# Bioinformatics

## CS300

## The Great Review

**Fall 2017**

**Oliver Bonham-Carter**

# Course Summary

\subsection*{\textbf{Academic Bulletin Description}}

An introduction to the development and application of methods, from the computational and information sciences, for the investigation of biological phenomena. In this interdisciplinary course, students integrate computational techniques with biological knowledge to develop and use analytical tools for extracting, organizing, and interpreting information from genetic sequence data. Often participating in team-based and hands-on activities, students implement and apply useful bioinformatics algorithms. During a weekly laboratory session students employ cutting-edge software tools and programming environments to complete projects, reporting on their results through both written assignments and oral presentations. Prerequisites: BIO 221 and FSBIO 201, or CMPSC 111. Distribution Requirements: QR, SP.

# Course Objectives

```
\subsection*{\textbf{Course Objectives}}

Students successfully completing this class will have developed:
\begin{enumerate}
  \item A "big-picture" view of bioinformatics.

  \item An understanding of the objectives and limitations of
bioinformatics.

  \item An understanding of the biological foundations of bioinformatics
(genes and genomes, gene expression, etc.).

  \item An understanding of the computational foundations of
bioinformatics (programming, databases, etc.).

  \item An understanding of how genetic information is obtained and
processed.

  \item The ability to use basic bioinformatics software tools to study
genetic information.
\end{enumerate}
```

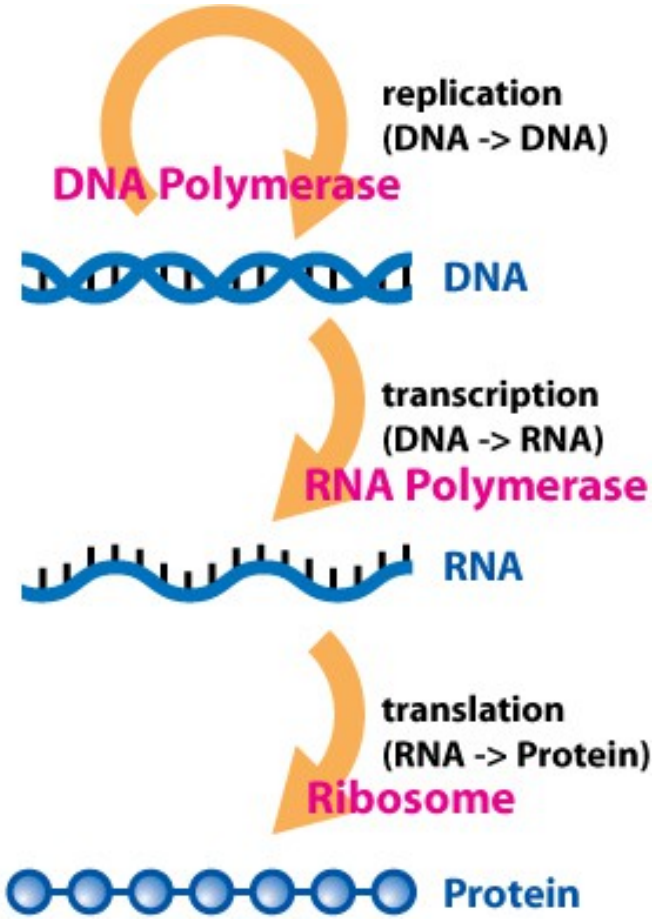
# How Did We Meet Our Objectives?

**Let’s go back and revisit some of our discussions and slides.**

# We Started With ...

**The Central Dogma Of Biology**

# The Central Dogma of Molecular Biology



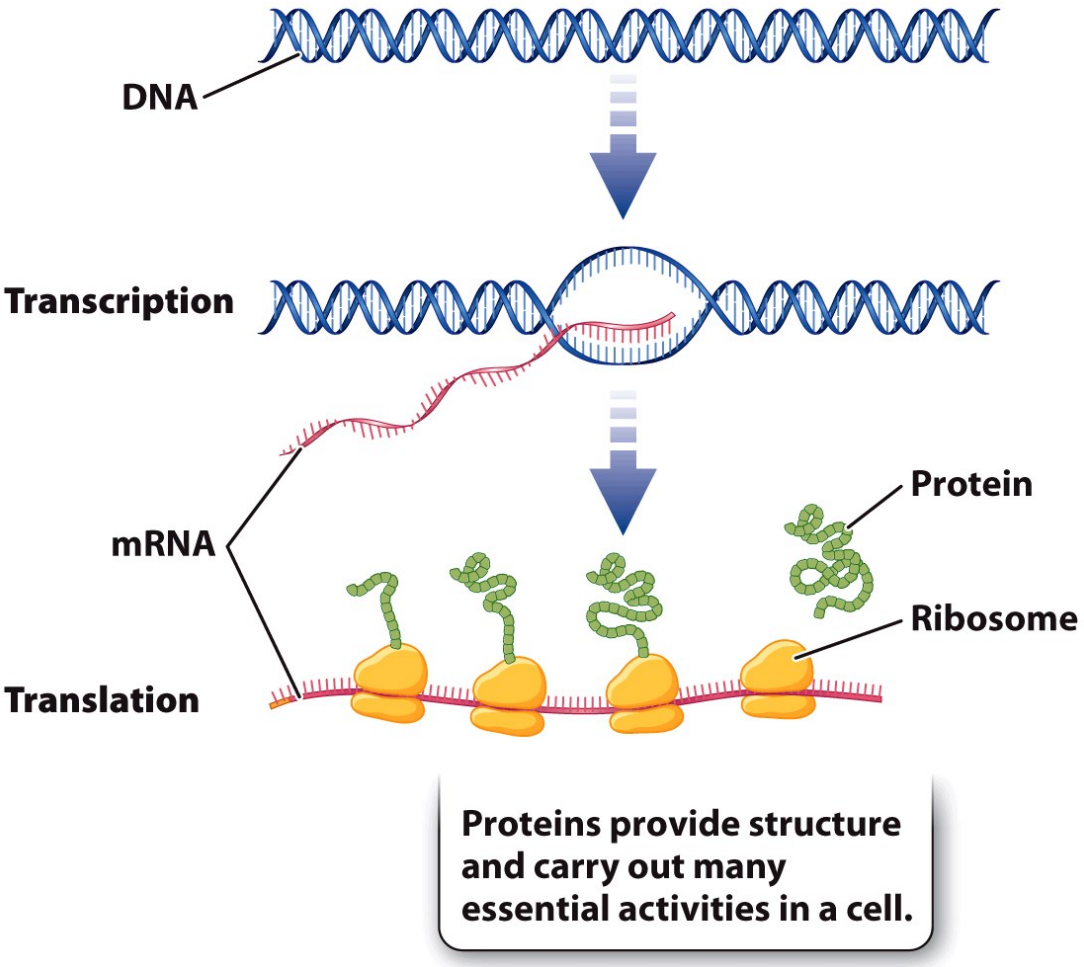
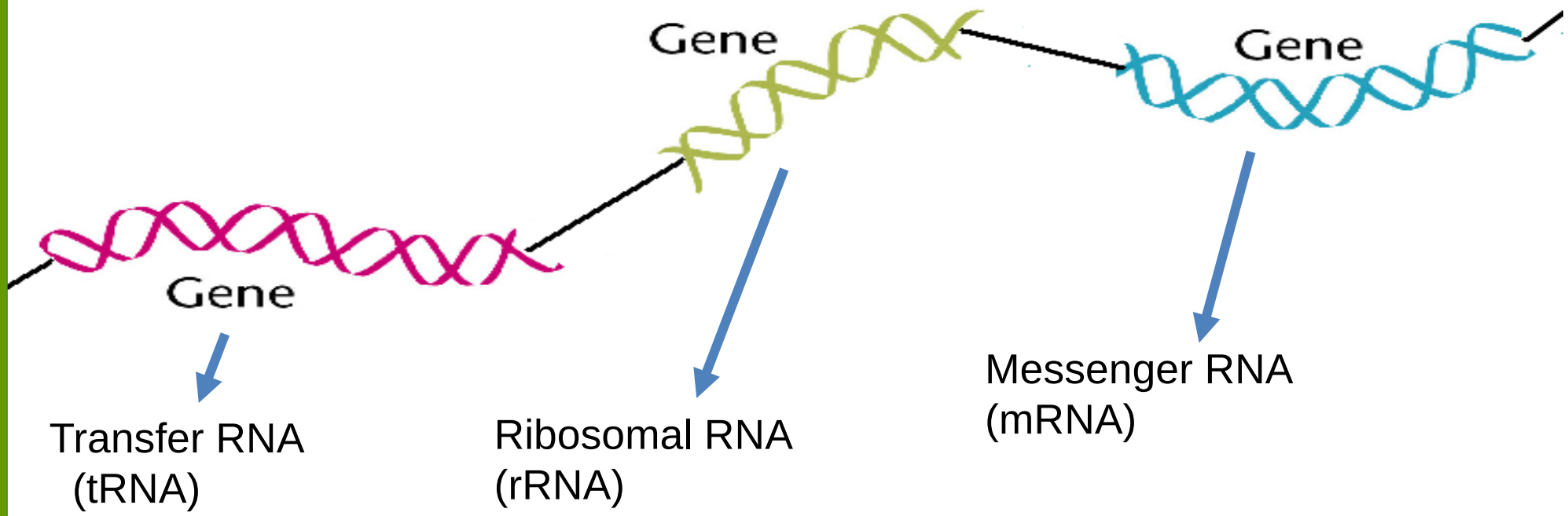




**Figure 3.3**
*How Life Works*

# Transcription

- Transcribe specific regions of DNA – genes
  - Human genome ~25,000 genes (just 1.5% of genome)
- RNA is the direct product of transcribing a gene (DNA)
  - DNA -> RNA
  - same language (nucleotides)



# The Genetic Code: RNA into Protein

- Triplet code
  - Combinations of three nucleotides code for one amino acid
  - Three nucleotides = codon
- Redundancy
  - Sometimes >1 codon codes for same amino acid
  - 20 amino acids, 64 possible codons

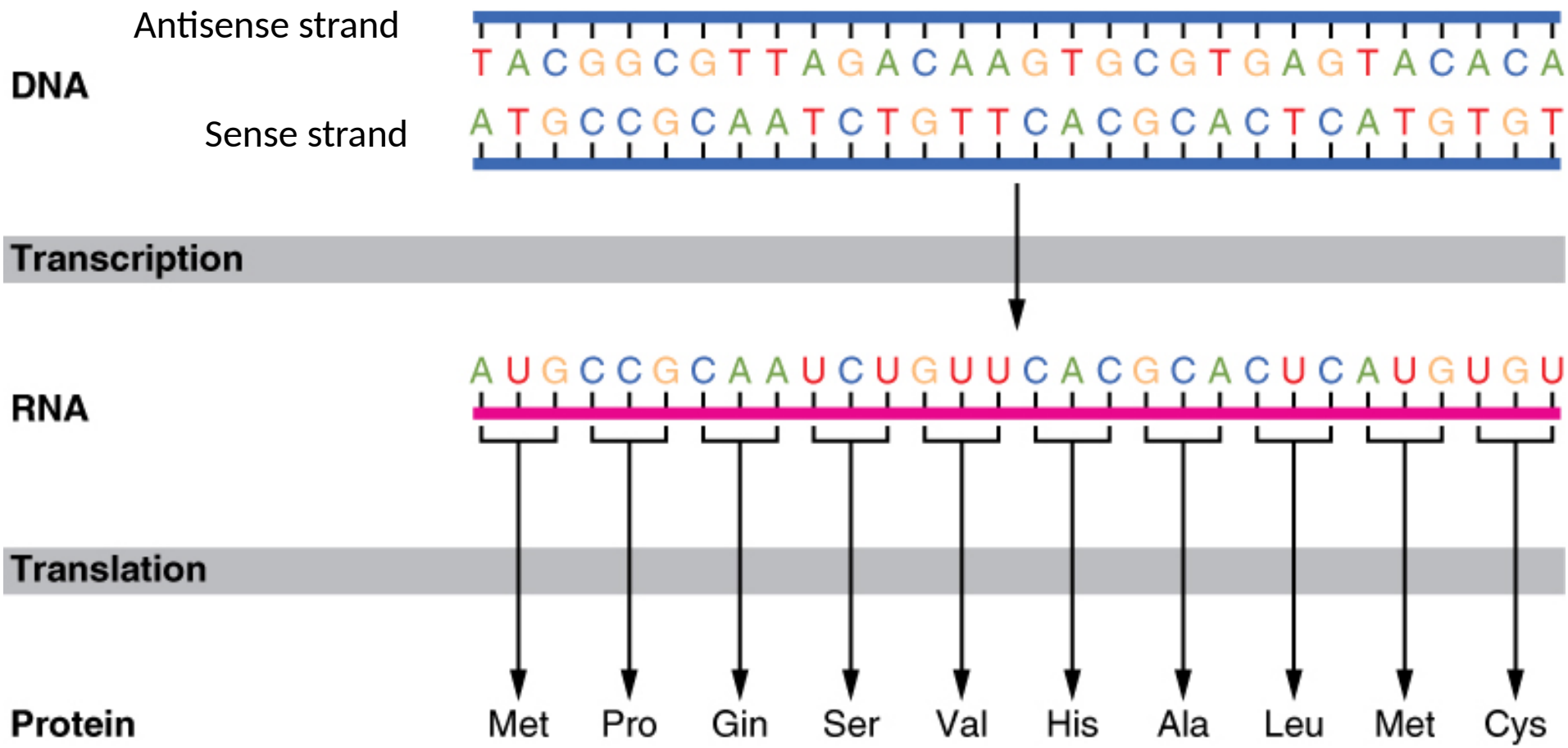
- Start and Stop codons
  - First codon of many transcripts is “AUG”, which codes for methionine
  - Codons UAA, UAG, and UGA indicate the end of the transcript

**Standard genetic code**

| 1st base | 2nd base | | | | | | | | 3rd base |
|---|---|---|---|---|---|---|---|---|---|
| | T | | C | | A | | G | | |
| T | TTT | (Phe/F) Phenylalanine | TCT | (Ser/S) Serine | TAT | (Tyr/Y) Tyrosine | TGT | (Cys/C) Cysteine | T |
| | TTC | | TCC | | TAC | | TGC | | C |
| | TTA | (Leu/L) Leucine | TCA | | TAA[B] | Stop (*Ochre*) | TGA[B] | Stop (*Opal*) | A |
| | TTG | | TCG | | TAG[B] | Stop (*Amber*) | TGG | (Trp/W) Tryptophan | G |
| C | CTT | (Leu/L) Leucine | CCT | (Pro/P) Proline | CAT | (His/H) Histidine | CGT | (Arg/R) Arginine | T |
| | CTC | | CCC | | CAC | | CGC | | C |
| | CTA | | CCA | | CAA | (Gln/Q) Glutamine | CGA | | A |
| | CTG | | CCG | | CAG | | CGG | | G |
| A | ATT | (Ile/I) Isoleucine | ACT | (Thr/T) Threonine | AAT | (Asn/N) Asparagine | AGT | (Ser/S) Serine | T |
| | ATC | | ACC | | AAC | | AGC | | C |
| | ATA | | ACA | | AAA | (Lys/K) Lysine | AGA | (Arg/R) Arginine | A |
| | ATG[A] | (Met/M) Methionine | ACG | | AAG | | AGG | | G |
| G | GTT | (Val/V) Valine | GCT | (Ala/A) Alanine | GAT | (Asp/D) Aspartic acid | GGT | (Gly/G) Glycine | T |
| | GTC | | GCC | | GAC | | GGC | | C |
| | GTA | | GCA | | GAA | (Glu/E) Glutamic acid | GGA | | A |
| | GTG | | GCG | | GAG | | GGG | | G |

# Translation

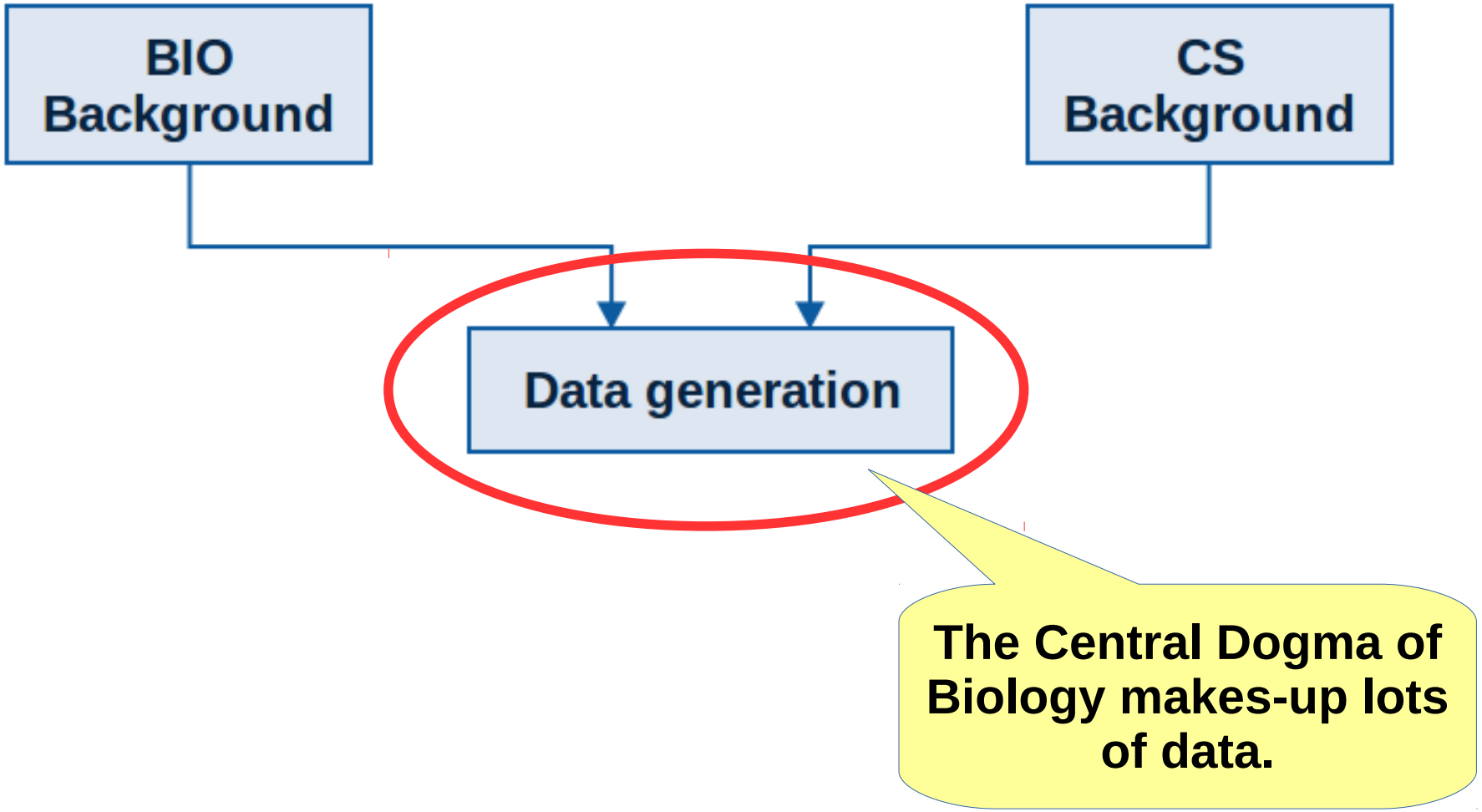
- The information from DNA is rewritten in a new language: RNA



# We Talked About...

***Data Generation*:**
**Or where the data comes from for research in bioinformatics**

# Course Outline



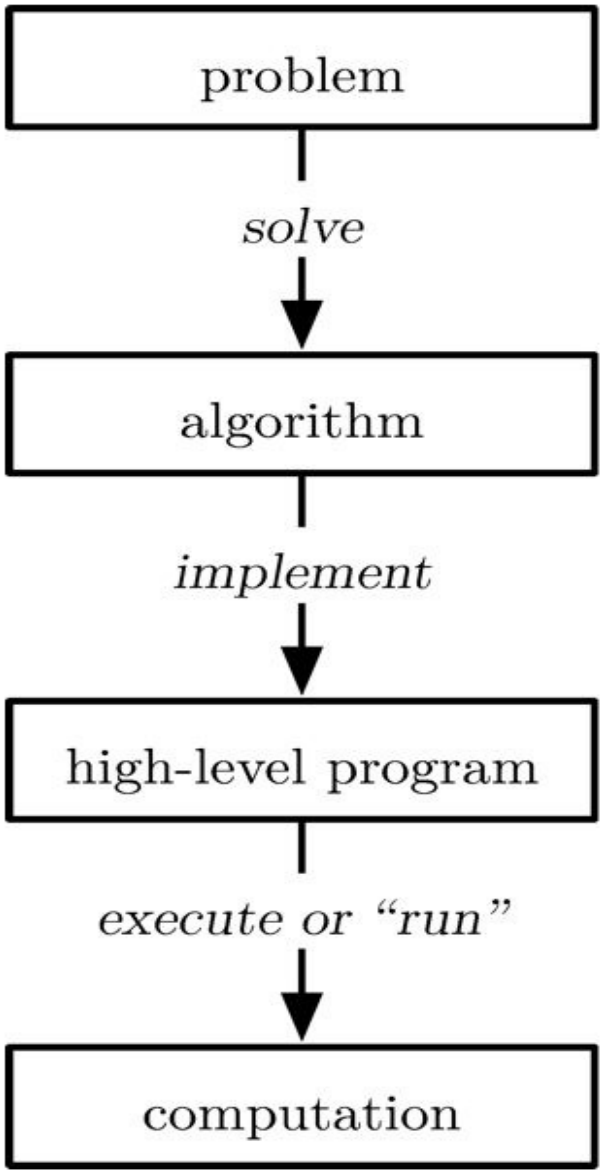
# Computation

## Python – overall view

| | | | |
|---|---|---|---|
| Learning curve | | ☺☺☺☺☺ | Easy to learn, yet powerf |
| Readibility of a python program | | ☺☺☺☺☺ | |
| Community, availability of open source modules | | ☺☺☺☺ | (for bioinformatics, CPAN sligthly bigger) |
| Programming paradigms | | ☺☺☺☺☺ | Multi paradigm (Object Oriented, structured, functional, etc..) |
| Execution speed | | 😐😐😐 | Interpreted language; importance of programm effort over computer effo |

Notes:

- This talk is full of tables like this
- They only reflect *my* opinion (biologist with 3-4 years experience)

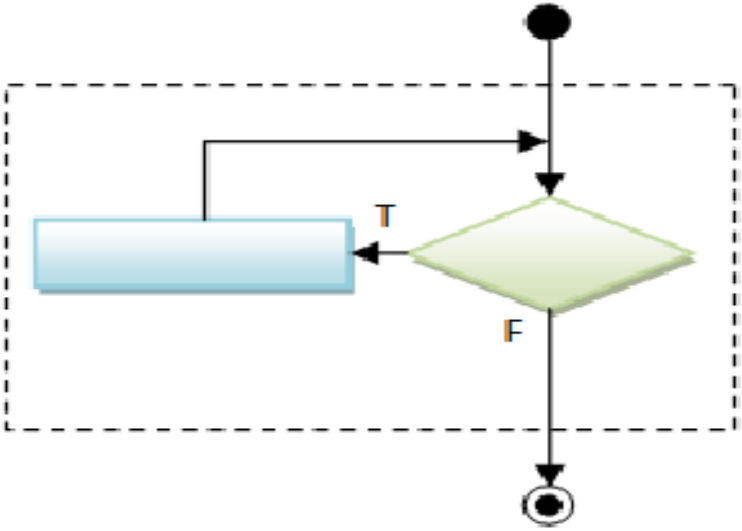


# Computation: How to Use

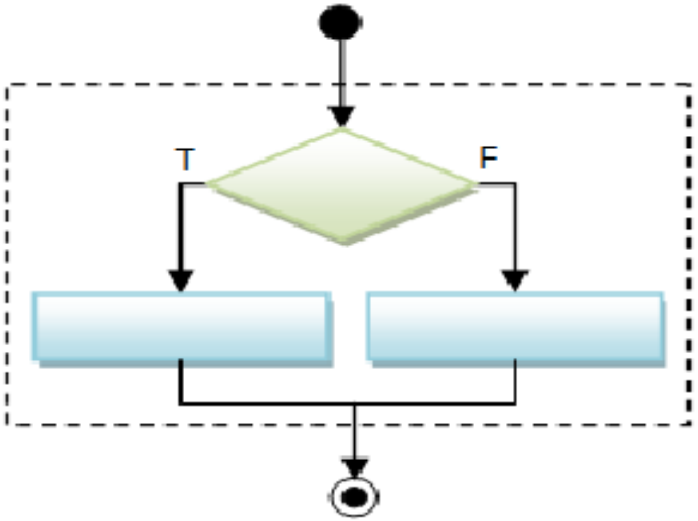
Bitbucket

Linux

python




Loop (Iteration)

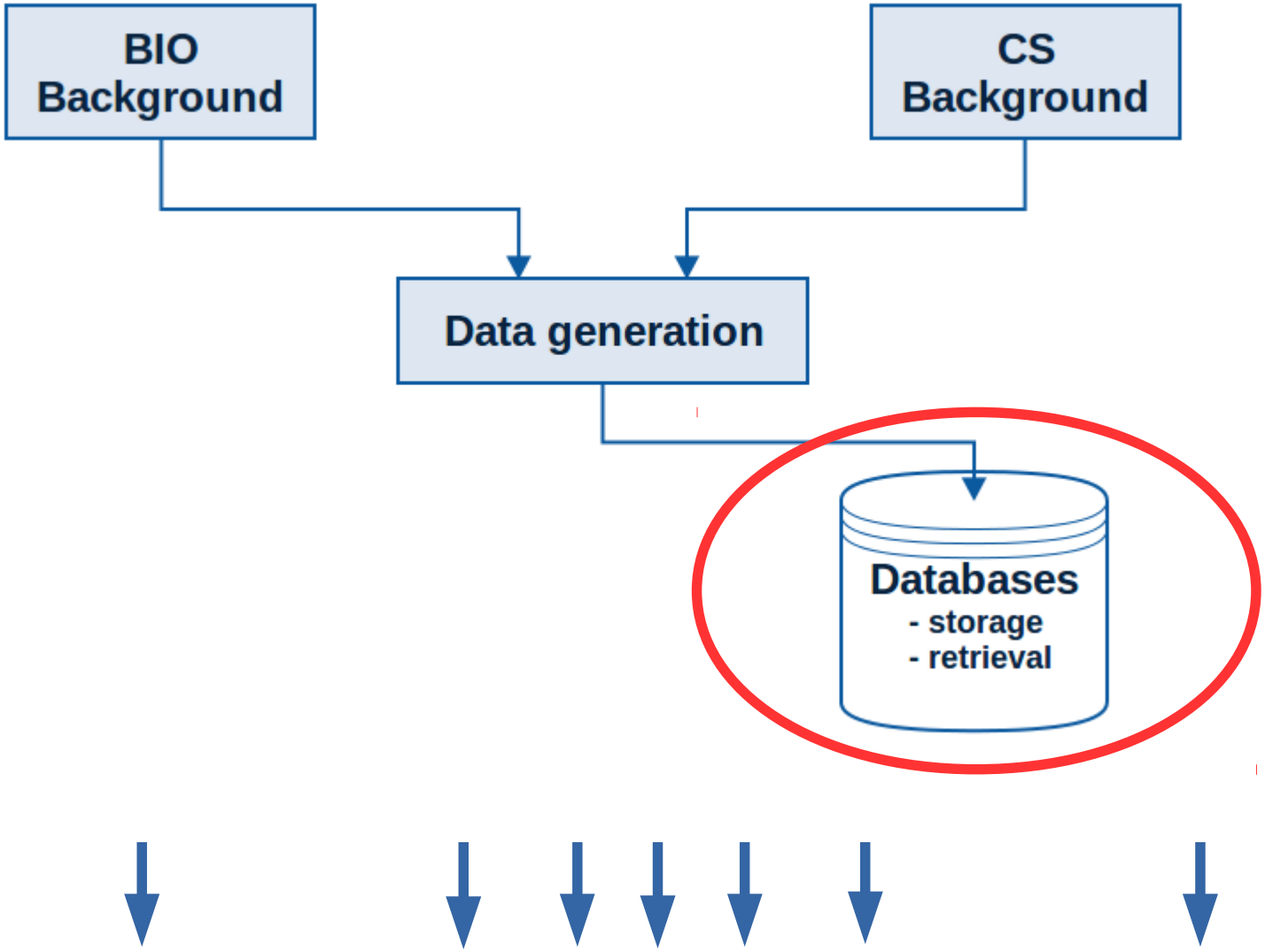



Conditional (Decision)

# We Talked About...

***Databases*:**
**Places where data is stored for further research**

# Course Outline



# Biological Data and Databases

- To learn how to use a Web-based genomic databases and tools.
- To understand the types of information stores in genomic databases.
- To learn how to use different interfaces to find and retrieve genomic information.
- Write Python program to find patterns (start and stop codons) in DNA sequences

# Biological Databases

NCBI Resources How To Sign In to NCBI

NCBI
National Center for Biotechnology Information

All Databases | Search

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

## Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

**Submit**
Deposit data or manuscripts into NCBI databases

**Download**
Transfer NCBI data to your computer

**Learn**
Find help documents, attend a class or watch a tutorial

**Develop**
Use NCBI APIs and code libraries to build applications

**Analyze**
Identify an NCBI tool for your data analysis task

**Research**
Explore NCBI research and collaborative projects

**Popular Resources**
- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

**NCBI News & Blog**

NCBI to assist in Southern California genomics hackathon in January
30 Nov 2017
From January 10-12, 2018, the NCBI will help with a bioinformatics hackathon in

December 6th NCBI Minute: Keeping Current and Getting Help with NCBI Resources
30 Nov 2017
In the next NCBI Minute on Wednesday

November 28th NCBI Minute: An update

# NCBI Browser

# Biological Databases

UniProt

UniProtKB

Advanced

Search

BLAST Align Retrieve/ID mapping Peptide search

Help Contact

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB

UniProt Knowledgebase

Swiss-Prot (556,196)

Manually annotated and reviewed.

TrEMBL (98,705,220)

Automatically annotated and not reviewed.

UniRef

Sequence clusters

UniParc

Sequence archive

Proteomes

Supporting data

Literature citations

Taxonomy

Subcellular locations

Cross-ref. databases

Diseases

Keywords

News

Forthcoming changes

There are currently no changes planned

UniProt release 2017_11

Sex determination in insects: 50 ways to achieve sex-specific splicing

UniProt release 2017_10

Of smell and social life

UniProt release 2017_09

News archive

Getting started

Text search

Our basic text search allows you to search all the resources available

UniProt data

Download latest release

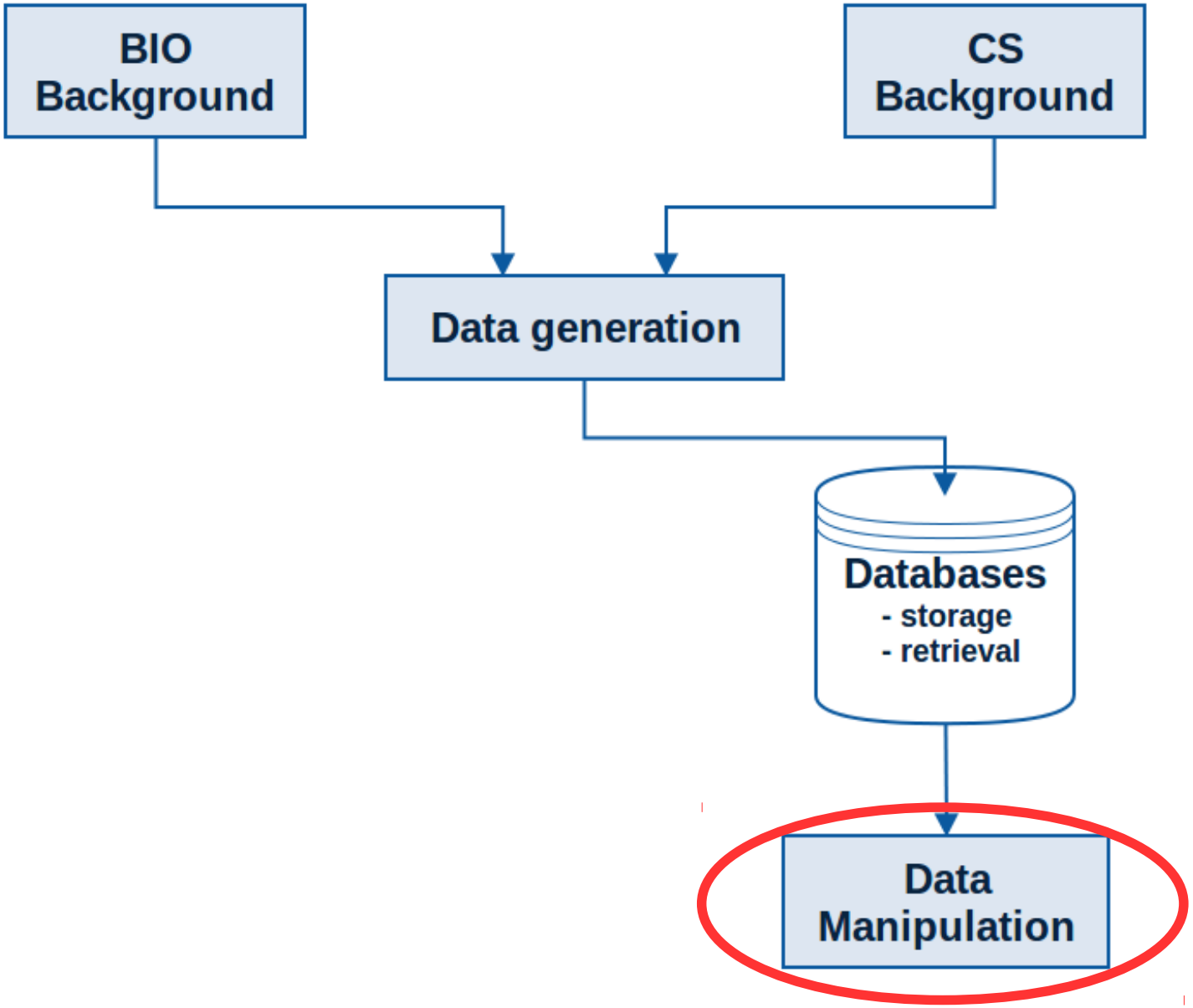
Get the UniProt data

# UniProt Browser

# We Talked About...

***Data Manipulation*:**
**How we begin to find meaning in the data**

# Course Outline



# Data Manipulation

- To become familiar with tools that can be used to manipulate sequences in a variety of ways and make basic comparisons between sequences
- Understand the structure and orientation of string representations of DNA and protein sequences
- Gain experience with string manipulation using Python and its application to DNA and protein sequence data.
- Understand how genetic information is computationally decided and appreciate the important complications in working with sequences (introns/exons, start codons, template/nontemplate strand orientation, etc)

# Databases...

ALLEGHENY COLLEGE

## Filter by

Reviewed (556,196)
Swiss-Prot

Unreviewed (98,705,220)
TrEMBL

## Popular organisms

Human (161,042)

Rice (122,677)

A. thaliana (89,135)

Mouse (83,100)

Zebrafish (59,673)

Other organisms

Go

## View by

Results table

Taxonomy

Keywords

Gene Ontology

BLAST | Align | Download | Add to basket | Columns

1 to 25 of 99,261,

| Entry | Entry name | Protein names | Gene names | Organism | Length |
|---|---|---|---|---|---|
| Q91G88 | 006L_IIV6 | Putative KilA-N domain-containing ... | IIV6-006L | Invertebrate iridescent virus 6 (IIV-6) (Chilo iridescent virus) | 352 |
| Q6GZW6 | 009L_FRG3G | Putative helicase 009L | FV3-009L | Frog virus 3 (isolate Goorha) (FV-3) | 948 |
| Q91G70 | 026R_IIV6 | Uncharacterized protein 026R | IIV6-026R | Invertebrate iridescent virus 6 (IIV-6) (Chilo iridescent virus) | 59 |
| Q6GZU9 | 027R_FRG3G | Uncharacterized protein 027R | FV3-027R | Frog virus 3 (isolate Goorha) (FV-3) | 970 |
| Q197D7 | 023R_IIV3 | Uncharacterized protein 023R | IIV3-023R | Invertebrate iridescent virus 3 (IIV-3) (Mosquito iridescent virus) | 106 |
| Q91G65 | 032R_IIV6 | Uncharacterized protein 032R | IIV6-032R | Invertebrate iridescent virus | 100 |
| Q6GZU3 | 033R_FRG3G | Transmembrane protein 033R | | | |

UniProt Browser

# Pulling Data From Databases...

## Homo sapiens genomic DNA, chromosome 21q

GenBank: BA000005.3

FASTA Graphics

Go to:

```
LOCUS       BA000005            33543332 bp    DNA     linear   CON 12-JUL-2008
DEFINITION  Homo sapiens genomic DNA, chromosome 21q.
ACCESSION   BA000005
VERSION     BA000005.3
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1
```

## Homo sapiens genomic DNA, chromosome 21q

GenBank: BA000005.3

GenBank Graphics

```
>BA000005.3 Homo sapiens genomic DNA, chromosome 21q
CATGTTTCCACTTACAGATCCTTCAAAAAGAGTGTTTCAAAACTGCTCTATGAAAAGGAATGTTCAAC
TGTGAGTTAAATAAAAGCATCAAAAAAAAGTTTCTGAGAATGCTTCTGTCTAGTTTTTATGTGAAGAT
TTCCATTTTCTCTATAAGCCTCAAAGCTGTCCAAATGTCCACTTGCAGATACTACAAAAAGAGTGTTT
AAAGTGCTCAATGAAAAGGAATGTTCAGCTCTGTGAGTTAAATGCAAACATCACAAATAAGTTTCTGA
ATGCTTCTGTCTAGTTTTTATGGGAAGATAATTCCGTGTCCAGCGAAGGCTTCAAAGCTTTCAAAATA
CACTTGCAAATTCTACAAAAAGAGTGTTTCAAAGCTGCTTTATCAAAAGAAAGTTTCAACTCTGTGAG
GAATGTGCACATCACAAAGAAGTTTCTGAGAATGCCTTCAGTCTGGTTTTTATGTGAAGATATTCCCT
```

Molecule processing

| Feature key | Position(s) | Description | Actions | Graphical vi |
|---|---|---|---|---|
| Transit peptide i | 1 – 77 | Mitochondrion<br>Sequence analysis | Add BLAST | |
| Chain i (PRO_0000024369) | 78 – 581 | Serine/threonine-protein kinase PINK1, mitochondrial | Add BLAST | |

Amino acid modifications

| Feature key | Position(s) | Description | Actions | Graphical view |
|---|---|---|---|---|
| Modified residue i | 228 | Phosphoserine; by autocatalysis<br>1 Publication | | |
| Modified residue i | 402 | Phosphoserine; by autocatalysis<br>1 Publication | | |

# Tools from Databases

NIH U.S. National Library of Medicine NCBI

BLAST® » blastn suite

Home Recent Results Saved

Sta l Nucleotide BLAST

blastn blastp blastx tblastn tblastx

## Enter Query Sequence

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

BA000005.3

Query subrange

From

To

Or, upload file Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

## Choose Search Set

Database Human genomic + transcript Mouse genomic + transcript Others (nr etc.):

Nucleotide collection (nr/nt)

Organism Optional

Enter organism name or id--completions will be suggested Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Optional Models (XM/XP) Uncultured/environmental sample sequences

Limit to Optional Sequences from type material

Entrez Query Optional

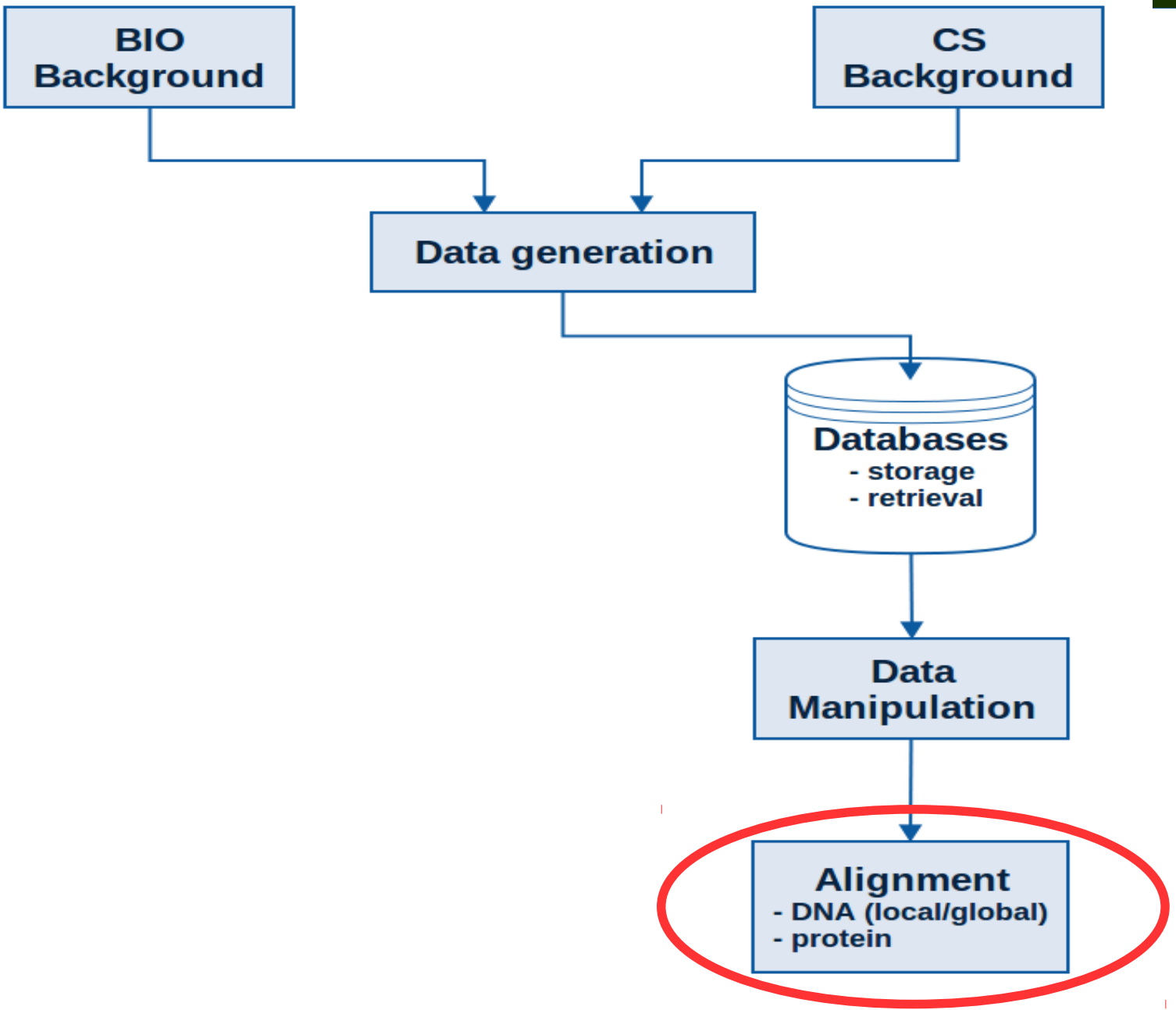
YouTube Create custom database

Enter an Entrez query to limit search

# We Talked About...

***Sequence Alignment*: Comparing sequences to discover similarities and differences**

# Course Outline



# Alignment to Compare ...

- DNA sequences
- Genes
- Proteins
- Organisms

- **Why do we compare these things?**
- **What is there to learn when we find that two *things* are the same? Not the same?**

# Sequence Alignment

## DNA - Nucleotides

- To understand the value of aligning genes and recognize the practical applications of this technique.
- To gain familiarity with the use of Web-based alignment tools to explore sequence similarity and understand how to modify their parameters.
- To know how the Needleman-Wunsch algorithm optimally aligns any two sequences.
- Understand how the Needleman-Wunsch algorithm can be modified to yield other alignments.

# Example Alignment score = 0

Let:

Match = +1

Mismatch = 0

Gap = -1

| | | C | A | C | G | T | A | T |
|---|---|---|---|---|---|---|---|---|
| | 0 | -1 | -2 | -3 | -4 | -5 | -6 | -7 |
| C | -1 | 1 | 0 | 1 | -2 | -3 | -4 | -5 |
| G | -2 | 0 | 1 | 0 | 0 | -1 | -2 | -3 |
| C | -3 | -1 | 0 | 2 | 1 | 0 | -1 | -2 |
| A | -4 | -2 | 0 | 1 | 2 | 1 | 1 | 0 |

```
CACGTAT
--CGCA-
```

```
CACGTAT
C--GCA-
```

```
CACGTAT
CGC--A-
```

# Pairwise Alignment
# Similarity and Relatedness

## Alignment of a gene from two closely related viruses

```
Hemagglutinin gene from virus A:   ATGAACGCAATACTCGTAGTT...
                                   ||||| ||||||||| ||||||
Hemagglutinin gene from virus B:   ATGAAGGCAATACTAGTAGTT...
```

Few Mismatches

## Alignment of a gene from two distantly related viruses

```
Hemagglutinin gene from virus A:   ATGAACGCAATACTCGTAGTT...
                                   ||| ||| ||| |||| |   |
Hemagglutinin gene from virus C:   ATGCACGAAATGCTCGGACCT...
```

Lots of Mismatches

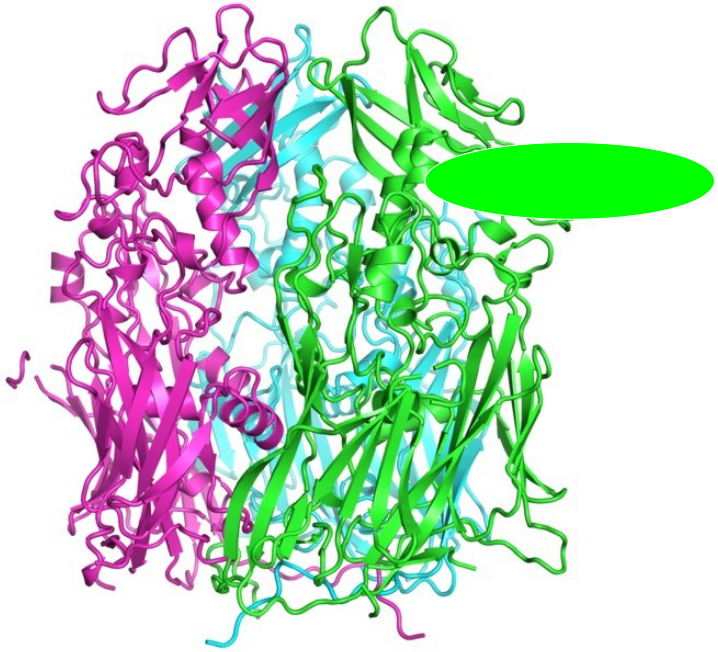
# Sequence Alignment

## Protein – Amino Acids

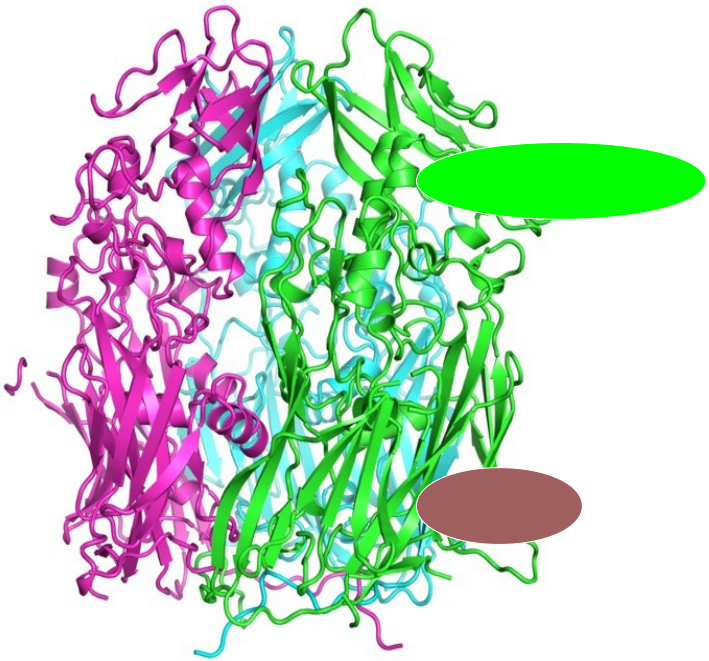
- Understand the use of a substitution matrix to score amino acid similarity in a protein sequence alignment.
- Gain experience using protein alignment to develop hypotheses about protein function based on sequence similarity.
- Know how protein alignment differs algorithmically from DNA alignment.
- Know how substitution matrix is developed and how different matrices might be used to produce better alignments in particular situations.

# Comparing Protein

- Two proteins (wildtype, non-wildtype) are compared to find causes of disorder.



Healthy



Unhealthy

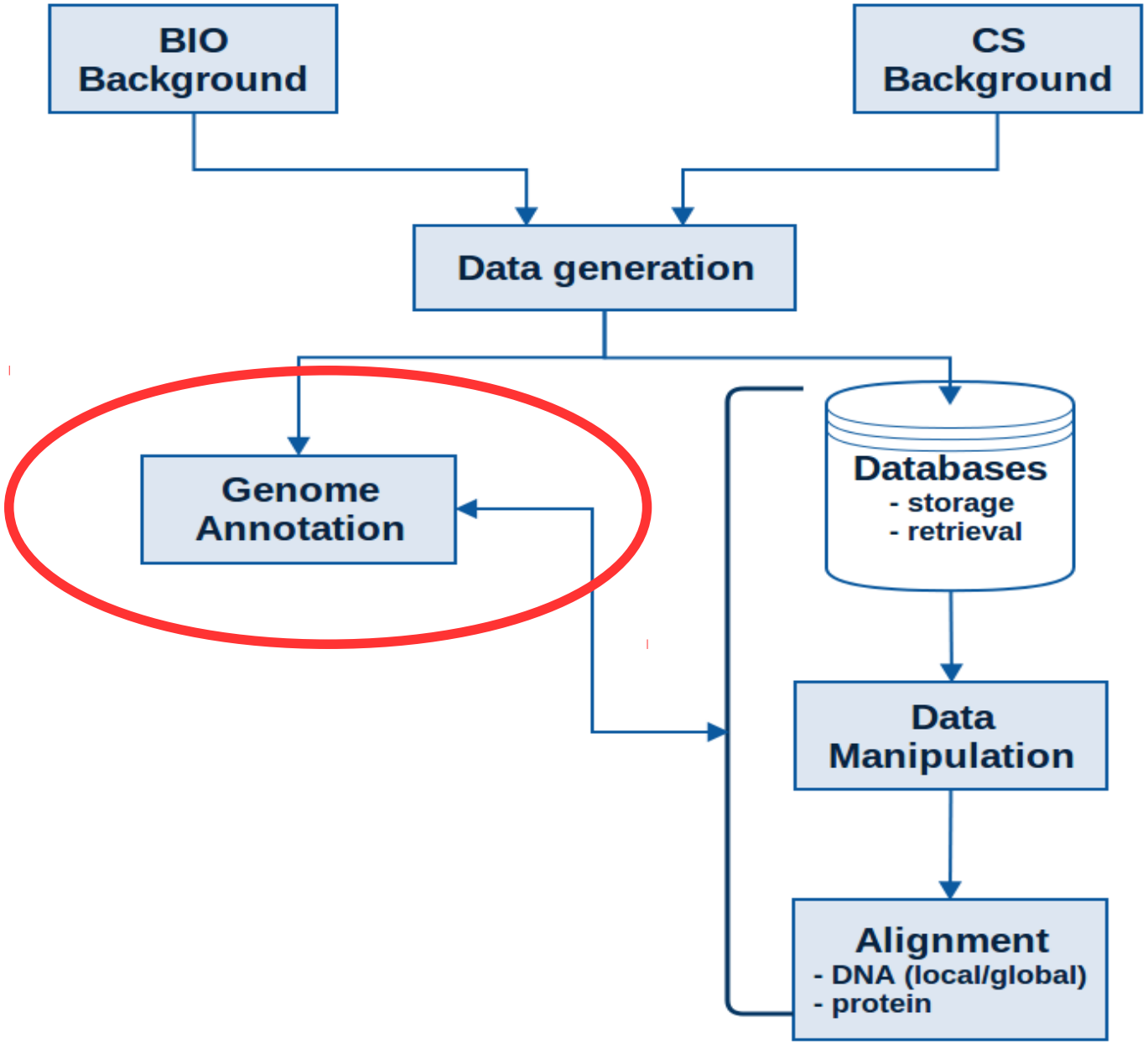
# Aligning Sequences To Locate Mutations

- A natural process that changes the DNA sequence
- A common process
  - during replication of the human genome a “typo” occurs every 100,000 or so nucleotides
  - that’s about 120,000 typos each time one of our cells divides
  - most are repaired

# We Talked About...

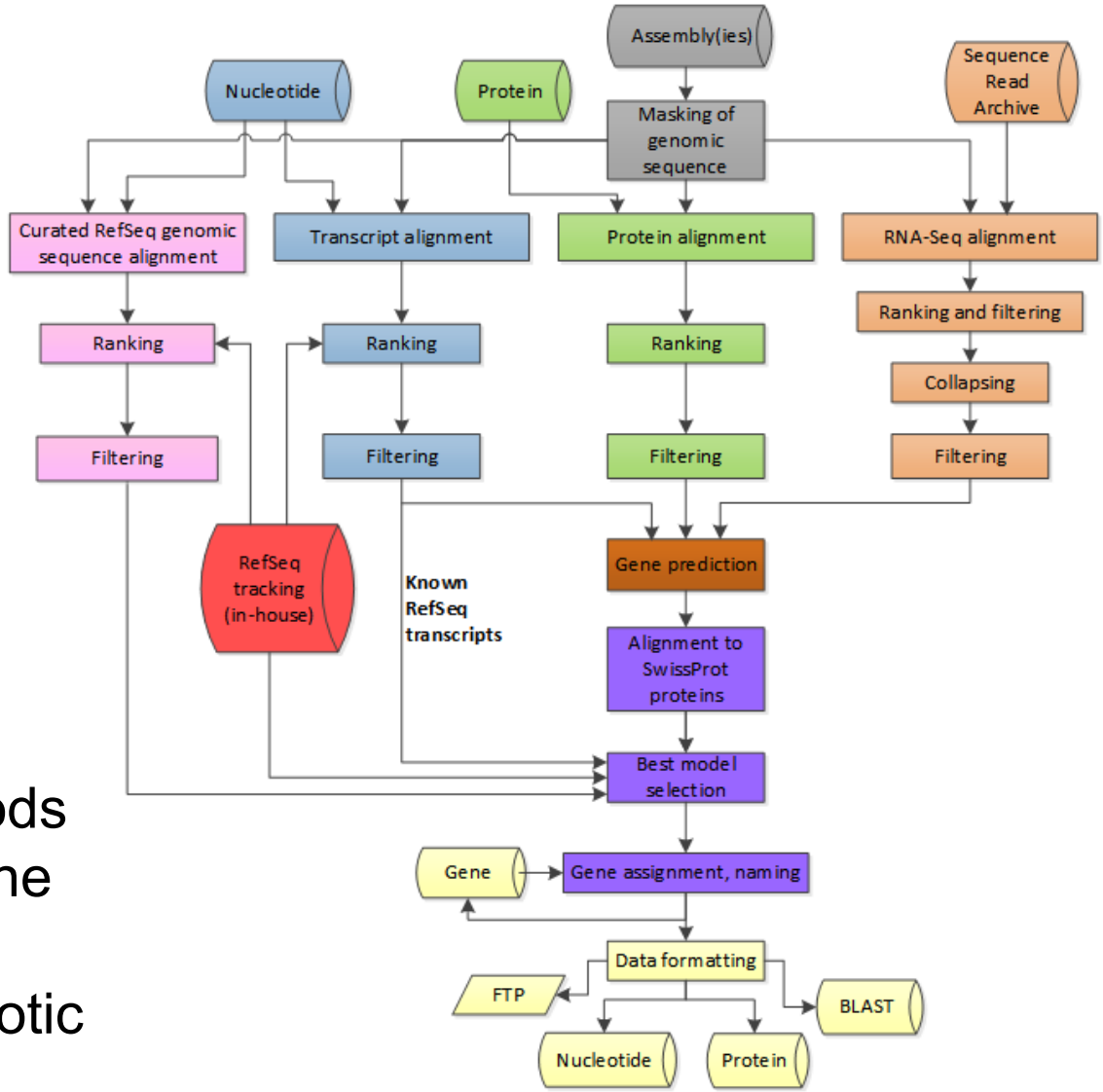
***Genome annotation*: Finding relevant regions in sequences**

# Course Outline



# Genome Annotation: Algorithms

- Alignment-based
- Sequence-based
- Content-based
- Probabilistic




- Be able to combine content-based and probabilistic methods of gene discovery to identify the most probable locations of introns and exons in a eukaryotic DNA sequence

# Genome Annotation

- Locate genes for proteins in sequences.

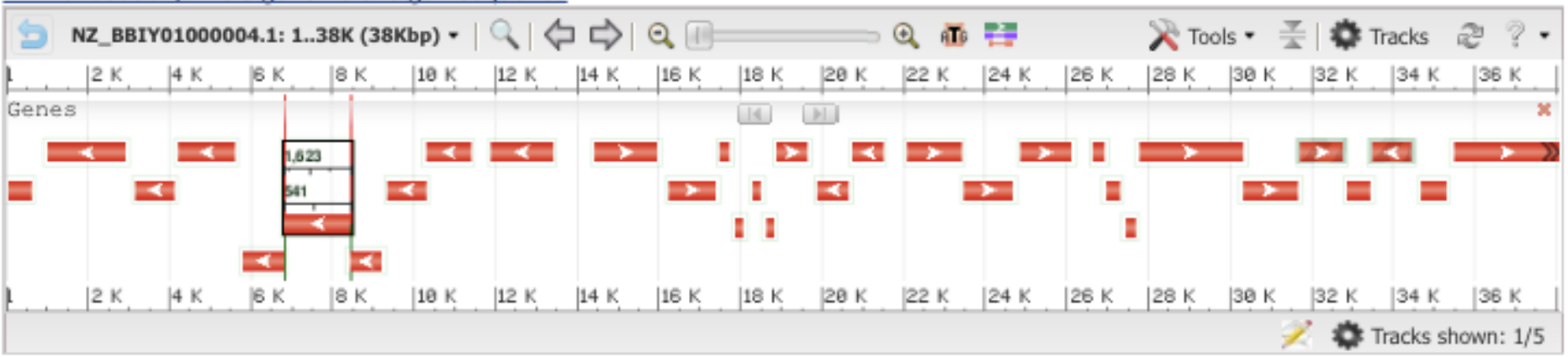
**Genome Assembly Annotation**

| Type | Name | RefSeq | INSDC | Size (Mb) | GC% | Protein | tRNA | Other RNA | Gene | Pseudogene |
|---|---|---|---|---|---|---|---|---|---|---|
| | master WGS | NZ_BBIY00000000.1 | BBIY00000000.1 | 0.74 | 27.6 | 901 | 27 | - | 928 | - |

**Genome Region**

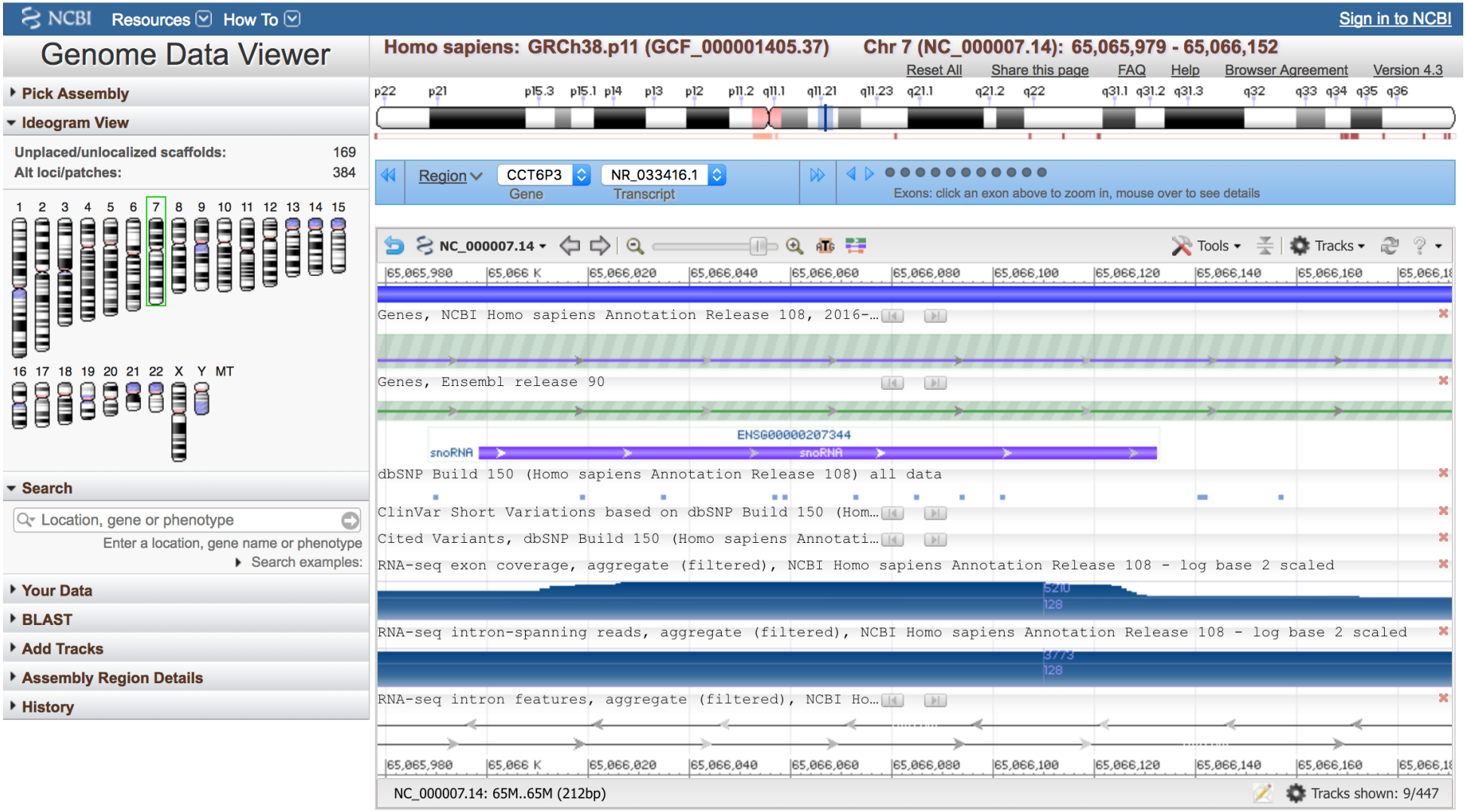
'Chrysanthemum coronarium' phytoplasma strain OY-V
BBIY01000004, whole genome shotgun sequence

**Go to nucleotide:** Graphics FASTA GenBank




https://www.ncbi.nlm.nih.gov/genome/browse/

# Genome Annotation

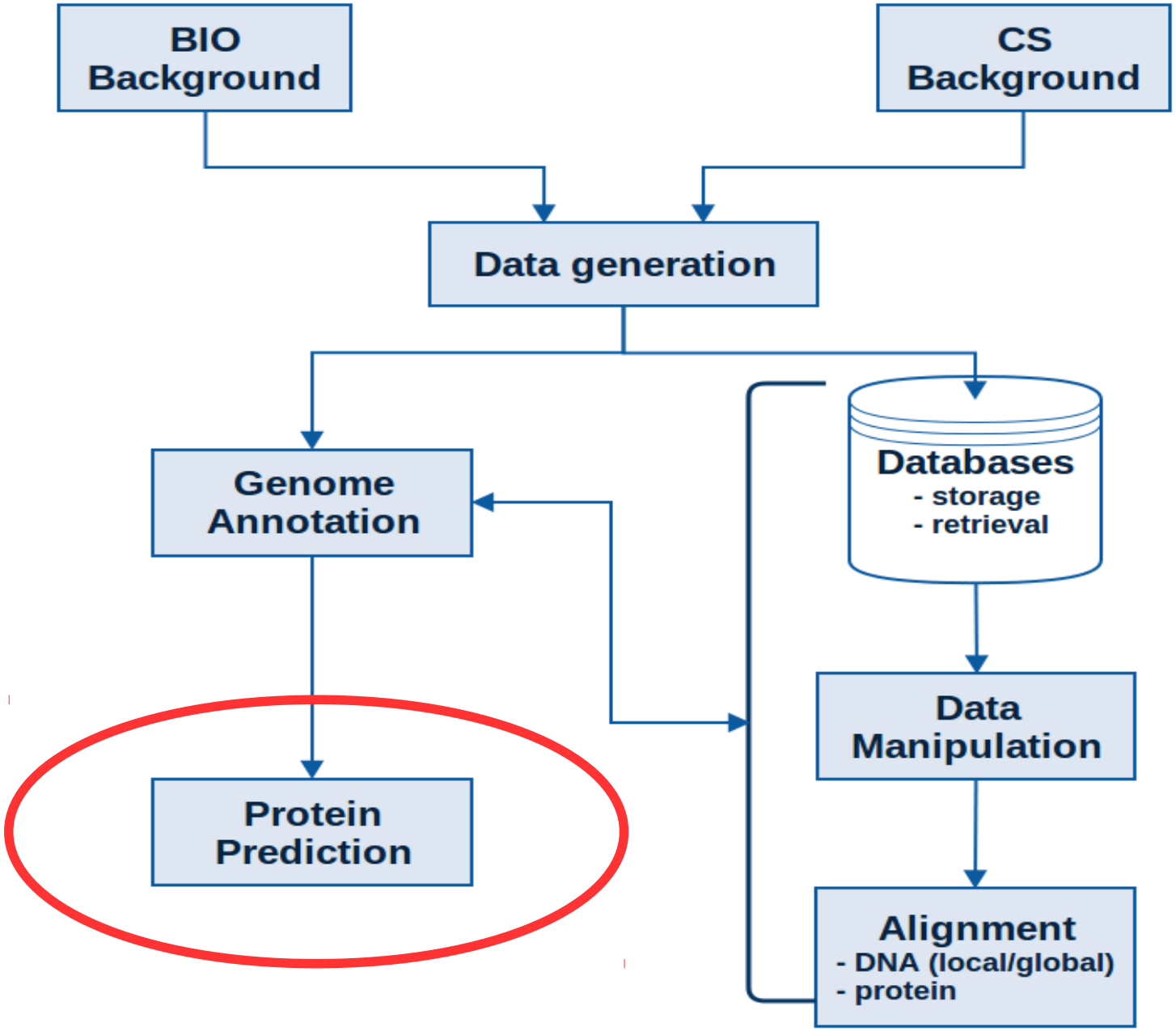



https://www.ncbi.nlm.nih.gov/genome/gdv/browser/?context=Nucleotide&acc=NR_145729.1

# We Talked About...

***Protein Prediction*: Determining what protein exist in a sequence and how they might *behave*.**

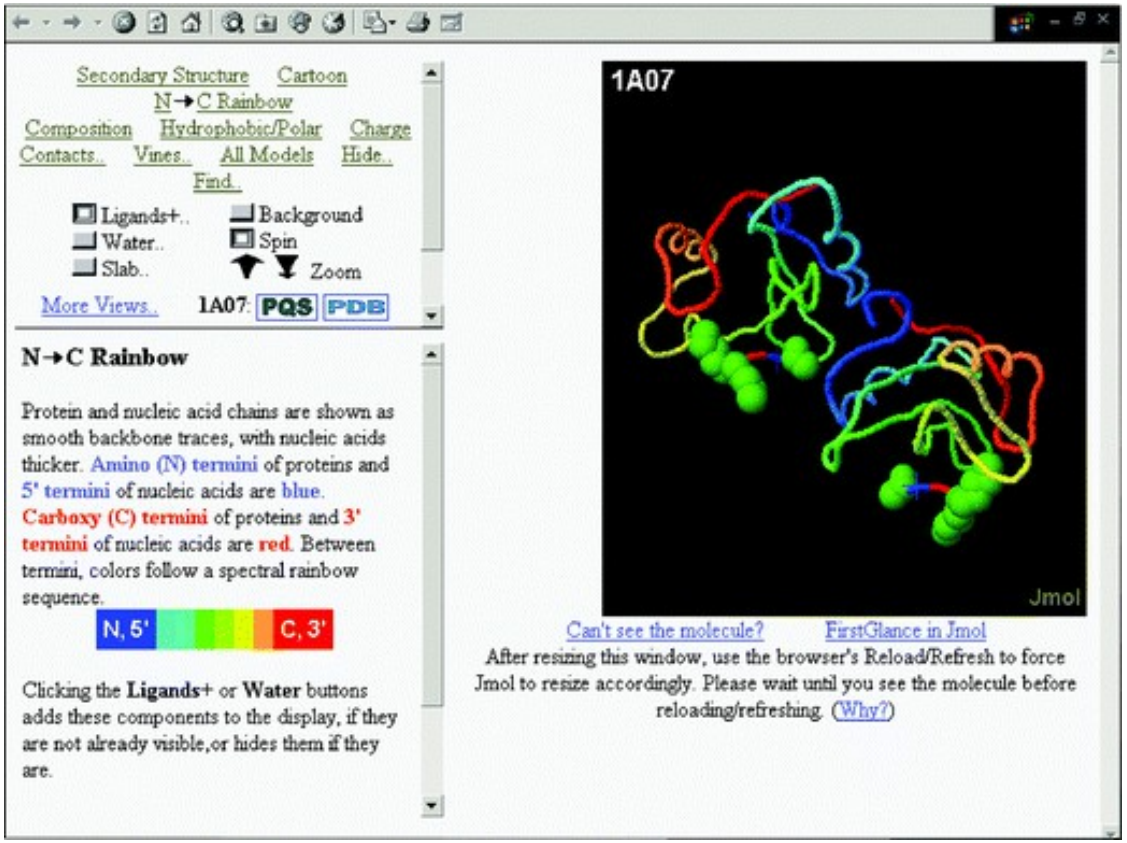
# Course Outline



# Protein Prediction

- Know how to use available tools to examine the experimentally determined structures of proteins and visualize structural an functional features

- Appreciate the value and limitations of predicting 3-D structure from sequence alone



# Protein Folding - Applications

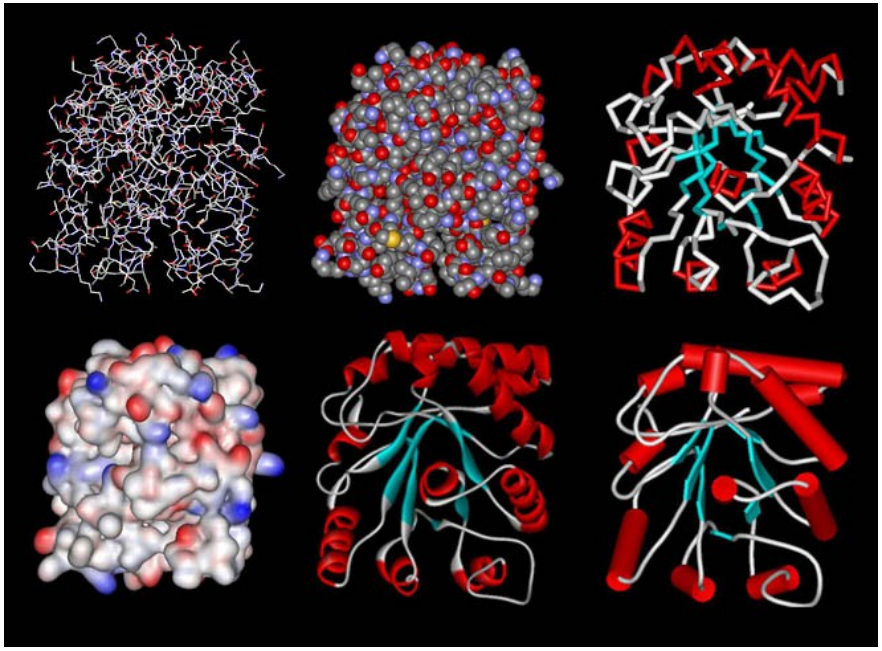
- **Protein must fold correctly to function**

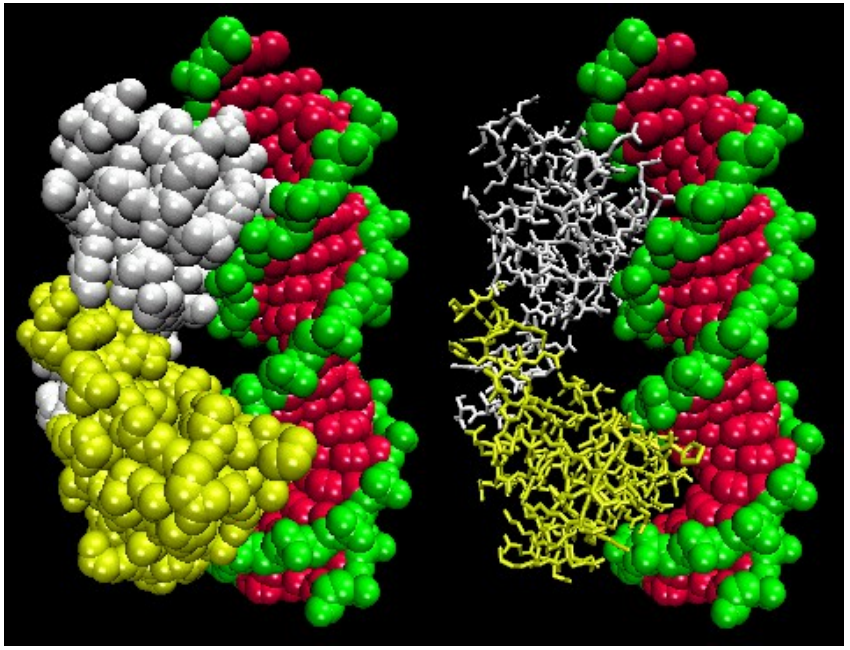
- Misfolded proteins
  - Accumulation – Huntington's and Parkinson's disease
  - Tagged for degradation – emphysema, cystic fibrosis
    - Pharmaceutical chaperones – fold mutated proteins to render them functional

- Antiviral drug development
  - Antibiotics vs antivirals
    - Bacteria – cells
    - Viruses – invade host's cells

# Protein DataBase (PDB)

- Database for 3-D structural data of large biological molecules
- https://www.rcsb.org/
- Data is viewable using jmol.



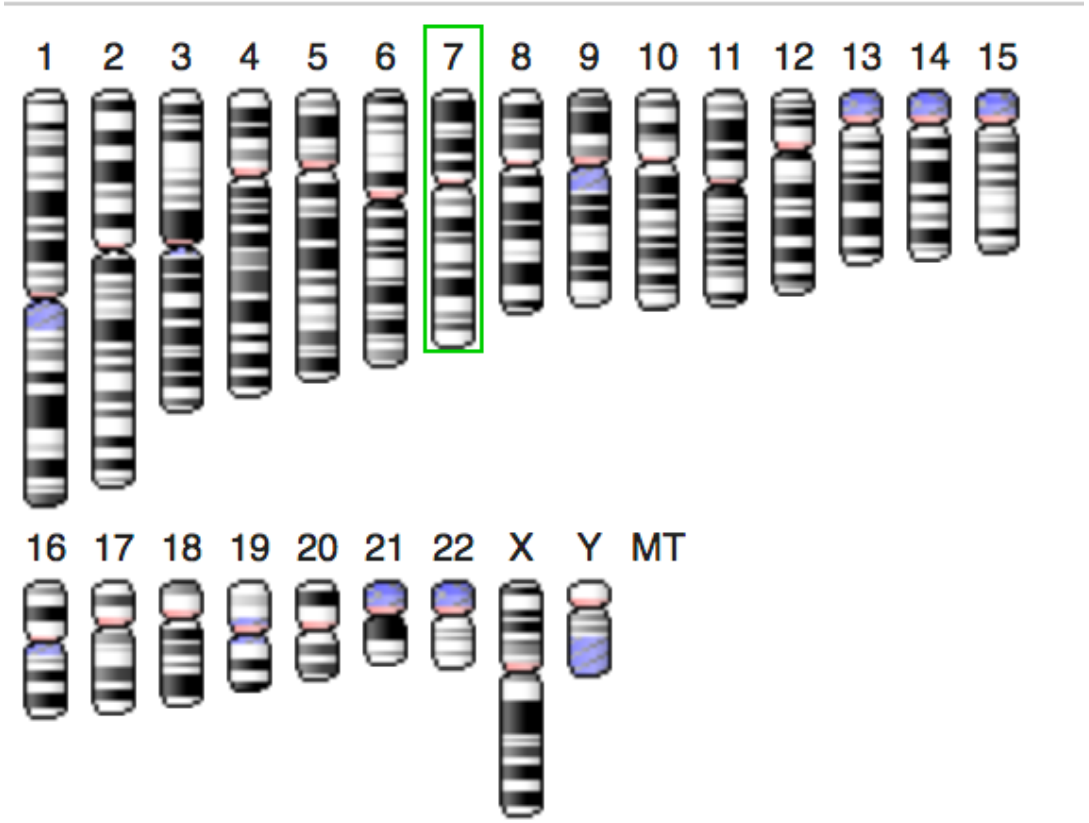


# In Closing ...

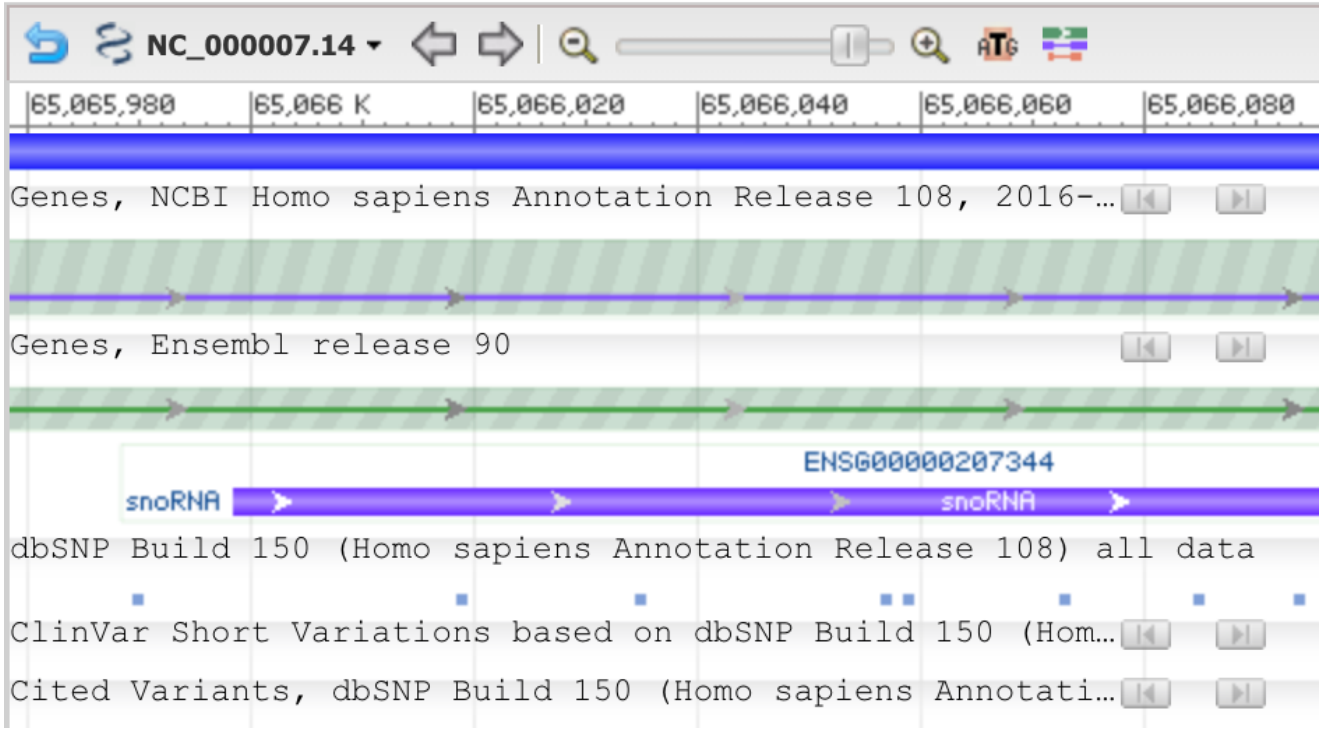
**Bioinformatics is diverse and exciting!**

# Bioinformatics Accomplishments

- ✓ A “big-picture” view of bioinformatics.
- ✓ An understanding of the objectives and limitations of bioinformatics.
- ✓ An understanding of the biological foundations of bioinformatics (genes and genomes, gene expression, etc.).



# Bioinformatics Accomplishments



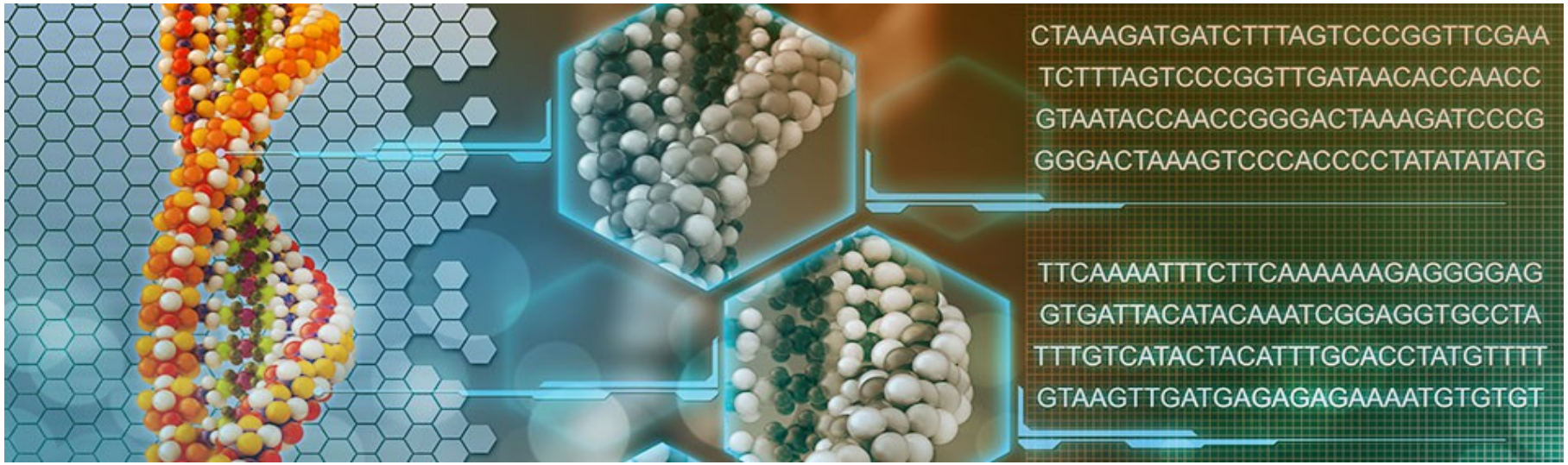

- An understanding of the computational foundations of bioinformatics (programming, databases, etc.).
- An understanding of how genetic information is obtained and processed.
- The ability to use basic bioinformatics software tools to study genetic information.

# What's Next?

**Bioinformatics could provide you with a satisfying career and plenty of room to advance**

# The Value of the Bioinformatics Skills

There is a great need for Bioinformaticians!



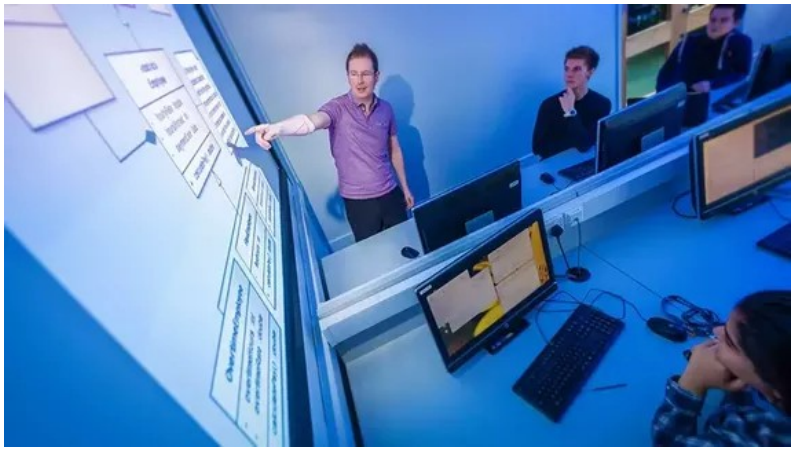
# Skills in Careers

- Biologists:
  - Computational skills
  - Mathematical /statistical
  - Programming for Automation

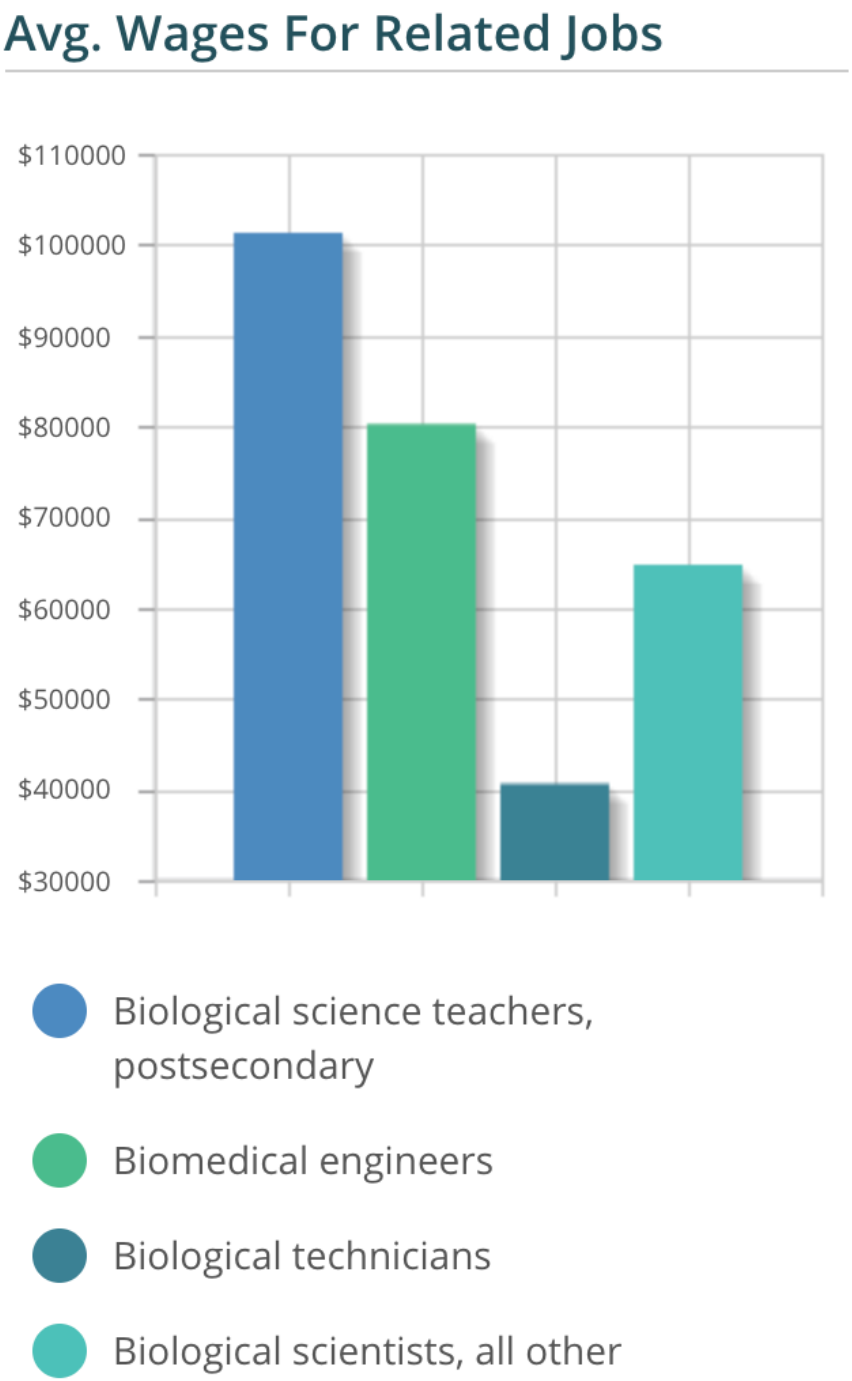

- Computer scientists
  - BioMedical skills
  - Understanding of biological systems and mechanisms
  - Early detection of disease by data
  - Modeling of therapeutic remedies
  - Others

# Skills in Careers

- Software (bioinformatics) engineer
- Research scientist in biotechnology
- Data scientist
- Project manager (pharmaceuticals, medical, etc)
- Computational immunologist
- Medical doctor (in clinical and research applications)



# High Paying Careers



# High Paying Careers

## Bioinformatics Research Scientist Salaries

36,327 Salaries  Updated Aug 10, 2015

All Industries | All Company Sizes | All Years of Experience

Average Base Pay

**$90,214** /yr

Not enough reports to show salary distribution

| $59K | $90K | $127K |
|---|---|---|
| Low | Average | High |

Additional Cash Compensation

| | |
|---|---|
| Average | $xx,xxx |
| Range | $xx,xxx |

**How much does a Bioinformatics Research Scientist make?**

The national average salary for a Bioinformatics Research Scientist is $90,214 in United States. Filter by... More

https://www.glassdoor.com/Salaries/bioinformatics-research-scientist-salary-SRCH_KO0,33.htm

# High Paying Careers

## Bioinformatics Scientist Salaries

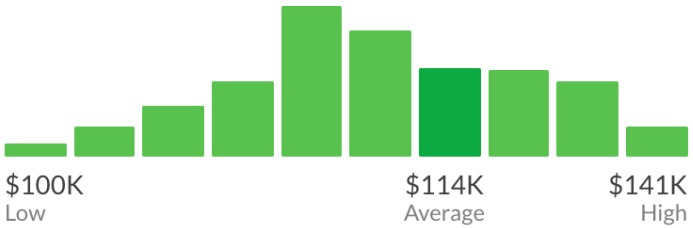
About This Data

287 Salaries  Updated Nov 27, 2017

All Industries | All Company Sizes | All Years of Experience

Average Base Pay

**$113,545** /yr




Additional Cash Compensation

| | |
|---|---|
| Average | $9,721 |
| Range | $1,511 - $17,411 |

**How much does a Bioinformatics Scientist make?**
The national average salary for a Bioinformatics Scientist is $113,545 in United States. Filter by location... More

Salaries for Related Job Titles

| | |
|---|---|
| Bioinformatics Analyst | $76K |
| Bioinformatics Research Scient... | $90K |
| Senior Scientist, Computation... | $129K |
| Bioinformatics Engineer | $101K |

https://www.glassdoor.com/Salaries/bioinformatics-scientist-salary-SRCH_KO0,24.htm

# Bioinformatician Jobs

- Research scientist
- Bioinformatician
- Bioinformatics programmer
- Software Developer
- Analyst
- Statistician
- Physician
- Project manager
- Database developer and administrator
- Technical assistant and technical sales representative
- or any jobs where biologists are currently hired

(some of these may require graduate education)

# Bioinformatics Jobs and Internships

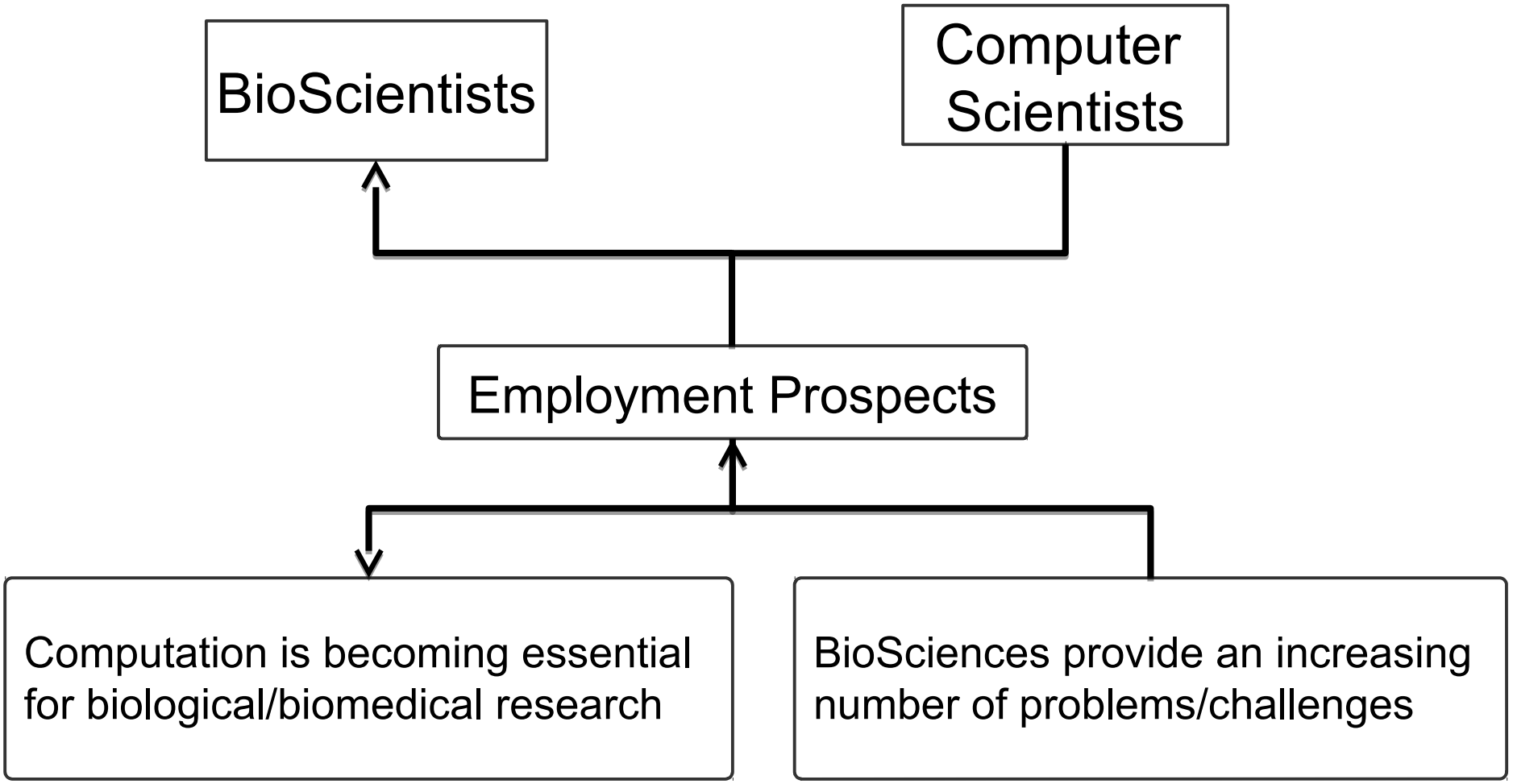
**Resources**

http://www.iscb.org/iscb-careers-job-database
http://www.bioinformatics.org/jobs/
http://www.bioplanet.com/
http://www.bio-itworld.com/BioIT/JobOpenings.aspx
http://www.biospace.com/

www.glassdoor.com
http://www.jobs-salary.com/jobs.php?

Campus Resources

# The Value of the Bioinformatics Skills



# In Bioinformatics, there is ...

SO MUCH

MORE