Introduction to bioinformatics and computational biology

What is the goal?

- To develop computer algorithms and theory to interpret large biological data and to understand complex biological systems
- An interdisciplinary enterprise:
  - Biology
  - Chemistry
  - Physics
  - Statistics /applied math
  - Computer Science
  - Engineering

What is Bioinformatics?

- **Bioinformatics** is defined as the application of tools of computation and analysis to the capture and interpretation of biological data. It is an interdisciplinary field, which harnesses computer science, mathematics, physics, and biology.
- **Bioinformatics** is the new science at the interface of molecular biology and computer science that seeks to develop better ways to explore, analyze, and understand this vast wealth of genomic data.

What is Bioinformatics?

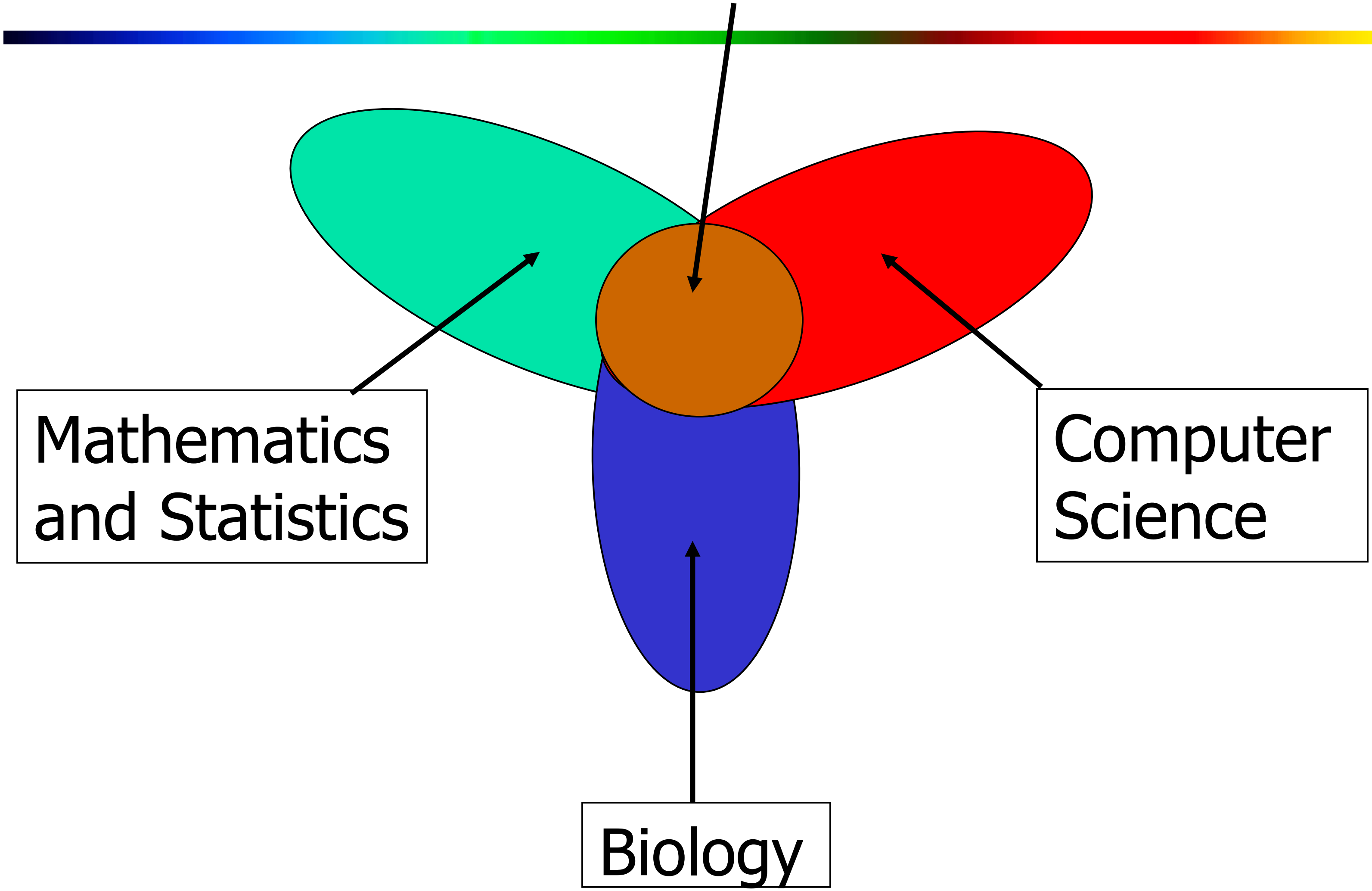
- **Other Definitions**

  - **Bioinformatics** is the field of science in which biology, computer science, and information technology merge into a single discipline. (**The National Centre for Biotechnology Information (NCBI 2001))**

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

http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html

What is Bioinformatics?

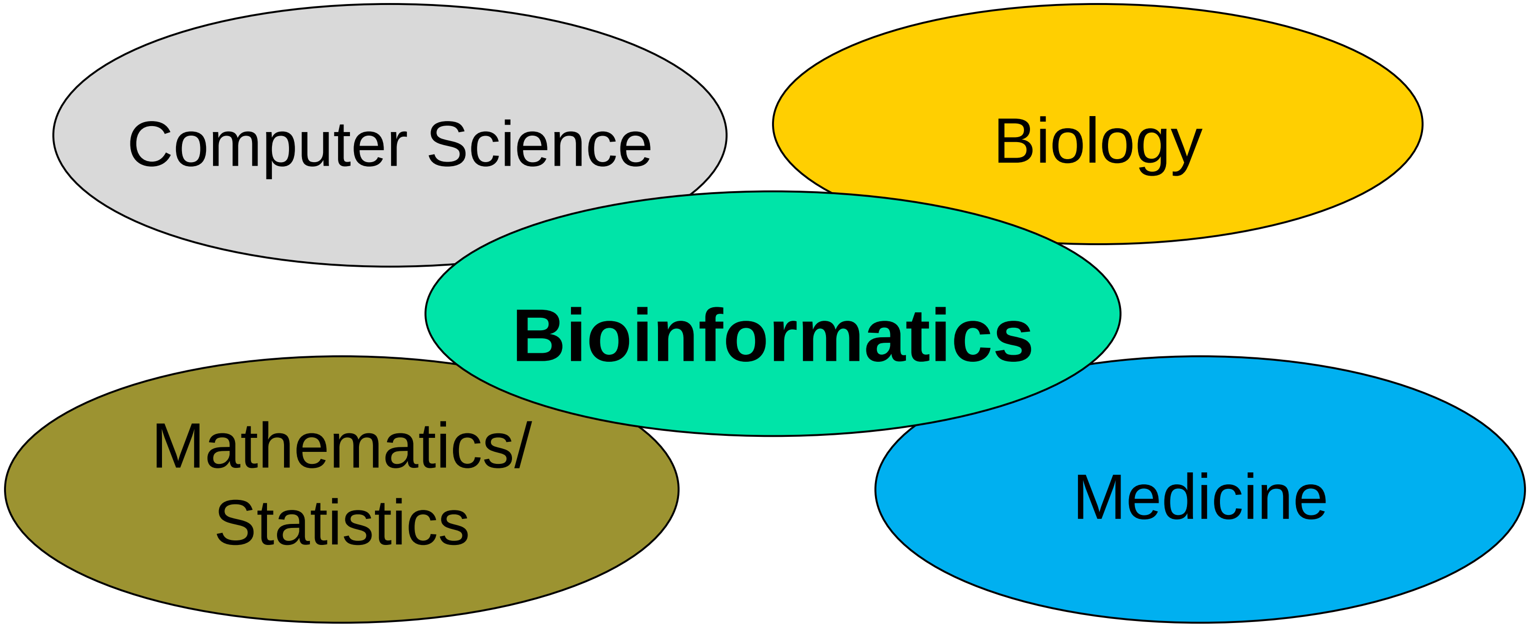


Mathematics and Statistics

Computer Science

Biology

What is Bioinformatics?



Areas of current and future development of bioinformatics

- Molecular biology and genetics
- Phylogenetic and evolutionary sciences
- Different aspects of biotechnology including pharmaceutical and microbiological industries
- Medicine
- Agriculture
- Eco-management

Bioinformatics and Data Mining
What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (**non-trivial**, implicit, **previously unknown** and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: **a misnomer**?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

Bioinformatics and Data Mining

- Bioinformatics involves:
  - the design of new algorithms and statistics for large molecular biology data sets analysis .
  - the design and construction of software that enable information extraction.
  - the analysis and interpretation of biological data including nucleotide and amino acid sequences, and protein structures using developed software tools.
- We can imply that Bioinformatics is the mining of biological datasets

Computational Biology

- **Computational biology** is the application of mathematical methods and computer algorithms to biological problems.

- Simply put, it is about studying biology using computational techniques, which further the understanding of the science.

- Bioinformatics can be regarded as a branch of computational biology.

Why Bioinformatics?

- It provide an avenue for **a better understanding of complex disorders such** as PD, increasing our ability to work toward improved treatments or cures.

- It solves **existing biological questions** using existing and newly developed algorithms and software tools.

- Provides a **better data management and visualization analysis** for the growing amount of available datasets.

Why Bioinformatics?

- It provide an avenue for **a better understanding of complex disorders such** as PD, increasing our ability to work toward improved treatments or cures.
- Rapid explosion in our ability to acquire biological data

  How can we find robust patterns in these data?

- Recognition that biological phenomena are enormously complex and biological problems benefit from interdisciplinary approaches

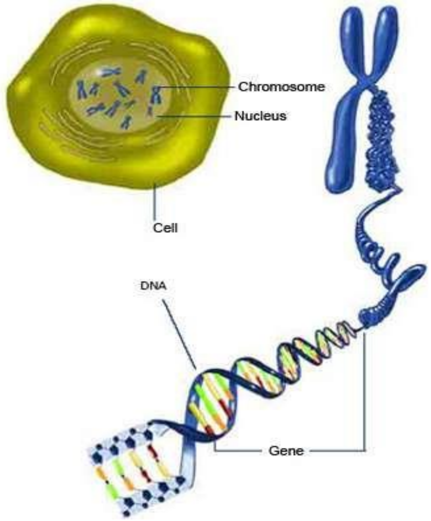
How can we understand, predict, and manipulate these systems?

Why Bioinformatics?

- Exponential **growth of investments**
- Constant **deficit of trained professionals**
- **Diversification of bioinformatics applications**

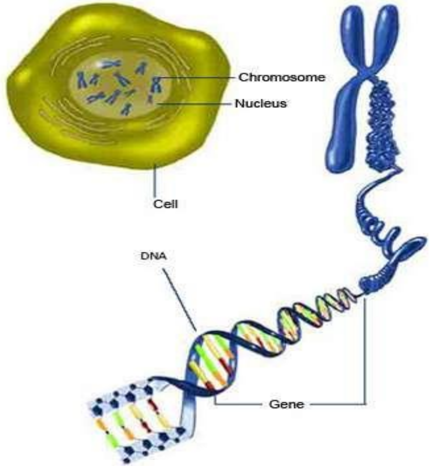
Key Concepts

- The **genome** is the complete genetic material of an organism. It contains all the information needed to build and maintain the organism



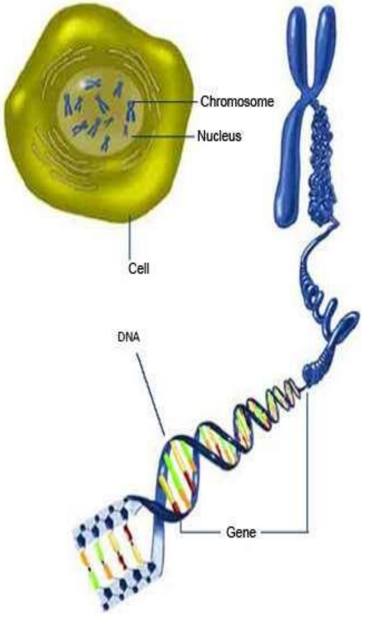
Key Concepts

- A **chromosome** is a continuous strands of DNA wrapped around a protein scaffold
- The human genome consists of 23 pairs of chromosomes, one member of each pair coming from each parent



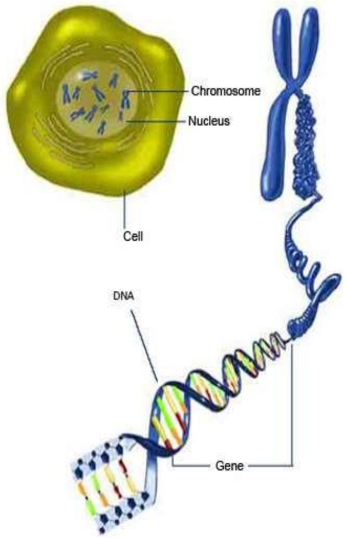
Key Concepts

- A **gene** is a segment of DNA that has information for making a specific type of protein



Key Concepts

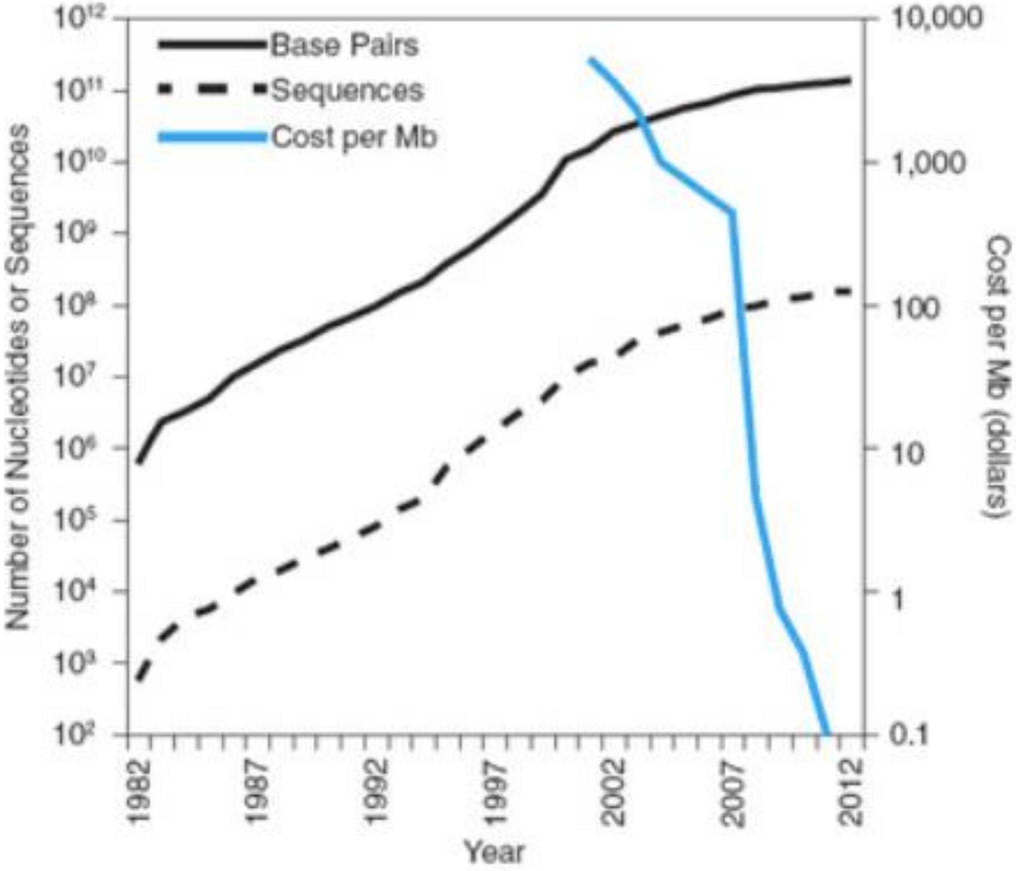
- The **ribosome** is a large molecular machine that serves as the site where biological proteins are synthesized



Types of Data available

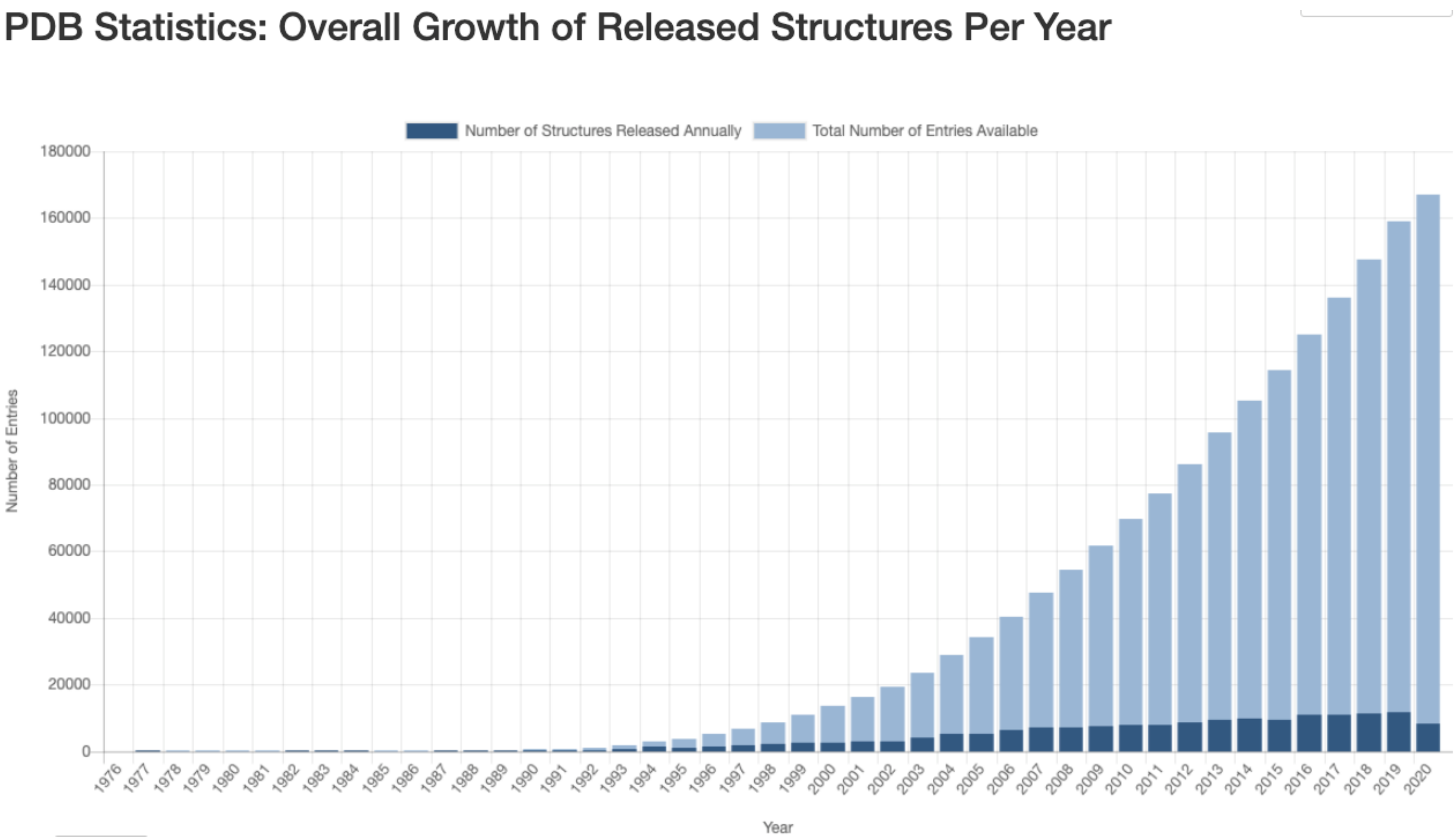
- DNA/RNA Sequence Data (ACTG-DNA, ACUG-RNA)
- Protein Sequence Data
- Protein structure Data
- Protein function
- Gene expression Data
- Biomolecular interactions
- Single Nucleotide Polymorphisms (SNPs)
- Molecular pathways
- Disease information
- And many more …

GenBank Data Growth



Clair, C. S., & Visick, J. E. (2013).

Growth of PDB



PDB Statistics: Overall Growth of Released Structures Per Year

https://www.rcsb.org/stats/growth/growth-released-structures

Popular Databases and Type of Data Stored

Database	Data Stored
GenBank	DNA/RNA Sequence Data
UniProtKB	Protein Sequence Data
Protein DataBank	Protein structure Data
Gene Ontology	Protein function Data
Gene Expression Omnibus (GEO)	Gene expression Data
dbSNP	SNPs
KEGG PATHWAY	Molecular pathways

Other Databases

- GSDB: Genome Structure Database
  - Store Three-Dimensional Chromosome and Genome Structure

Problems affecting Analysis

- Database Data Quality
  - Additional Noise (ERRORS) stemming from
    - Incorrect interpretation of experiments and incorrect handling
    - Incorrect parameter entry in public databases
    - Spelling mistakes
    - Annotation errors
    - Frame shift errors

Problems affecting Analysis

- Database Redundancy
  - Data from different experimental approach for same organism
  - Sequence discrepancies and variation
  - As a result of redundancy if data used for training and testing are closely related. Then the developed model will lack ability to generalize.

“The Million Dollar Question”

- What do we do with these huge amount of data?
- Alternatively, should our question be, What **can we get from** these huge and growing amount of data?

What do we do with these huge amount of data?

- Store
- Search
- Analyze, Annotate, visualize and Recognize patterns
- Build biological models and make Predictions

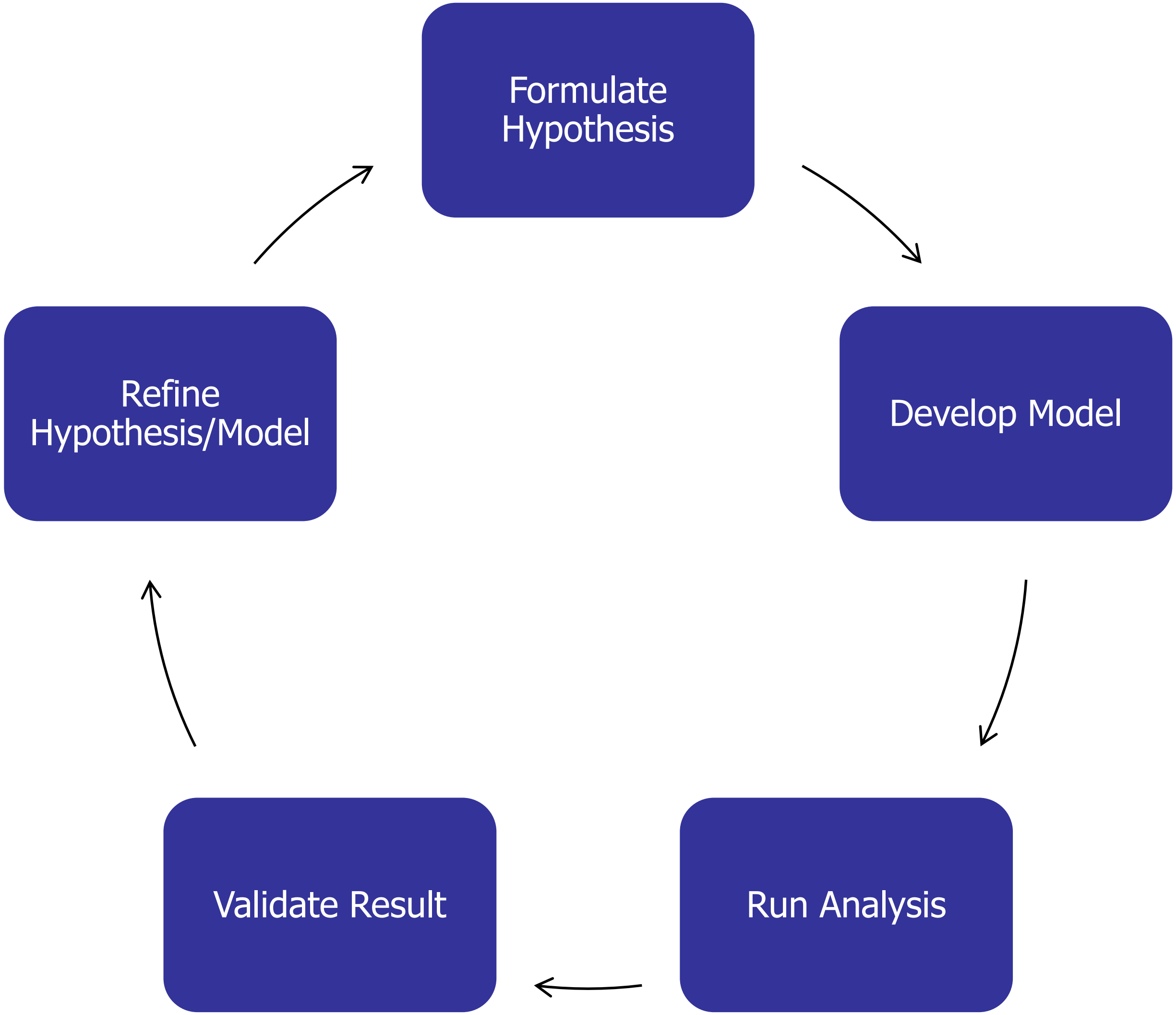
What do we do with these huge amount of data?

- Need to have some biological knowledge to know the questions to ask.
- Need to have knowledge of the data to know what possible questions it could answer.
  - What are the research areas/topics in bioinformatics?

Common Tools used for Analysis

- BLAST(Basic Local Alignment Search Tool**)**
  - Sequence similarity(“Homology”) searching
- Clustal Omega(Protein)/ MUSCLE (DNA Alignment)
  - Sequence Alignment
- GenScan
  - To identify complete gene structures in genomic DNA.
- UniFold
  - RNA Structure
- JPred
  - Protein Structure Prediction
- PHYLIP
  - Phylogenetics: Inferring evolutionary relationships among biological entities.

Methodology for Tool development in Bioinformatics



Formulate Hypothesis

Develop Model

Run Analysis

Validate Result

Refine Hypothesis/Model

References

- Clair, C. S., & Visick, J. E. (2013). *Exploring bioinformatics*. Jones & Bartlett Publishers
- Bayat, A. (2002). Bioinformatics. Bmj, 324(7344), 1018-1022.
- http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html
- https://www.nature.com/subjects/systems-biology
- Bioinformatics: The Machine Learning Approach, Pierre Baldi and Soren Brunak, 2001, MIT press.
- Dr. Jianhua Ruan, Department of Computer Science, UTSA
- Dr. Jianlin Cheng, Deparmtment of Electrical Engineering and Computer Science, University of Missouri, Columbia