# Bioinformatics and Health Technology

## Lecture 1: Data Structures for Bioinformatics Primer

William Okech, PhD (Instructor)

2025

# Outline

1. Class Logistics
2. Introduction to Biological Data
3. Overview of Data Structures
4. Applications of Data Structures in Bioinformatics
5. Summary

# Outline

1. Class Logistics
2. Introduction to Biological Data
3. Overview of Data Structures
4. Applications of Data Structures in Bioinformatics
5. Summary

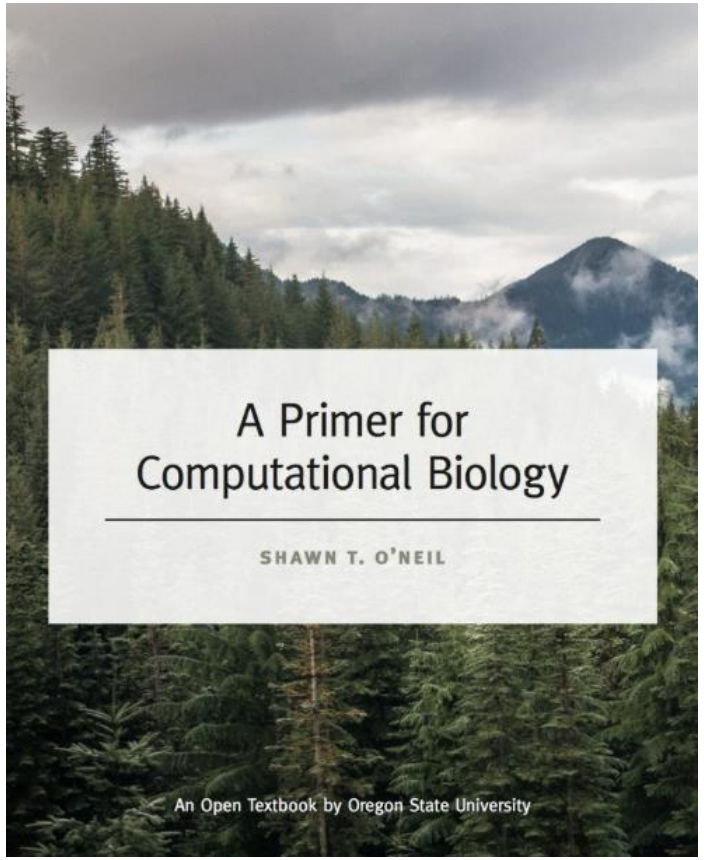
# Class Policies

**Expectations**: To ensure a safe learning environment please treat each other with respect.

**Attendance**: Course is run using an in-person format and attendance is mandatory unless explicit permission is obtained from the instructor.

**Assignments**: Late homework and lab reports will not be accepted unless extensions have been requested in advance.

# Reference Textbooks




**Chapter 25: Algorithms and Data Structures**
https://open.oregonstate.education/computationalbiology/

Bioinformatics

Laurent Gatto

Chapter 3 High-level data structures

https://uclouvain-cbio.github.io/WSBIM1322/sec-obj.html

**Data Structures and Algorithms for Bioinformatics**

Philip Machanick

# Outline

1. Class Logistics ✔
2. Introduction to Biological Data
3. Overview of Data Structures
4. Applications of Data Structures in Bioinformatics
5. Summary

# Biological Data is Everywhere!













# Data Lifecycle Management

Biological Samples

Scientists and Clinicians

Data Analysts

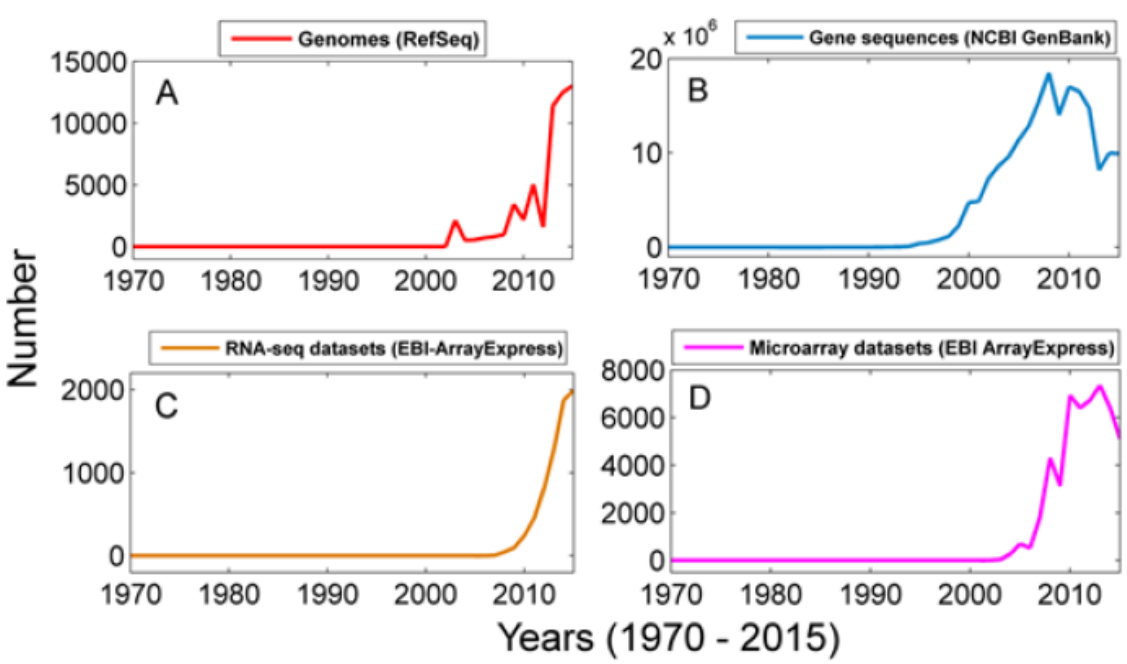
**BIOSTATISTICIANS AND DATABASE ADMINISTRATORS**

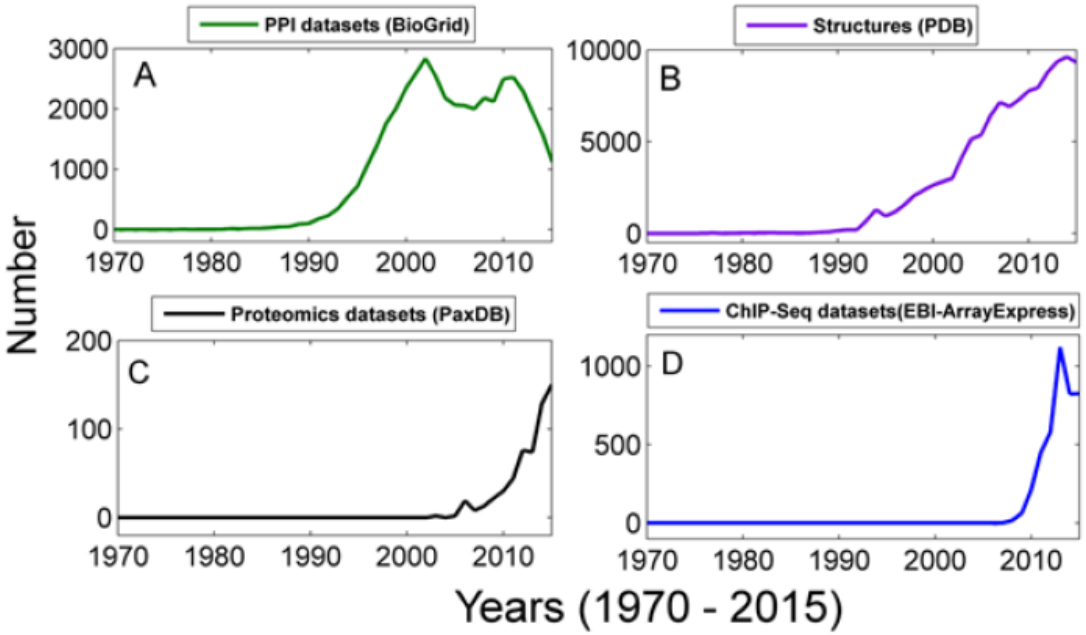


Note: Images obtained from https://bioart.niaid.nih.gov/

# Data Lifecycle Management

- Data Deluge
  - Increasing use of high throughput experimental techniques.
  - Generation of large amounts of diverse and heterogeneous biological data.
  - Challenges associated with the **management**, **storage**, and **analysis** of data.






Sarkar RR (2016) The Big Data Deluge in Biology: Challenges and Solutions. Global J Technol Optim 7: e103.

# Data Lifecycle Management

- What are the consequences of poor data management?
  - Decreased Reproducibility
  - Data Wastage
  - Cost

- Need for standardized formats → **Data Structures**

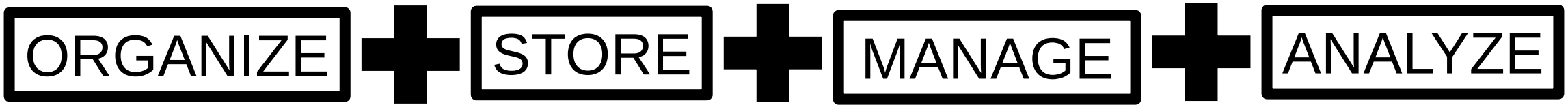
# Outline

1. Class Logistics ✓
2. Introduction to Biological Data ✓
3. Overview of Data Structures
4. Applications of Data Structures in Bioinformatics
5. Summary

# Data Structures

- An **organization** for a collection of data that ideally allows for fast access and certain operations on the data (O'Neil, 2019).

- A method of **storing** information (either single item or some kind of composite item), and it is also characterized by the **operations** that can be performed on it (Machanick, 2019).

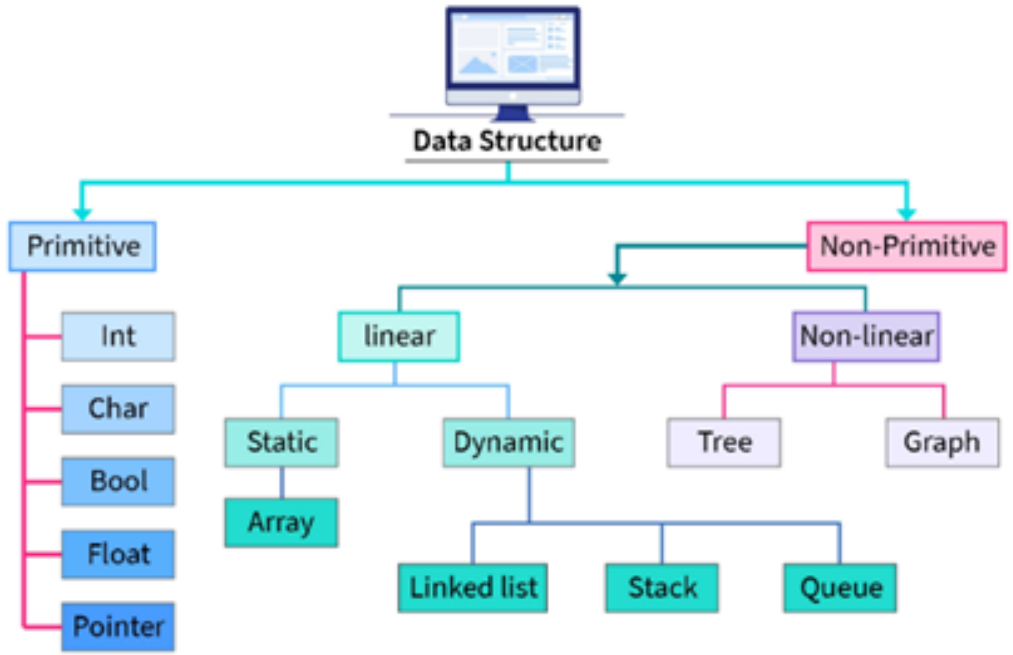



1. O'Neil, S. T. (2019). A primer for computational biology. Pressbooks. https://open.oregonstate.education/computationalbiology/
2. Machanick, P (2019). Data Structures and Algorithms for Bioinformatics. RAMpage Research

# Data Structures

**Primitive:** Integer, Character, Boolean, and Floats.

**Non-Primitive:** Linear, Non-Linear, and Other.






Classification of Data Structures (https://findtodaysnotes.wordpress.com/classification-of-data-structures/)

# Outline

1. Class Logistics ✓
2. Introduction to Biological Data ✓
3. Overview of Data Structures ✓
4. Applications of Data Structures in Bioinformatics
5. Summary and Quiz

# Data Structures in Bioinformatics

- Enable the efficient **storage, organization, management, and analysis** of large-scale biological data.
- Applications:
  - Genomic Sequence Analysis
  - Genome Assembly
  - Biological Network Analysis
  - Structural Bioinformatics
  - Phylogenetic Analysis
  - Data Integration and Mining
  - Gene Expression Analysis and NGS Data Processing
- Review of commonly used data structures in bioinformatics

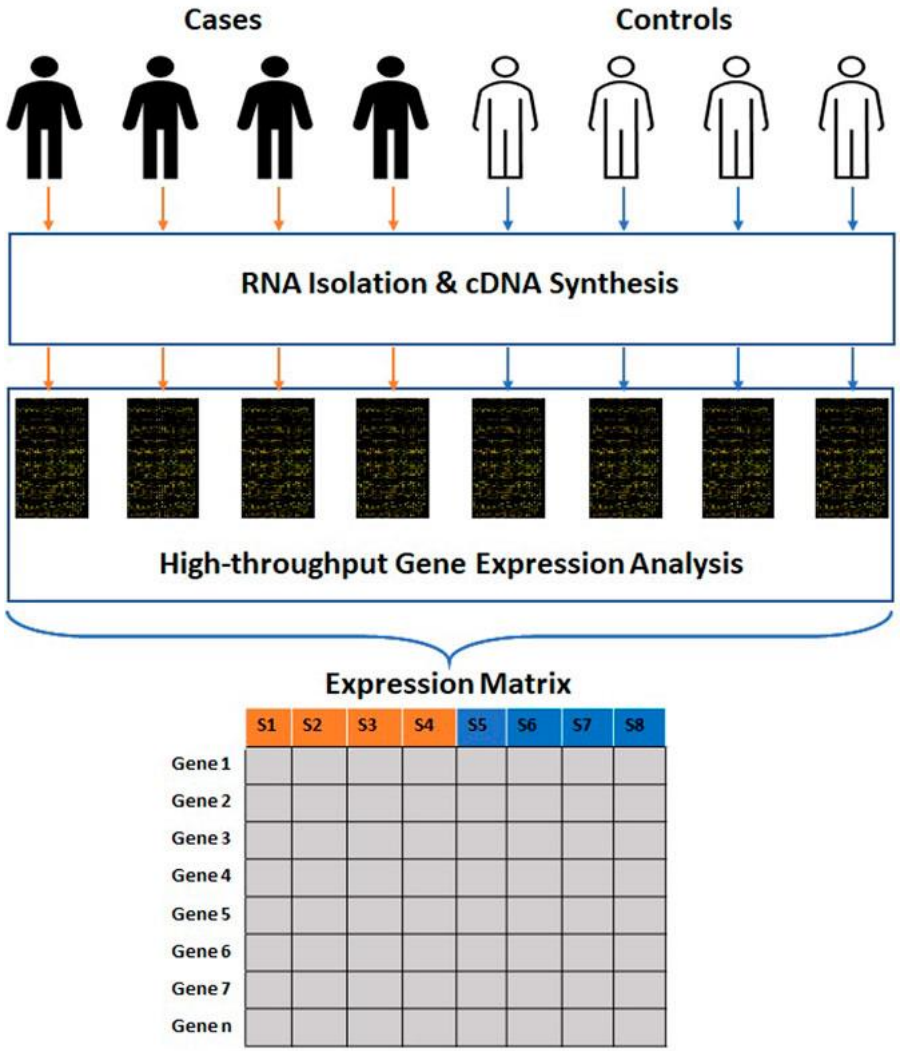
# Data Structures in Bioinformatics

## Arrays and Matrices

- Store collections of elements of the same data type.
- **Vectors**: One-dimensional arrays.
- **Matrices**: Two-dimensional arrays.

## Example Usage

- Gene expression matrices




Bhandari, N., Walambe, R., Kotecha, K., & Khare, S. P. (2022). A comprehensive survey on computational learning methods for analysis of gene expression data. Frontiers in Molecular Biosciences, 9.

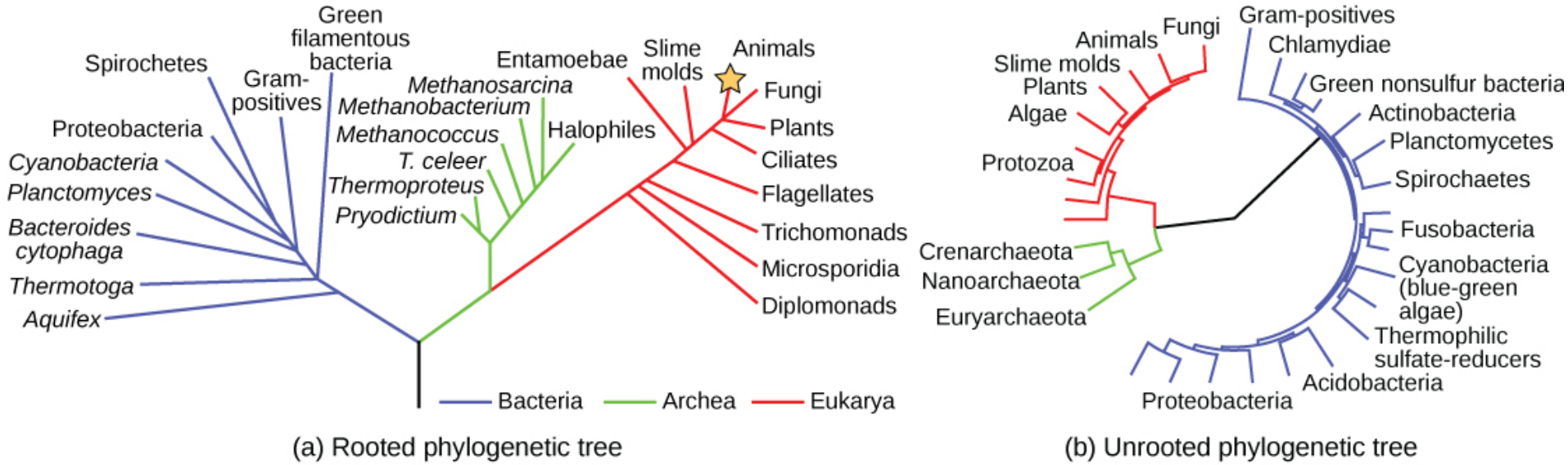
# Data Structures in Bioinformatics

## Trees

- Hierarchical structures with a root (parent) node and child nodes.

## Example Usage

- Phylogenetic Trees






Limeri, L., & Reid, J. (2023, January 11). Introductory Biology 2. Pressbooks. https://raider.pressbooks.pub/biology2/

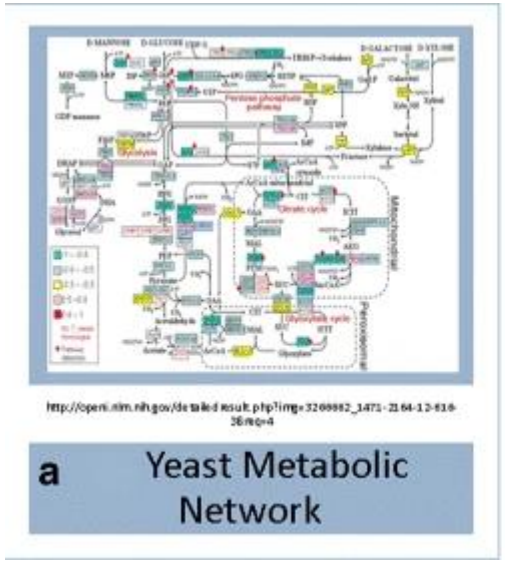
# Data Structures in Bioinformatics

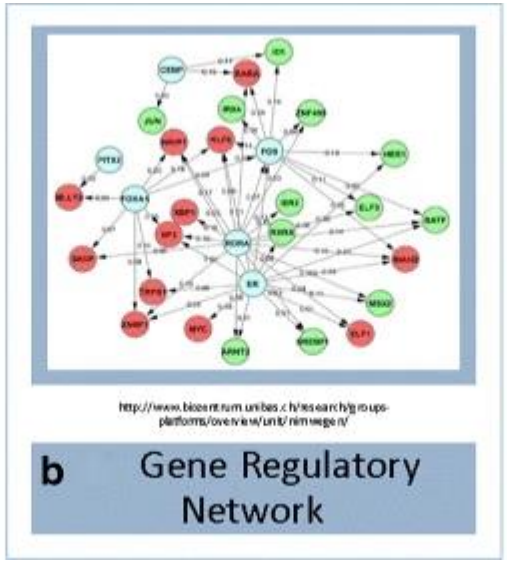
## Graphs

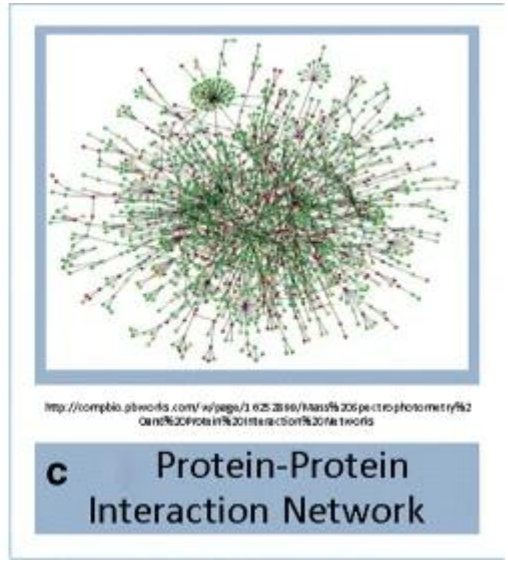
- Network of **nodes** connected to each other via **edges**.
- **Nodes** contain data fields.
- **Edges** act as an interaction between data fields.

## Example Usage

- Metabolic Network
- Gene Regulatory Network
- Protein-Protein Interaction Network






Kim, W., Haukap, L. NemoProfile as an efficient approach to network motif analysis with instance collection. BMC Bioinformatics 18 (Suppl 12), 423 (2017).

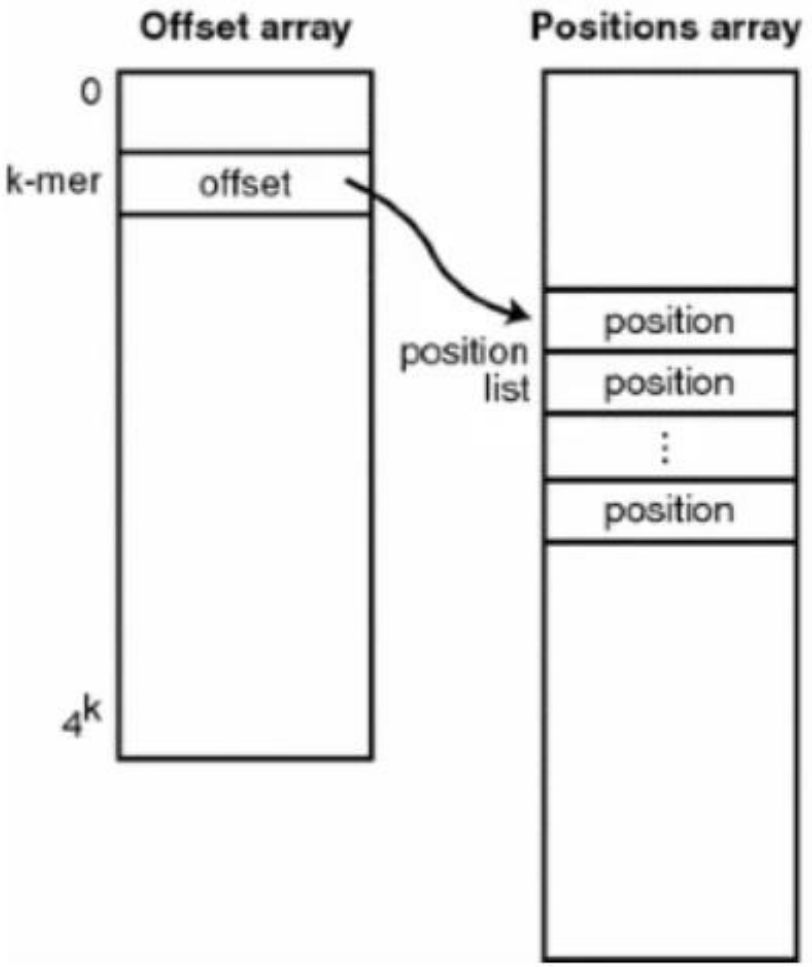
# Data Structures in Bioinformatics

## Hash Tables

- Store key-value pairs.
- Uses hash function to map keys to indices in an array.
- Enables fast look up of values based on their keys.

## Example Usage

- Genome indexing (Wu, 2016)






Wu, T.D. Bitpacking techniques for indexing genomes: I. Hash tables. Algorithms Mol Biol 11, 5 (2016).

# Data Structures in Bioinformatics

## Linked Lists

- Lists can store elements of different types.
- Linked lists are linear data structures where elements are stored in nodes. Each node contains data and a pointer to the next node.

## Example Usage

- Sequence Alignment and Genome Assembly
- Cyclic Metabolic Pathways (Krebs or Calvin)

Pathways of Metabolism | British Society for Cell Biology. (n.d.-b). https://bscb.org/learning-resources/softcell-e-learning/pathways-of-metabolism/

# Data Structures in Bioinformatics

## Stacks and Queues

- Linear and dynamic data structures that are used for ordering operations on data.
- **Stacks**: **Last In**, First Out.
- **Queues**: **First In**, First Out.

## Example Usage

- **Stacks**: Backtracking algorithms for aligning biological sequences.
- **Queues**: Parallel processing and managing bioinformatics pipelines.



Hendrix, D. A. (2019, October 3). Applied Bioinformatics. Pressbooks. https://open.oregonstate.education/appliedbioinformatics/

# Outline

1. Class Logistics ✓
2. Introduction to Biological Data ✓
3. Overview of Data Structures ✓
4. Applications of Data Structures in Bioinformatics
5. Summary

# Summary

1. **Data Deluge**: Exponential increase in the amount of biological data being generated.

2. Need for efficient methods to **organize, manage, store, and organize** data.

3. Selecting the appropriate data structures enables the efficient use of large-scale biological data.

# Outline

1. Class Logistics ✓
2. Introduction to Biological Data ✓
3. Overview of Data Structures ✓
4. Applications of Data Structures in Bioinformatics ✓
5. Summary ✓

# References


1. Sarkar RR (2016) The Big Data Deluge in Biology: Challenges and Solutions. Global J Technol Optim 7: e103.
2. O'Neil, S. T. (2019). A primer for computational biology. Pressbooks. (https://open.oregonstate.education/computationalbiology/)
3. Machanick, P (2019). Data Structures and Algorithms for Bioinformatics. RAMpage Research.
4. Classification of Data Structures (https://findtodaysnotes.wordpress.com/classification-of-data-structures/)
5. Limeri, L., & Reid, J. (2023, January 11). Introductory Biology 2. Pressbooks. (https://raider.pressbooks.pub/biology2/)
6. Kim, W., Haukap, L. NemoProfile as an efficient approach to network motif analysis with instance collection. BMC Bioinformatics 18 (Suppl 12), 423 (2017).
7. Wu, T.D. Bitpacking techniques for indexing genomes: I. Hash tables. Algorithms Mol Biol 11, 5 (2016).
8. Pathways of Metabolism | British Society for Cell Biology. (n.d.-b). (https://bscb.org/learning-resources/softcell-e-learning/pathways-of-metabolism/)
9. Hendrix, D. A. (2019, October 3). Applied Bioinformatics. Pressbooks. (https://open.oregonstate.education/appliedbioinformatics/)