 4 | Phylogenetics<br>Substitution matrices<br>Distance-based methods |
| Thursday | Feb 6 | Phylogenetic analysis |
| Tuesday | Feb 11 | Phylogenetics<br>Character-based methods |
| Thursday | Feb 13 | Phylogenetic analysis |
| Tuesday | Feb 18 | Genome comparisons |
| Thursday | Feb 20 | Genome similarity |
| Tuesday | Feb 25 | Functions of genes:<br>knowledge bases |
| Thursday | Feb 27 | EXAM 1 |
| Tuesday | Mar 4 | GO enrichment |
| Thursday | Mar 6 | Pattern matching algorithms |
| Tuesday | Mar 11 | Spring Break - No Class |
| Thursday | Mar 13 | Spring Break - No Class |
| Tuesday | Mar 18 | RNA secondary structure |
| Thursday | Mar 20 | RNA folding and motifs |
| Tuesday | Mar 25 | Hidden Markov Models<br>Gene Finding in prokaryotes |
| Thursday | Mar 27 | Open reading finder |
| Tuesday | Apr 1 | Intro to High Throughput Sequencing (HTS)<br>Technologies/platforms |
| Thursday | Apr 3 | HTS overview |
| Tuesday | Apr 8 | HTS<br>Alignment<br>Assembly |
| Thursday | Apr 10 | EXAM 2 |
| Tuesday | Apr 15 | HTS<br>RNA-seq |
| Thursday | Apr 17 | Alignment/assembly |

# Schedule

<u>Lectures - Tuesday</u>

- Slides - myCourses
- Lecture assignments - Perusal
- Weekly Quiz - myCourses

<u>Labs - Thursday</u>

- Dry Lab workshop
- Copy lab materials in your folder
- Document your steps
- Discussions - myCourses
- Assignments - myCourses

| SUN | MON | TUE | WED | THU | FRI | SAT |
|---|---|---|---|---|---|---|
| | | Lecture | | Lab<br>Discussion<br>Activity | Quiz Open<br>Discussion due | |
| | | Quiz due | | Activity due | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |

# Schedule

Lectures

- Slides - myCourses
- Lecture assignments - Perusal
- Weekly Quiz - myCourses

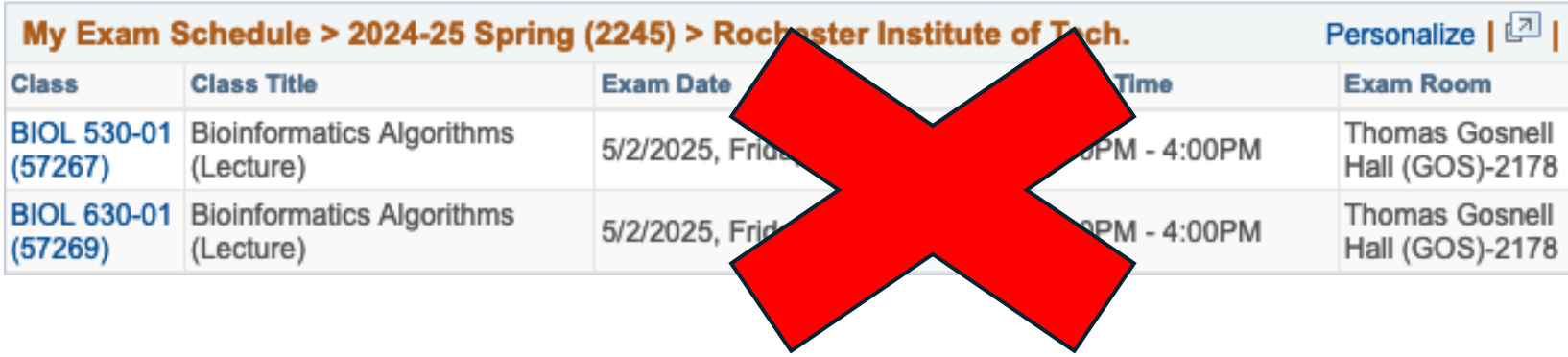
Labs

- Dry Lab workshop
- Copy lab materials in your folder
- Document your steps
- Discussions - myCourses
- Assignments - myCourses

# Project

| | | |
|---|---|---|
| Tuesday | Apr 22 | Final Project Q & A |
| Thursday | Apr 24 | Final Project Draft |
| Tuesday | Apr 29 | Final Project Q & A |
| Thursday | May 1 | Final Project due |

# No Final Exam!



My Exam Schedule > 2024-25 Spring (2245) > Rochester Institute of Tech. Personalize |

| Class | Class Title | Exam Date | Time | Exam Room |
|---|---|---|---|---|
| BIOL 530-01 (57267) | Bioinformatics Algorithms (Lecture) | 5/2/2025, Frid | PM - 4:00PM | Thomas Gosnell Hall (GOS)-2178 |
| BIOL 630-01 (57269) | Bioinformatics Algorithms (Lecture) | 5/2/2025, Frid | PM - 4:00PM | Thomas Gosnell Hall (GOS)-2178 |

# Grading

| | Percentage of final grade |
|---|---|
| **2 exams** | 30% |
| **Final Project** | 15% |
| **Quizzes** | 10% |
| **In-class Activities and Participation** | 20% |
| **Lab Discussions** | 20% |
| **Attendance** | 5% |
| **TOTAL** | **100%** |

Read the Syllabus in MyCourses!
*Academic accommodations

# Labs

- **Read Lab00**
- Terminal
- Install software (PHYLIP, MEGA, FigTree)
- Jupiter Notebook (Visual Studio Code)
- Check if websites work in your internet browser
- Server ***oedipus.bioinformatics.rit.edu***
  - Account up and operational
  - Password changed and remembered!