Bioinformatics

CS300

The Great Review

Fall 2022

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

- **Students successfully completing this class will have developed:**
  - A “big-picture” view of bioinformatics.
  - An ethical foundation
  - An understanding of the objectives and limitations of bioinformatics.
  - An understanding of the biological foundations (genes and genomes, gene expression, etc.).
  - An understanding of the computational foundations (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.

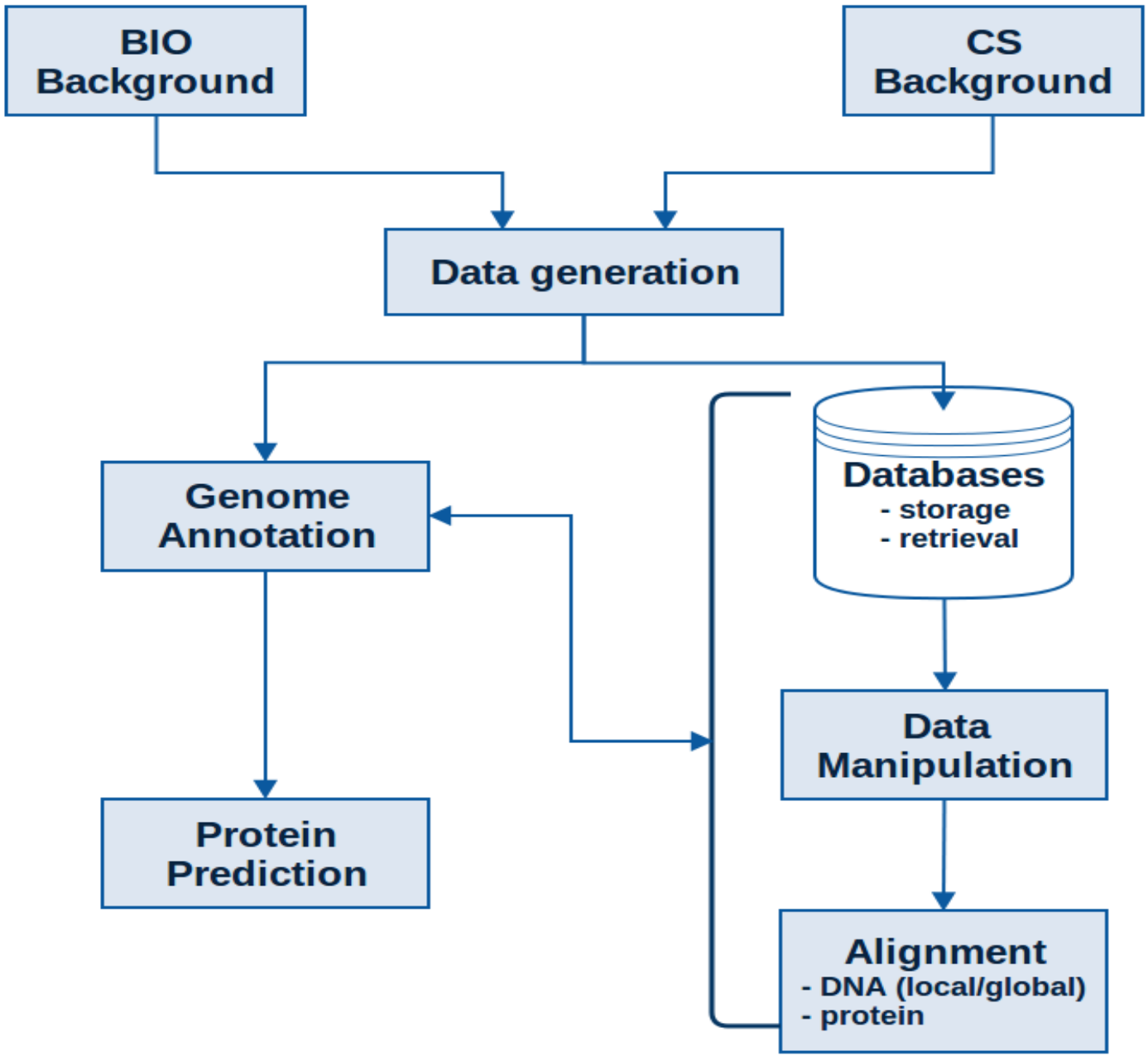
An **Ethically** Conscious Course

- We had labs and assignments in which ethical themes
- We met a **speaker** who came to talk about
  - The ethical use of data in the health sciences
  - The responsible use of bio-tech and automation

How Did We Meet These Objectives?

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

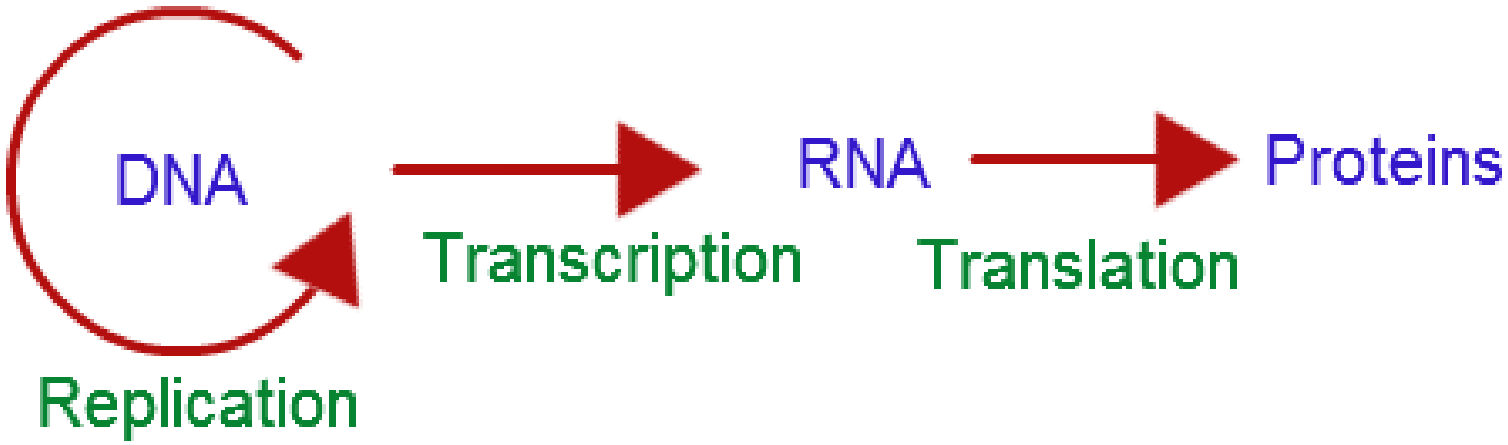
Course Outline



We Started With …

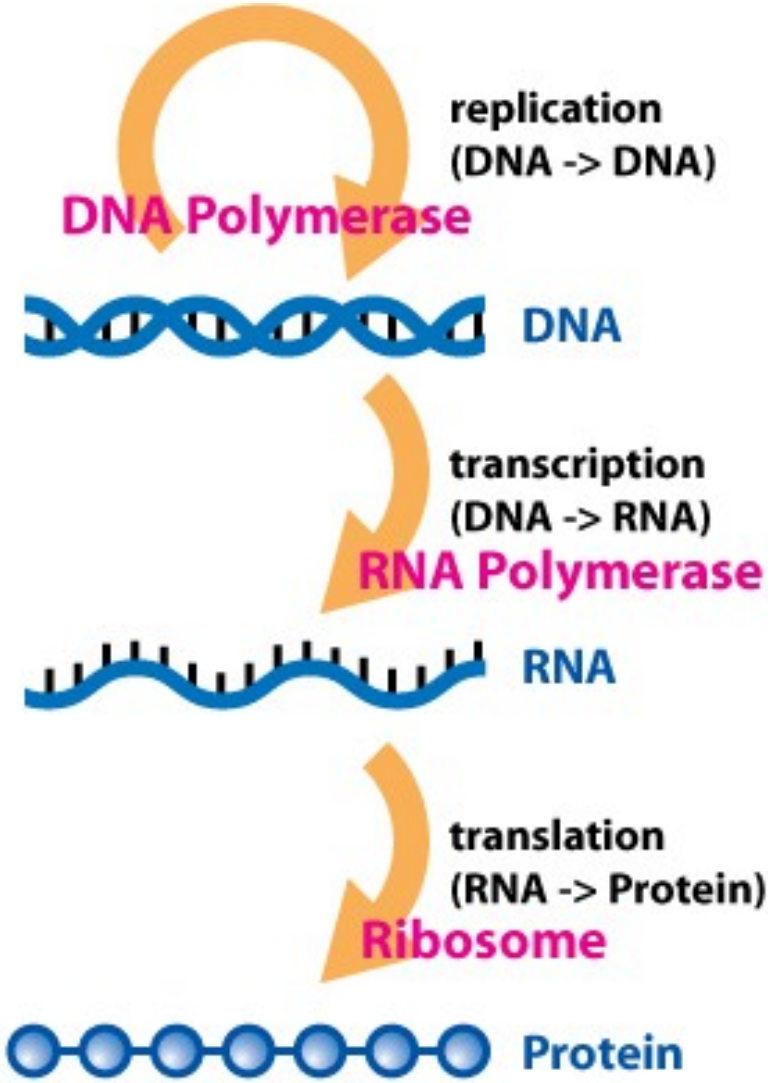
The Central Dogma Of Biology

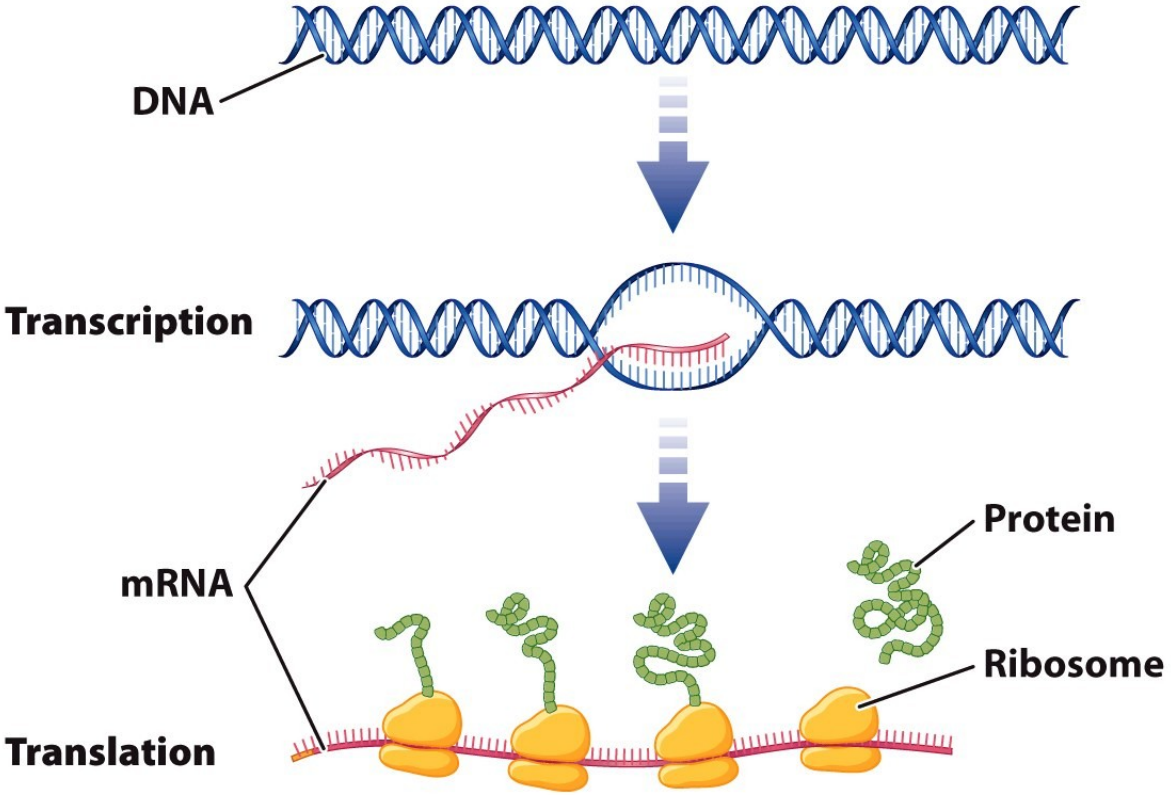
The Central Dogma Of Biology



“the coded genetic information hard-wired into DNA is transcribed into individual transportable cassettes, composed of messenger RNA (mRNA); each mRNA cassette contains the program for synthesis of a particular protein (or small number of proteins)“ - NCBI

The Central Dogma of Molecular Biology

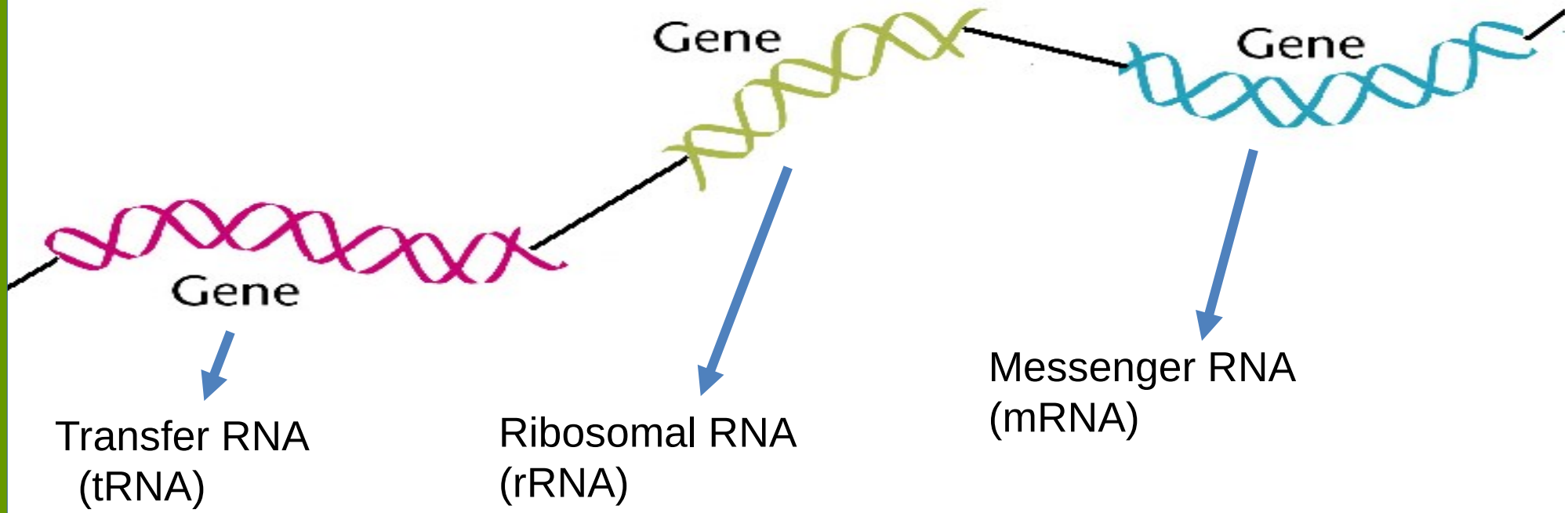




Proteins provide structure and carry out many essential activities in a cell.

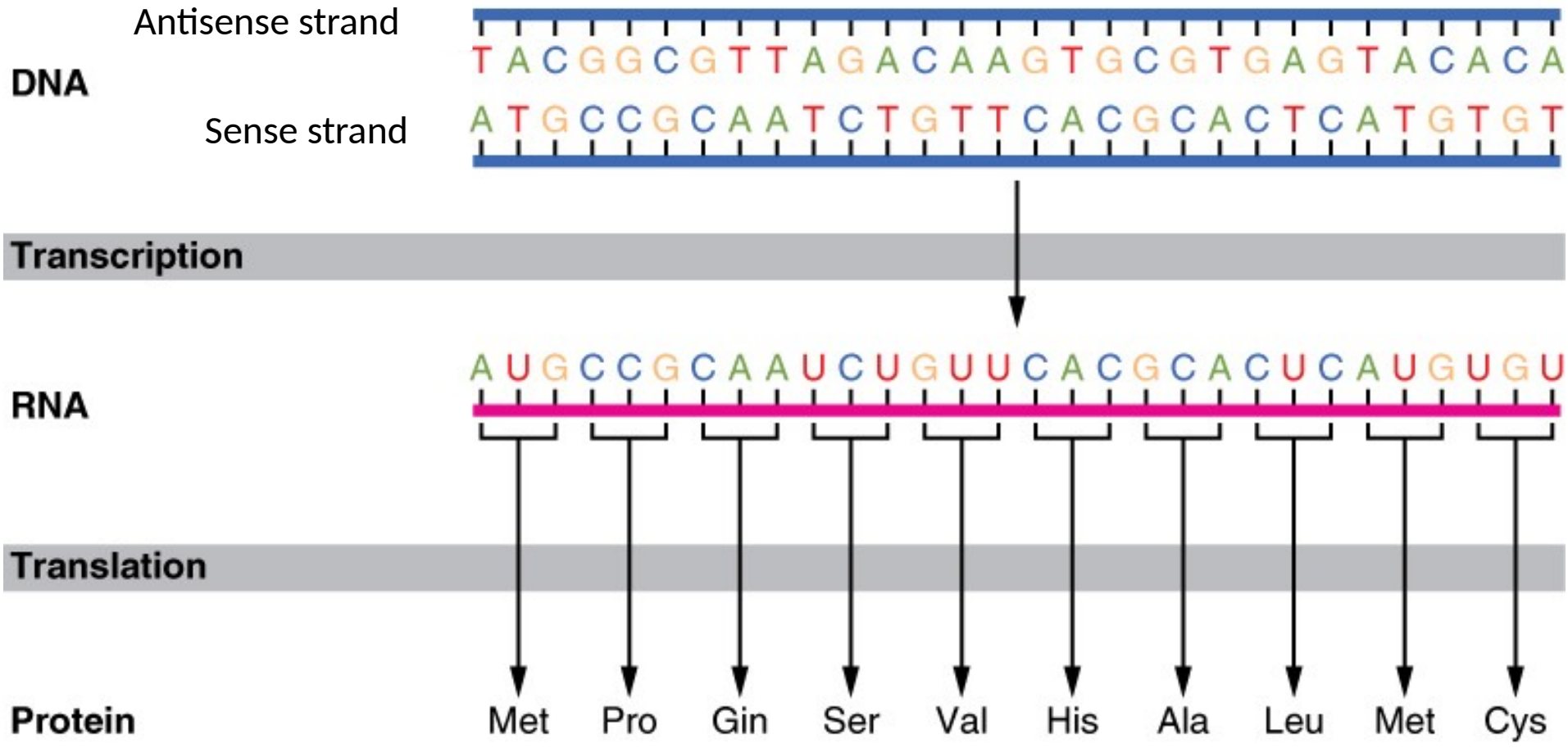
Transcription

- 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)



Transcription

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



Transcription

- 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

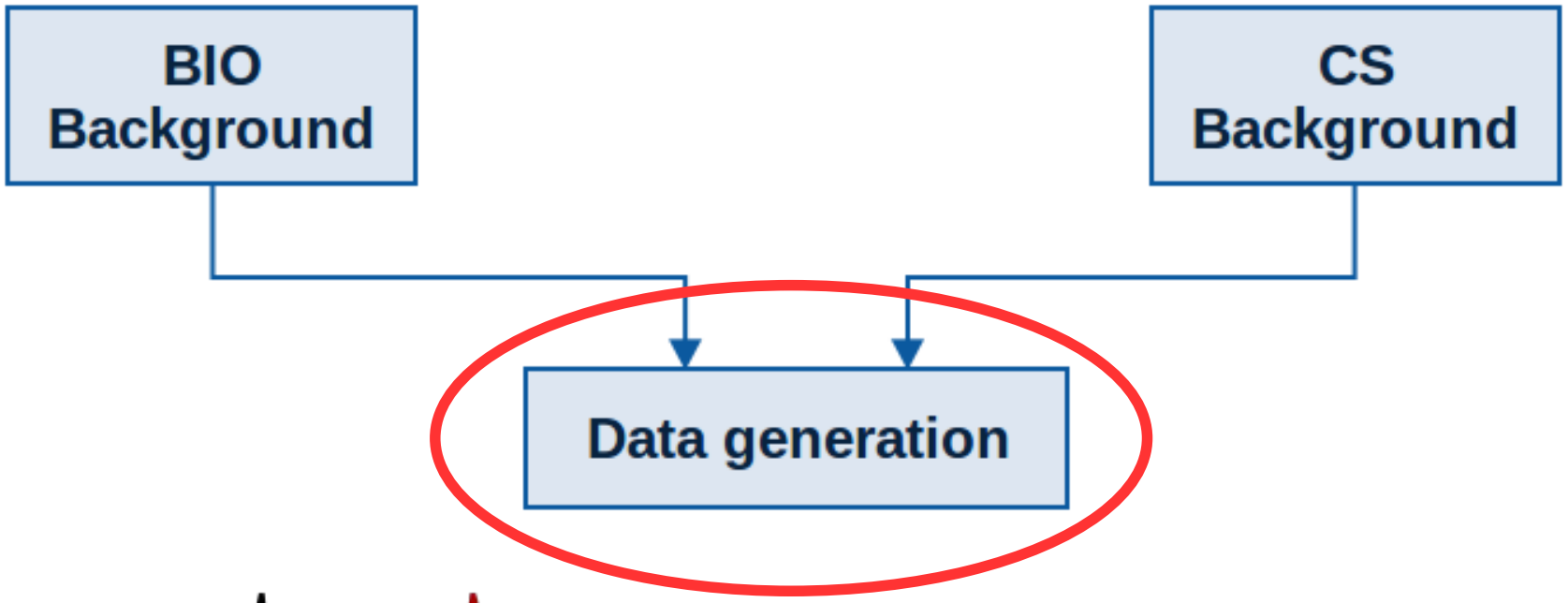
1st base	2nd base T		2nd base C		2nd base A		2nd base G		3rd base
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine	TCA		TAA[B]	Stop (*Ochre*)	TGA[B]	Stop (*Opal*)	A
	TTG		TCG		TAG[B]	Stop (*Amber*)	TGG	(Trp/W) Tryptophan	G
C	CTT		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

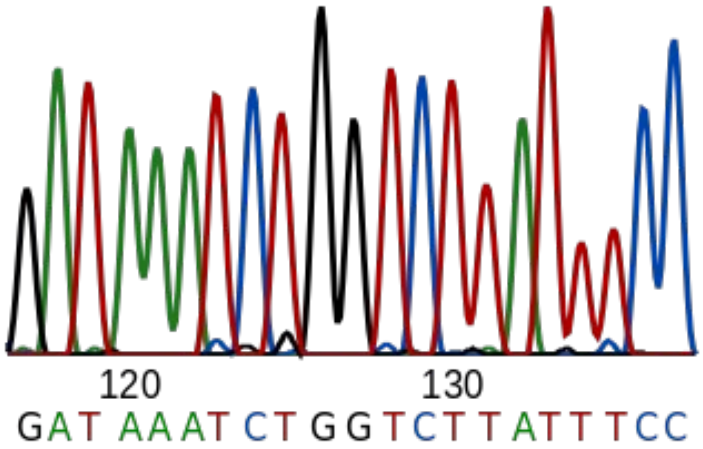
We Talked About…

ALLEGHENY COLLEGE

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

From The Course Outline

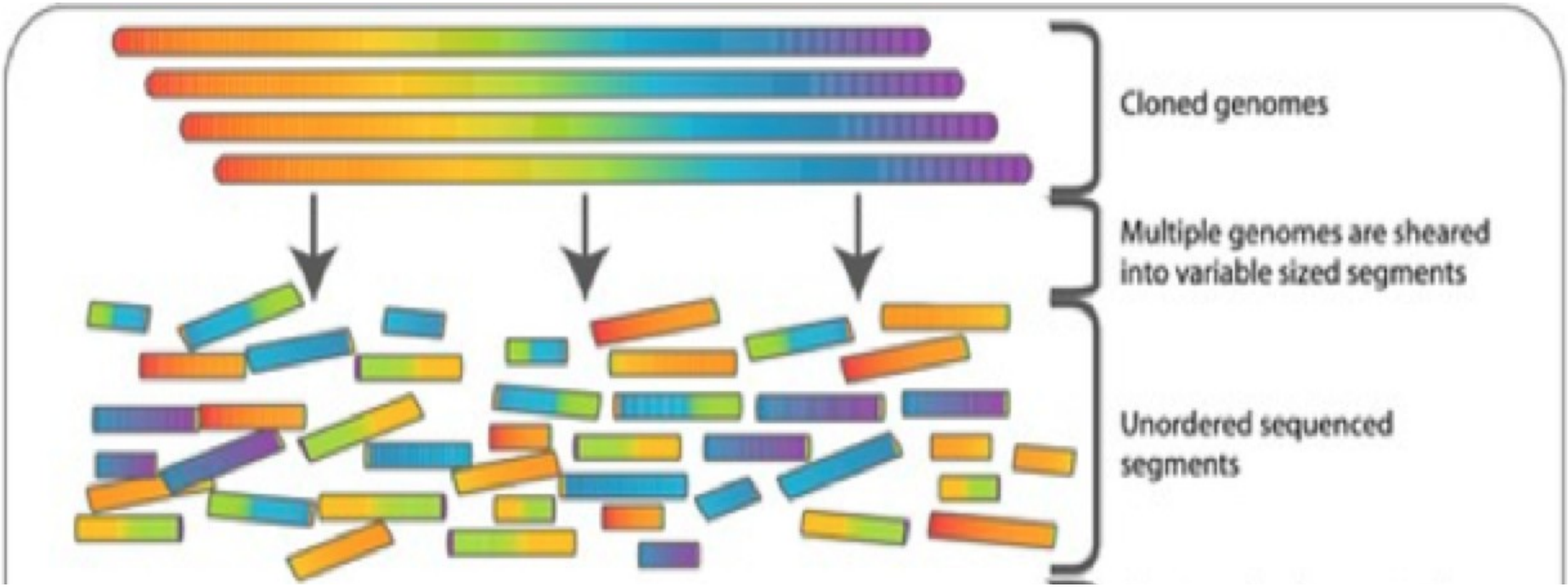


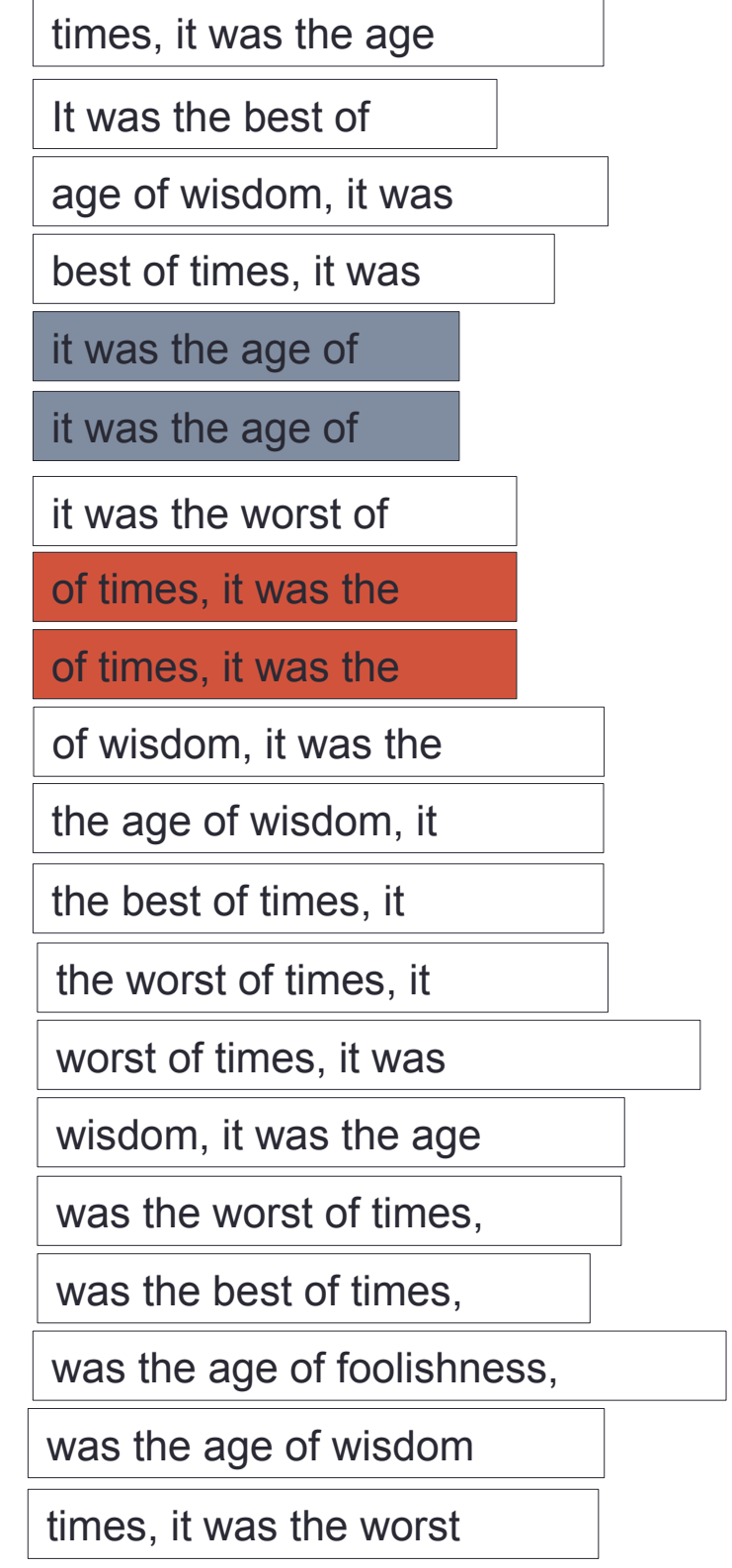


We get DNA data by combining biology and computer science

Genome Sequencing

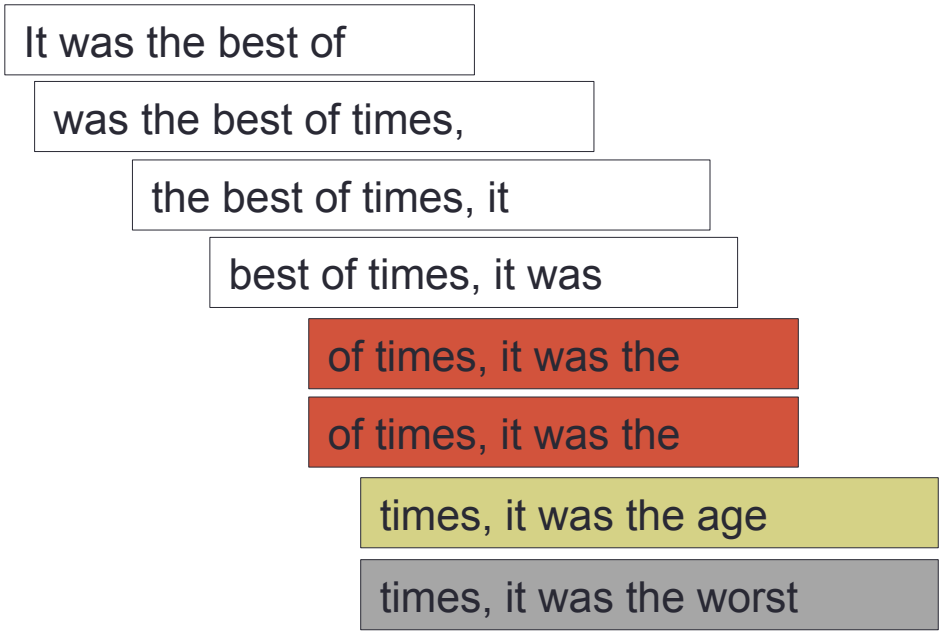
- The technology works by “exploding” DNA into smaller, manageable pieces
- Then it recombines pieces (*Reads*) into bigger pieces (C*ontigs*)
- And then it combines contigs bigger chunks like a jigsaw puzzle





Repeats pile up – actual placement of each individual fragment unknown

Repeats can cause ambiguity and prevent proper assembly



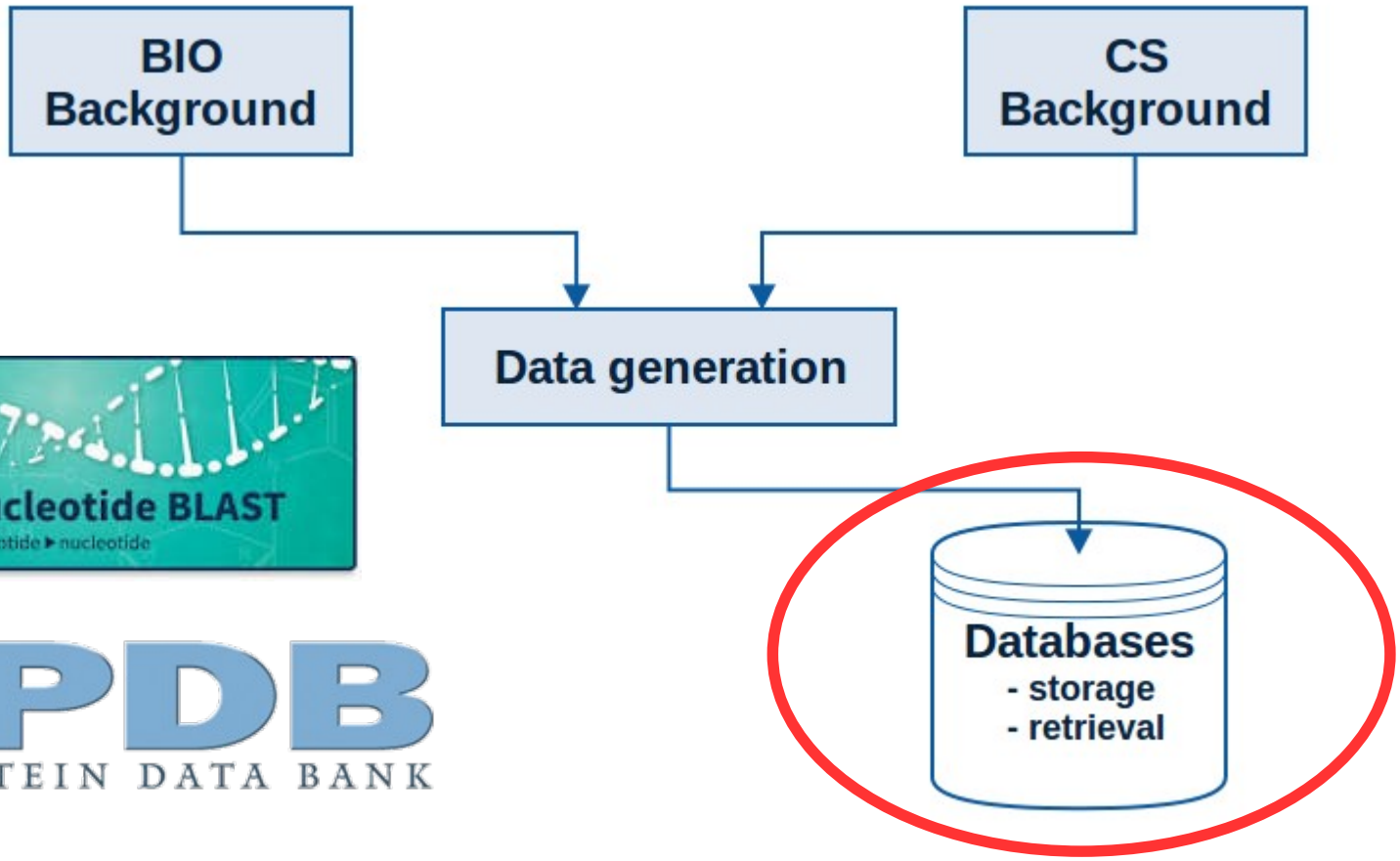
It was the best of times, it was the [age/worst]

Assembly Parameter:
100% identify across 4 words

We Talked About…

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

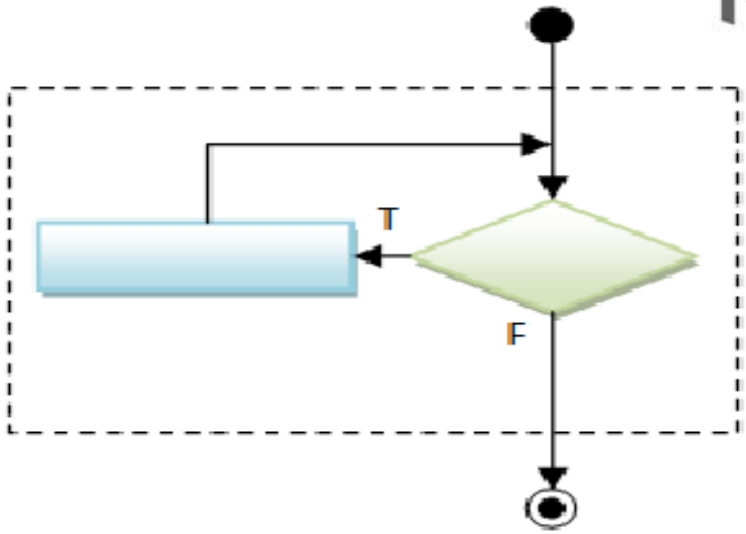
From The Course Outline



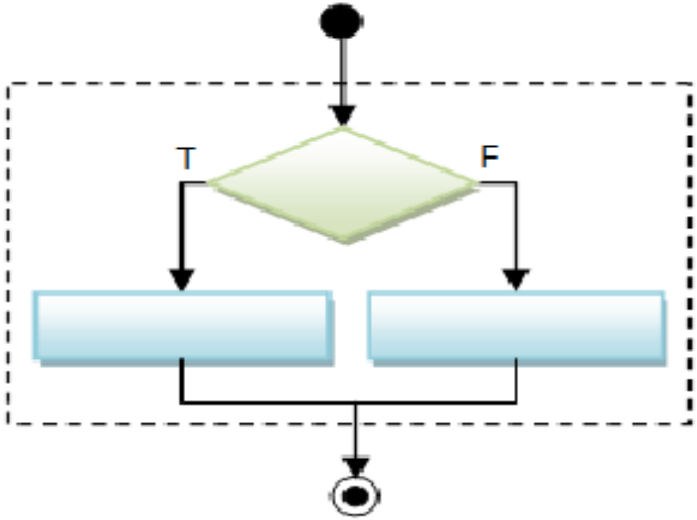
Databases (and Tools)

ALLEGHENY COLLEGE

GitHub



Loop (Iteration)



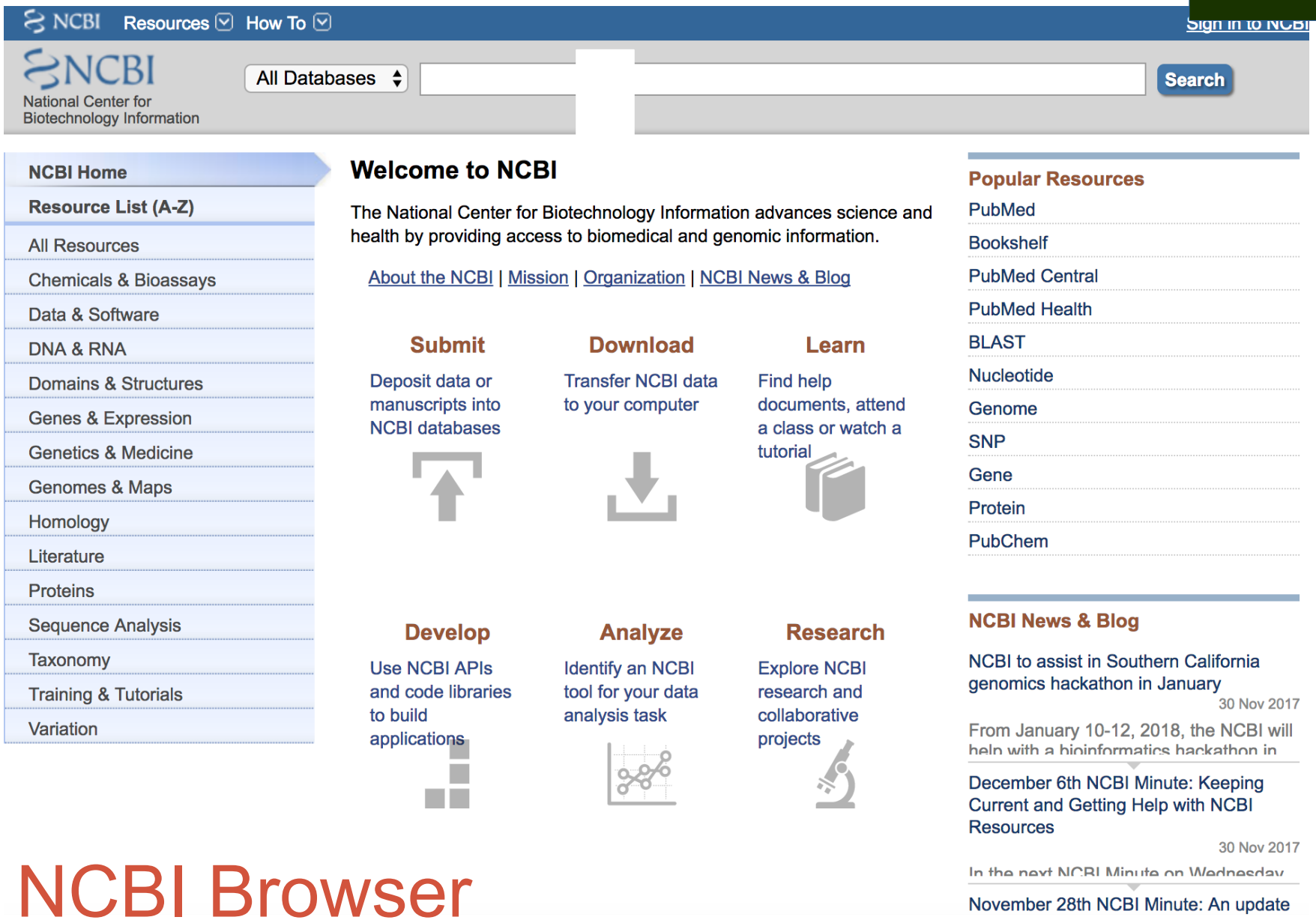
Conditional (Decision)

Biological Data and Databases

ALLEGHENY COLLEGE

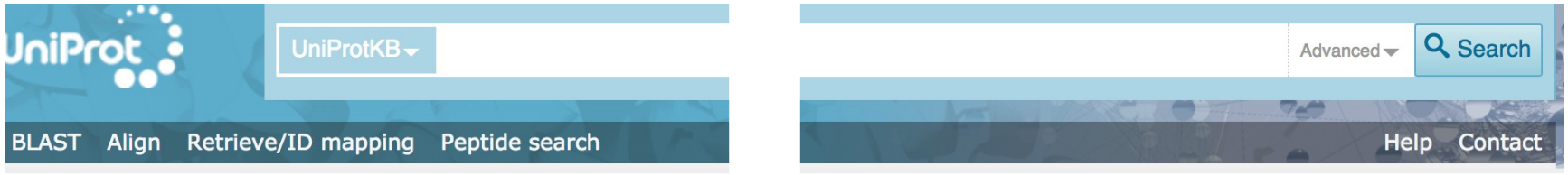
- To learn how to use a Web-based genomic databases and tools.
- To understand the types of information stored in genomic databases.
- To learn how to use different interfaces to find and retrieve genomic information.
- Write Python program to find patterns (start and stop codons) in DNA sequences

DB Interfaces: NCBI



NCBI Browser

DB Interfaces: UniProt



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

DB Interfaces: UniProt

- Domains give the protein special qualities:
  - Domain Names: *Alpha1, Alpha2, Alpha3, Ig-like C1-type*

Family & Domains[i]

Domains and Repeats

Feature key	Position(s)	Description	Actions	Graphical view	Length
Domain[i]	209 – 297	Ig-like C1-type	Add BLAST		89

Region

Feature key	Position(s)	Description	Actions	Graphical view	Length
Region[i]	25 – 114	Alpha-1	Add BLAST		90
Region[i]	115 – 206	Alpha-2	Add BLAST		92
Region[i]	207 – 298	Alpha-3	Add BLAST		92
Region[i]	299 – 309	Connecting peptide	Add BLAST		11

UniProt ID: P01899

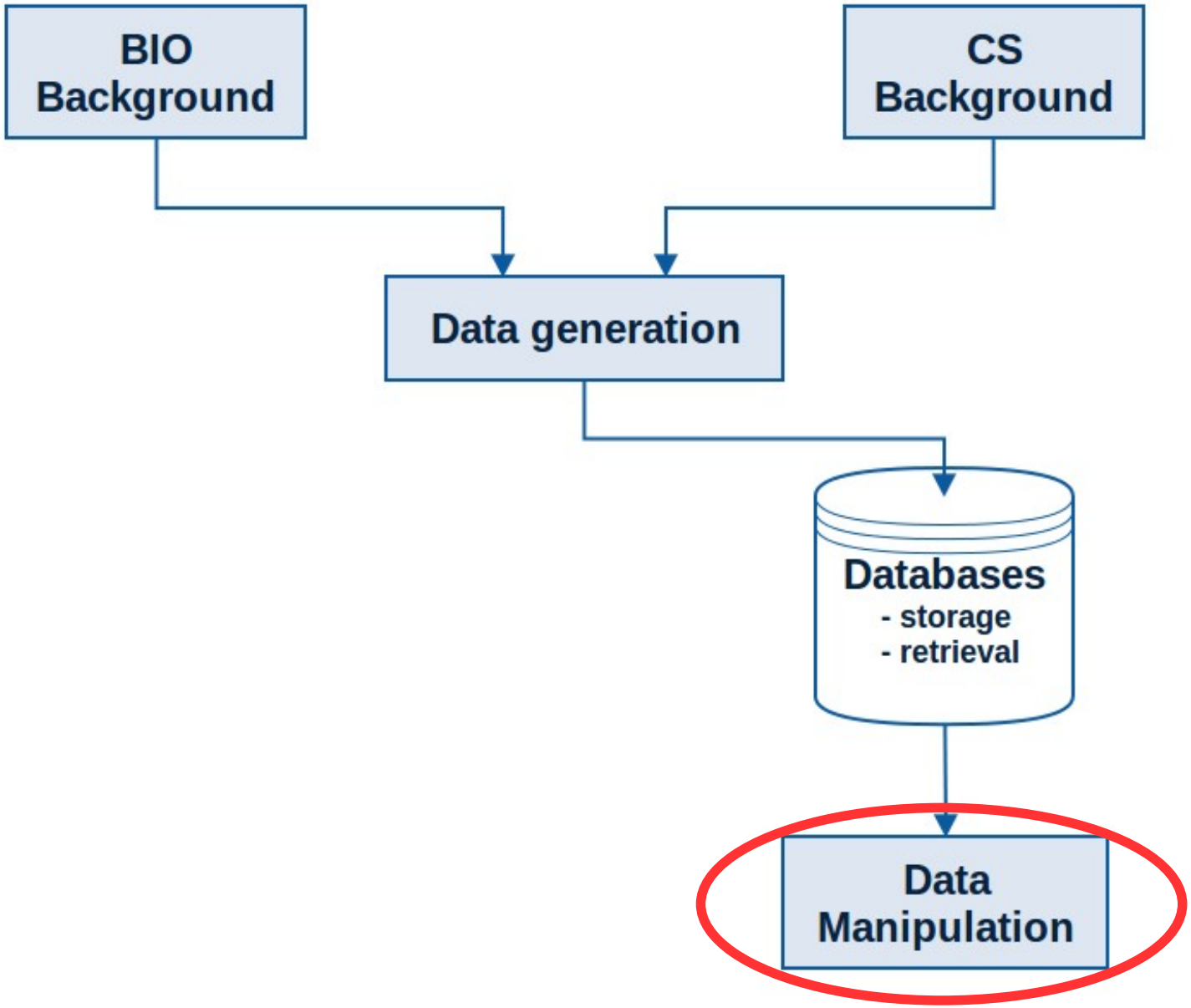
A Protein Knowledge Base

http://www.uniprot.org/uniprot/P01899#family_and_domains

We Talked About…

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

From The Course Outline



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	**utative KilA-N omain-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
```

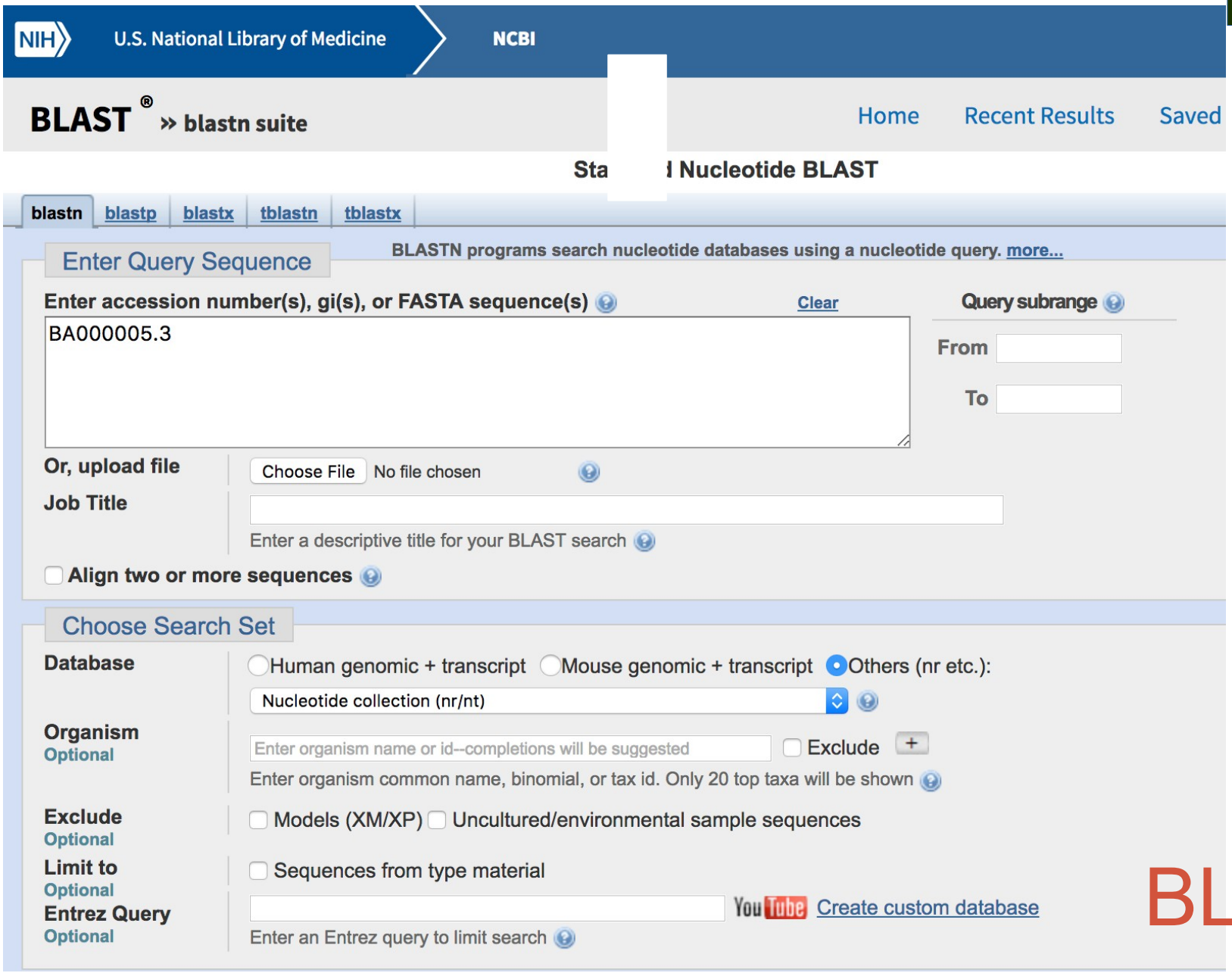
Molecule processing

Feature key	Position(s)	Description Actions	Graphical vi
Transit peptide i	1 – 77	Mitochondrion 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 1 Publication	
Modified residue i	402	Phosphoserine; by autocatalysis 1 Publication	

Using Tools With Databases

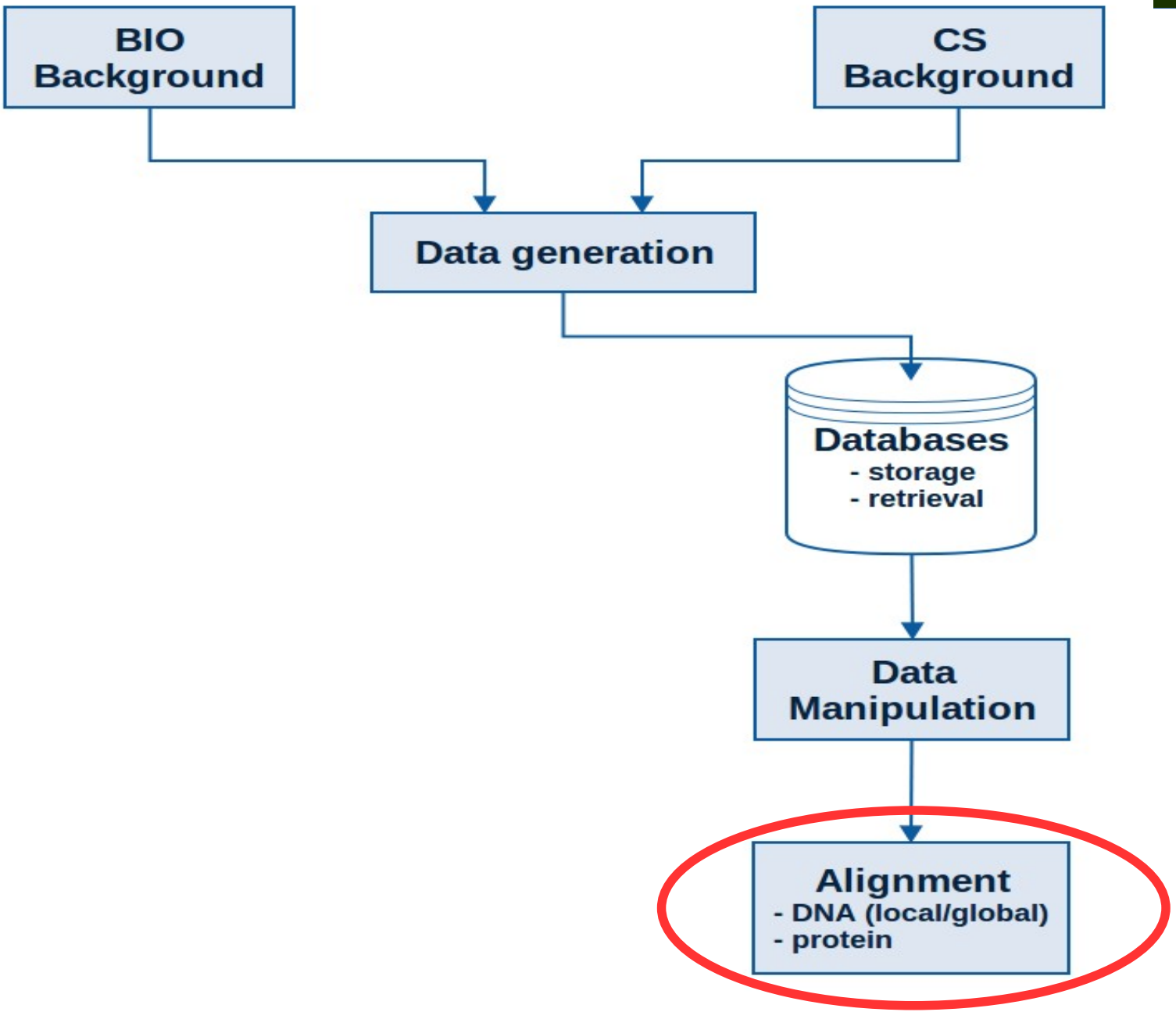


BLAST

We Talked About…

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

From The Course Outline



Comparing Things!

- DNA sequences
- Genes
- Proteins
- Organisms

- **How to compare these things?**
- **What do we learn by comparison?**

Sequence Alignment

ALLEGHENY COLLEGE

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.

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

Sequence Alignment: Needleman-Wunsch

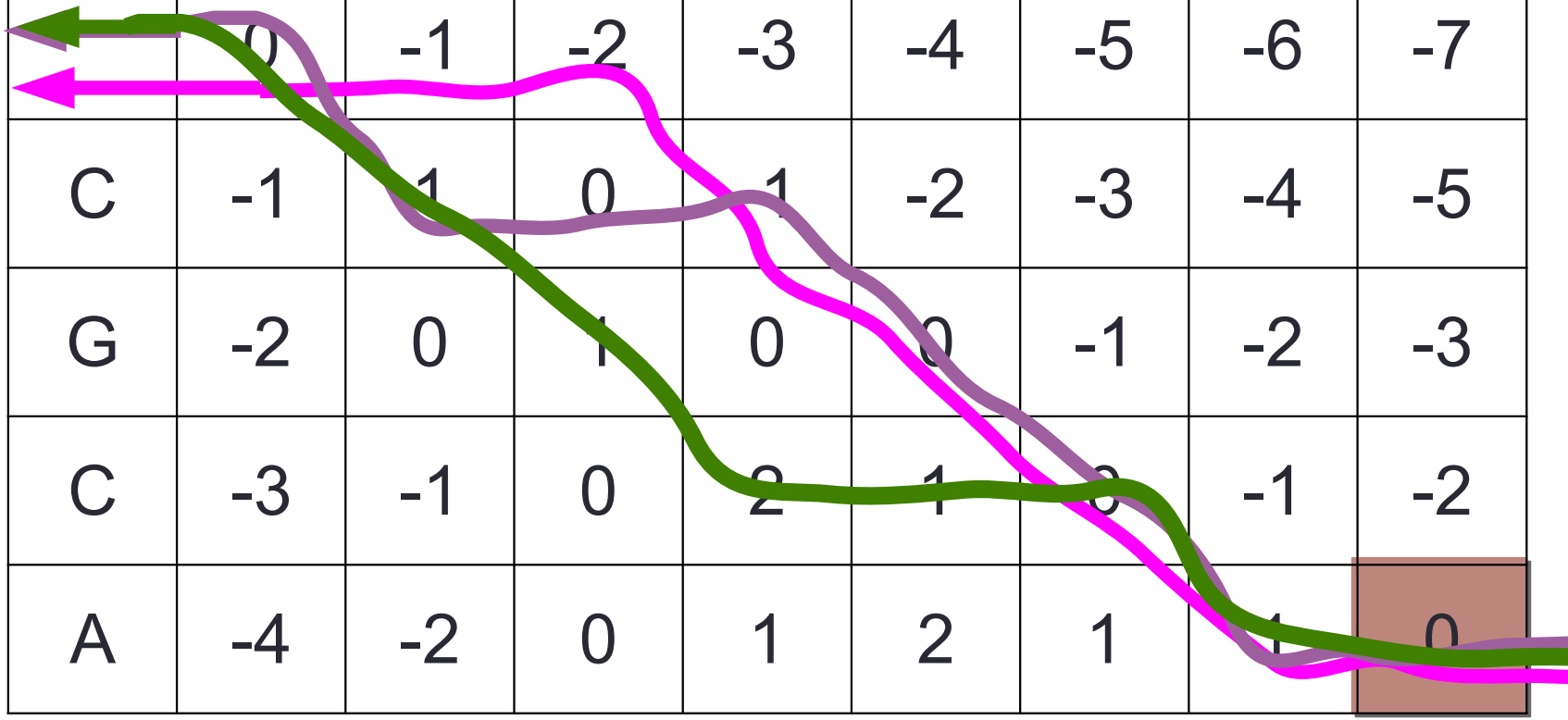
Let:

Match = +1

Mismatch = 0

Gap = -1

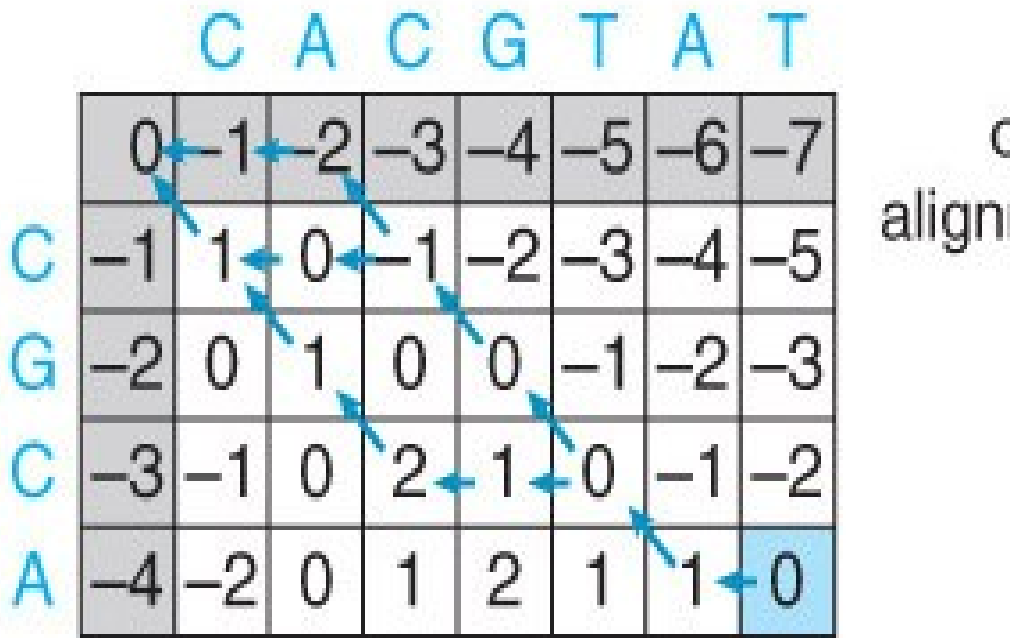
		C	A	C	G	T	A	T
	0	-1	-2	-3	-4	-5	-6	-7
C	-1	1	0	1	-2	-3	-4	-5
G	-2	0	1	0	0	-1	-2	-3
C	-3	-1	0	2	1	0	-1	-2
A	-4	-2	0	1	2	1	1	0



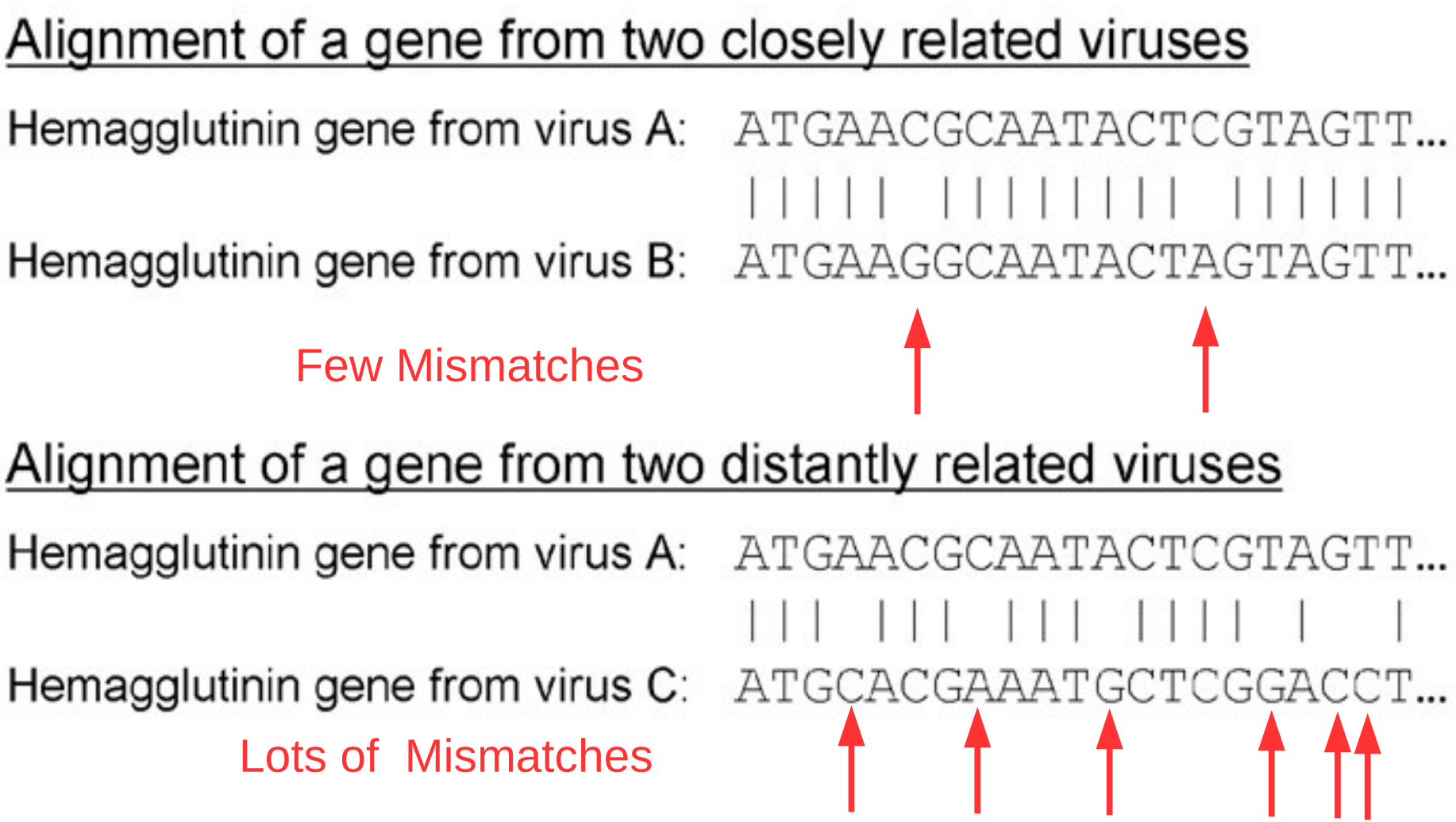
```
CACGTAT
--CGCA-

CACGTAT
C--GCA-

CACGTAT
CGC--A-
```

Global Pairwise Alignment

- Steps to begin
  - Initialization of the matrix
  - Calculation of the scores given for a character by character comparison.
  - Filling in a system of arrows in a trace-back matrix to uncover a path back to the start in the score matrix.
  - Deducing the alignment by following the arrows in the trace-back matrix.



Pairwise Alignment
Similarity and Relatedness



Sequence Alignment

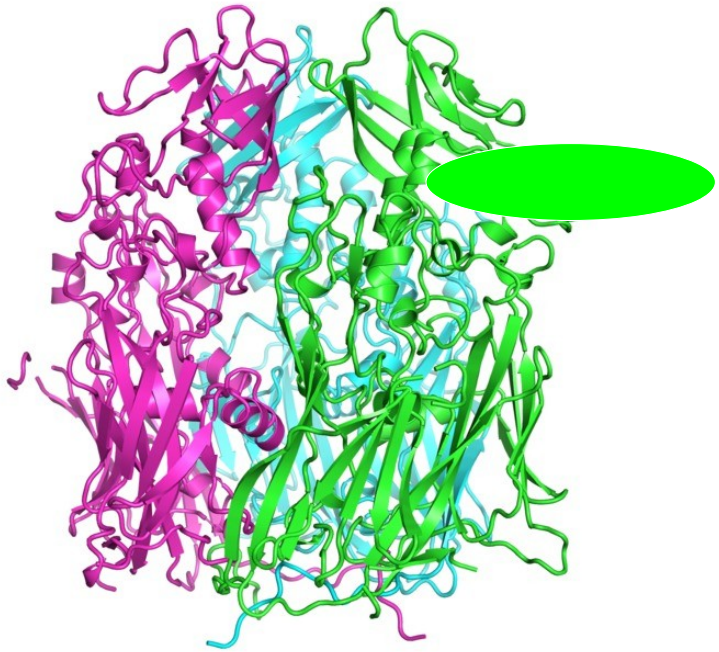
ALLEGHENY COLLEGE

Protein – Amino Acids

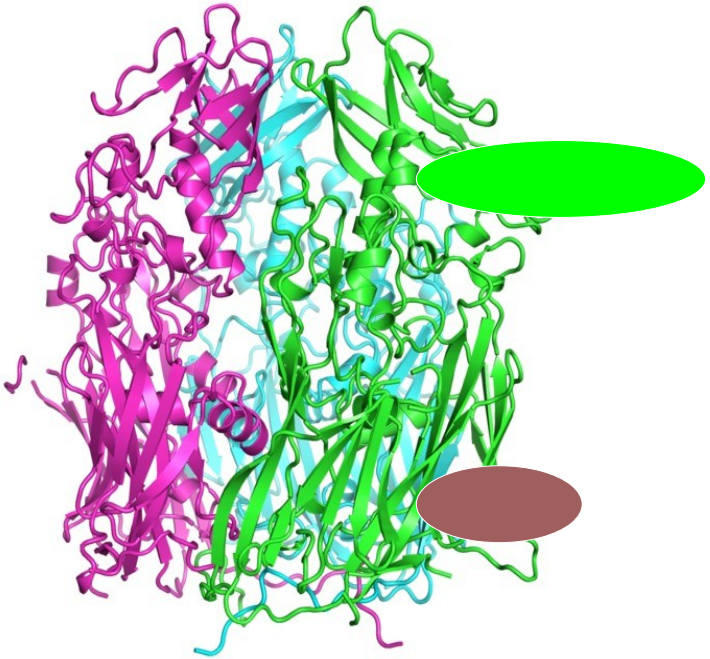
- Understand the use of a substitution matrix to score amino acid similarity in a protein sequence alignment.
- Gain experience using protein alignment to develop hypotheses about protein function based on sequence similarity.
- Know how protein alignment differs algorithmically from DNA alignment.
- 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

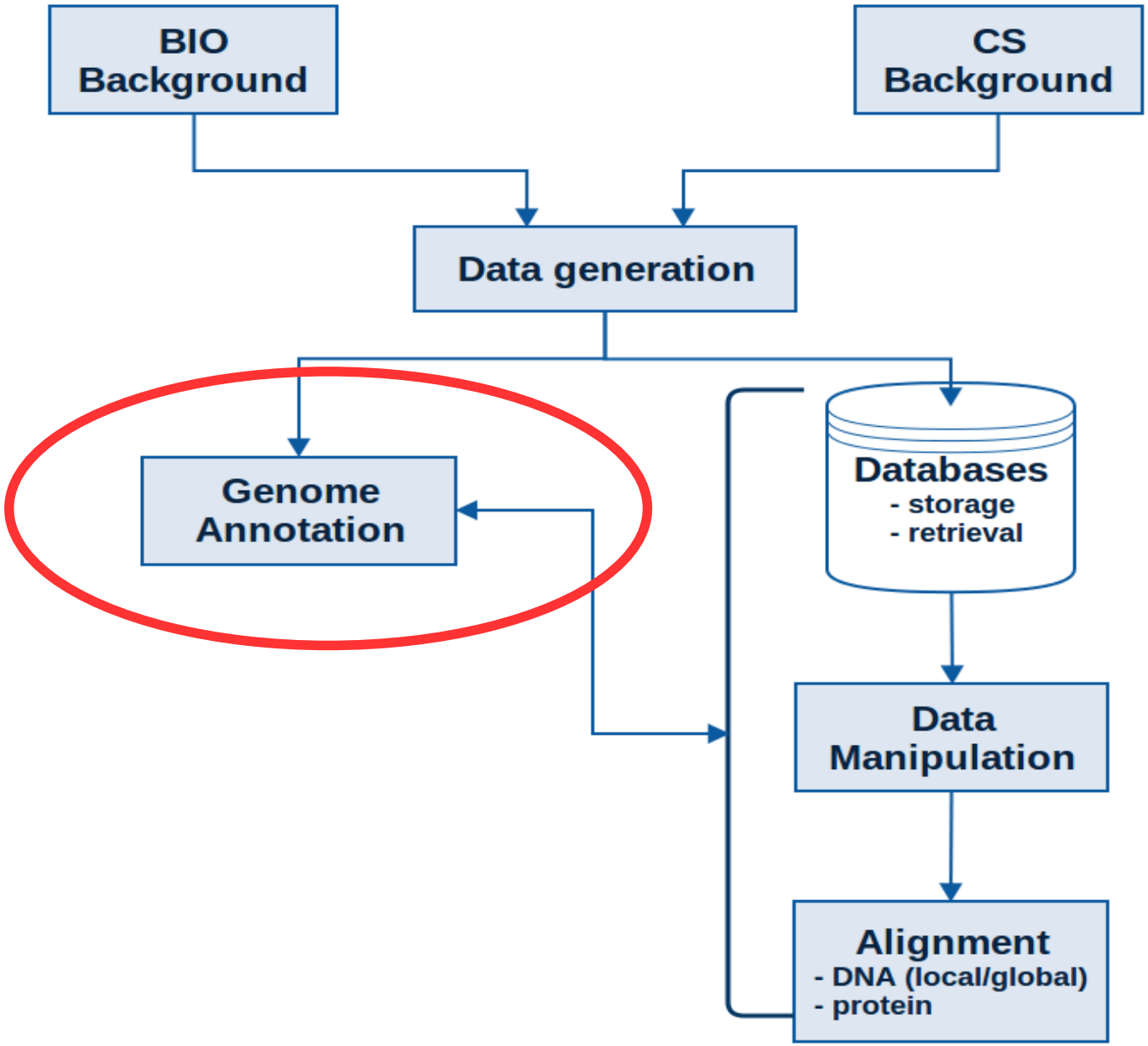


Unhealthy

We Talked About…

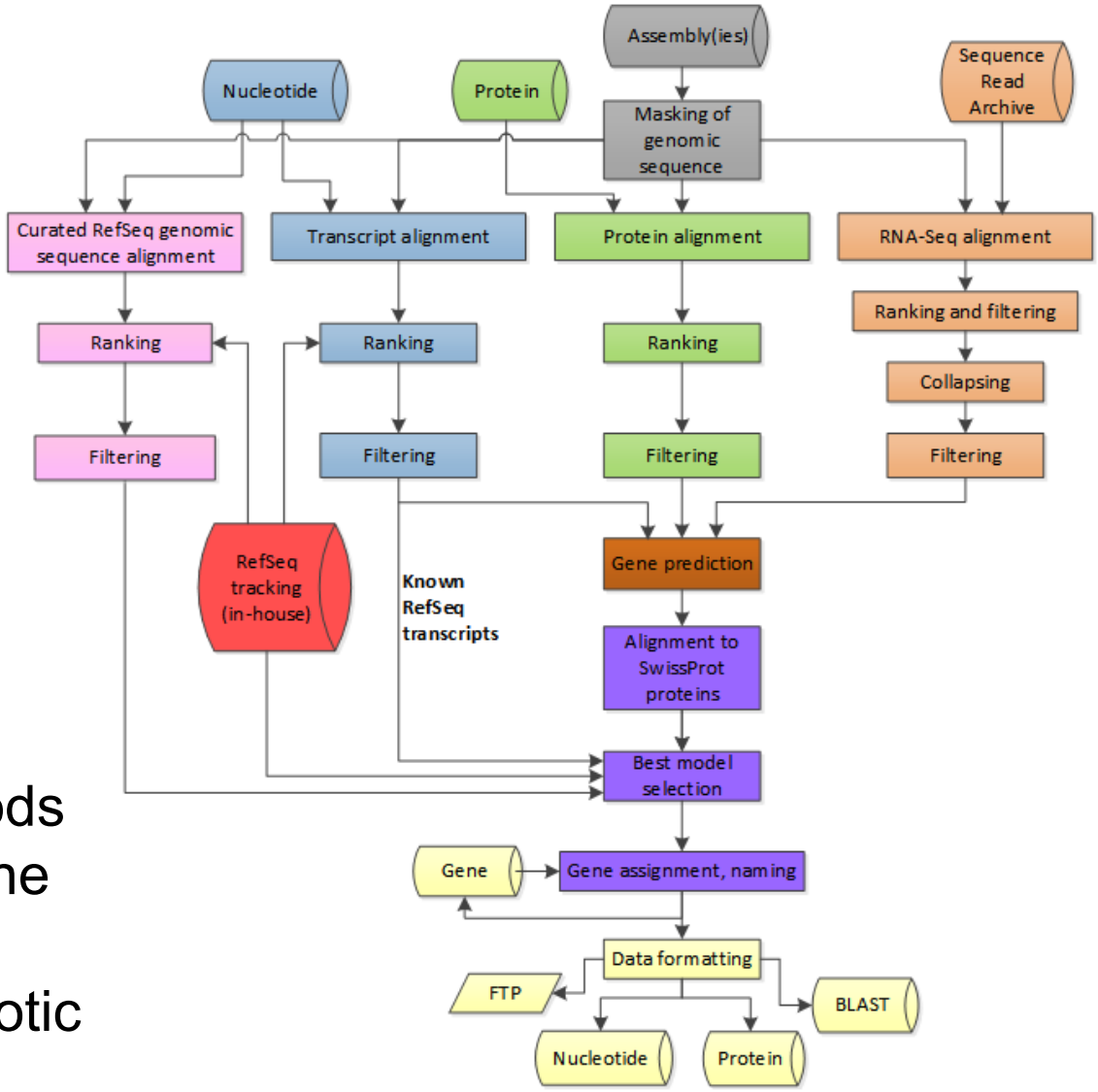
***Genome annotation*: Finding relevant regions in sequences**

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.

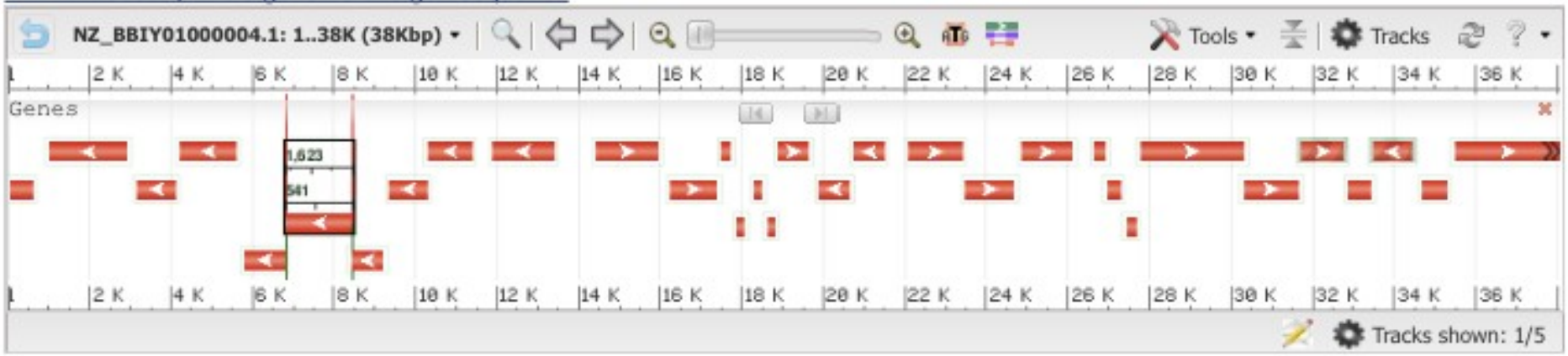
Genome Assembly Annotation

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

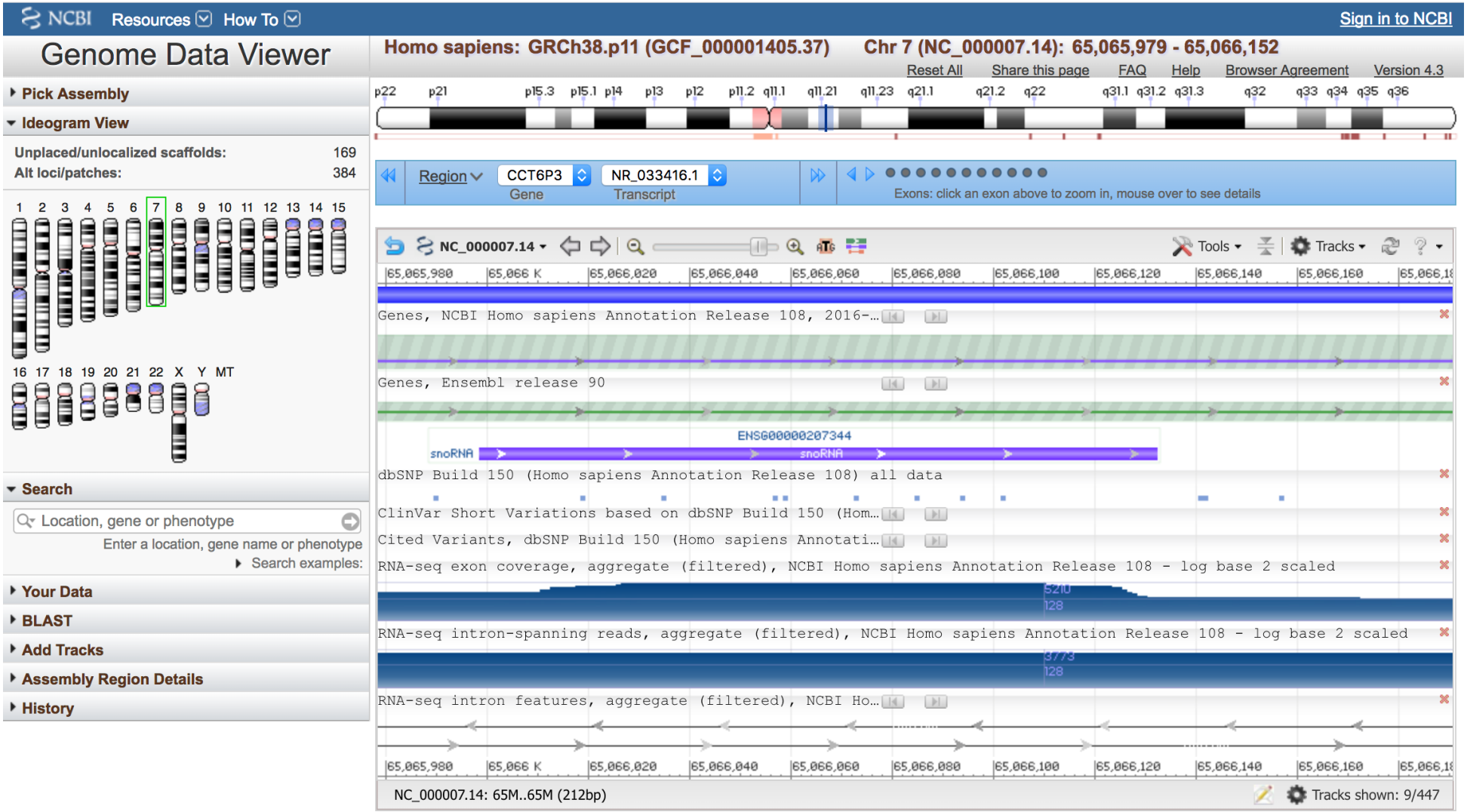
'Chrysanthemum coronarium' phytoplasma strain OY-V
BBIY01000004, whole genome shotgun sequence

Go to nucleotide: Graphics FASTA GenBank



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

Genome Annotation

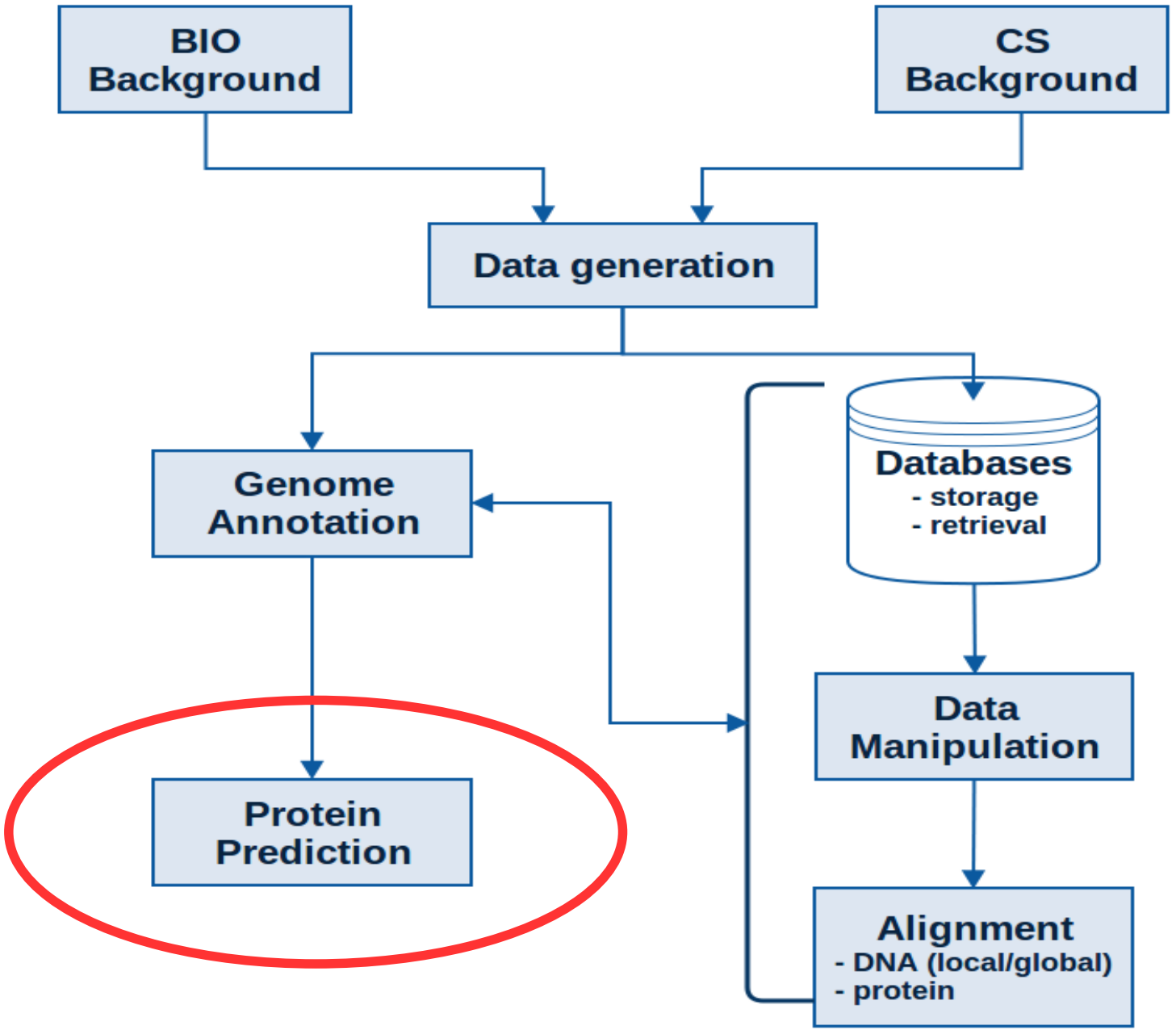


https://www.ncbi.nlm.nih.gov/genome/gdv/browser/?context=Nucleotide&acc=NR_145729.1

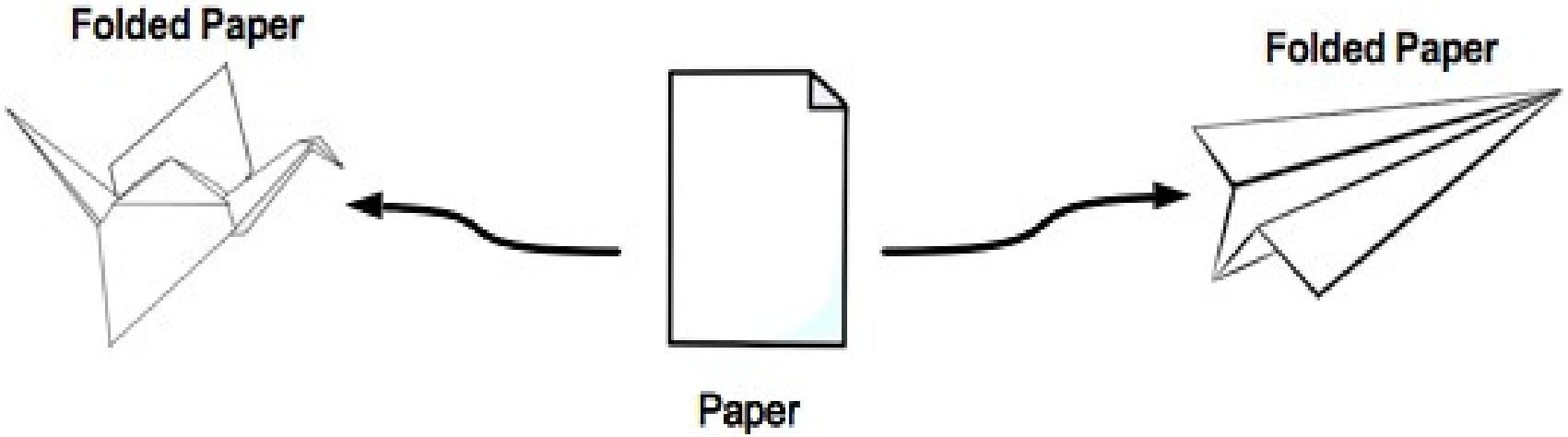
We Talked About…

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

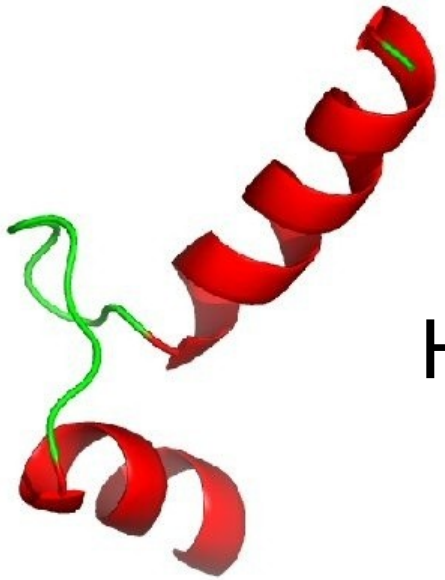
Course Outline



Properties From Folding

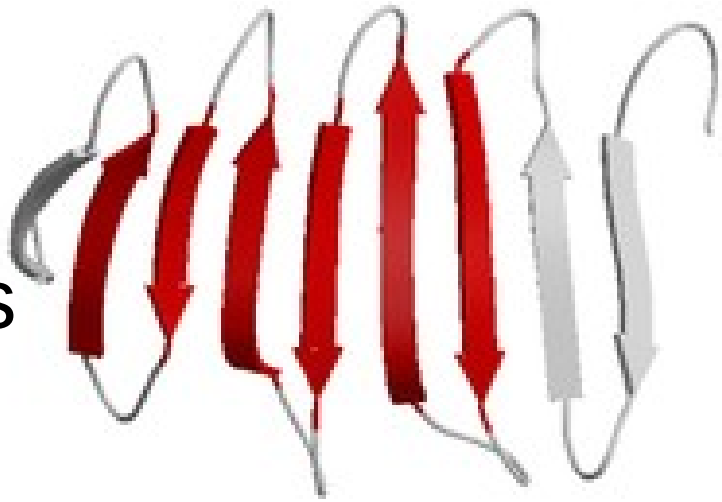


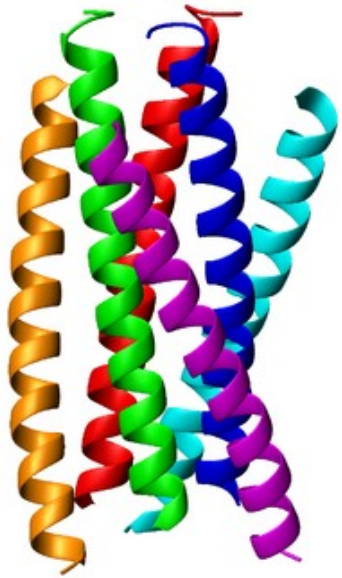
Parts of Protein (Structures)



Helices

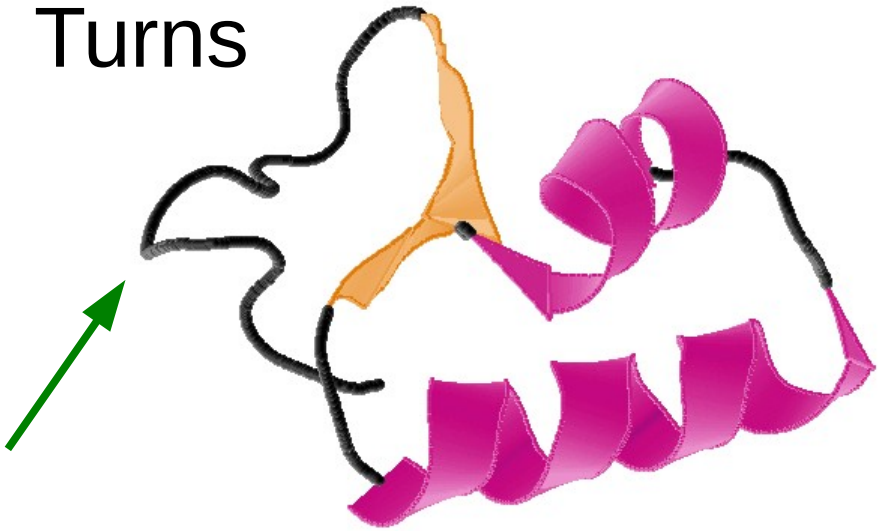
Sheets





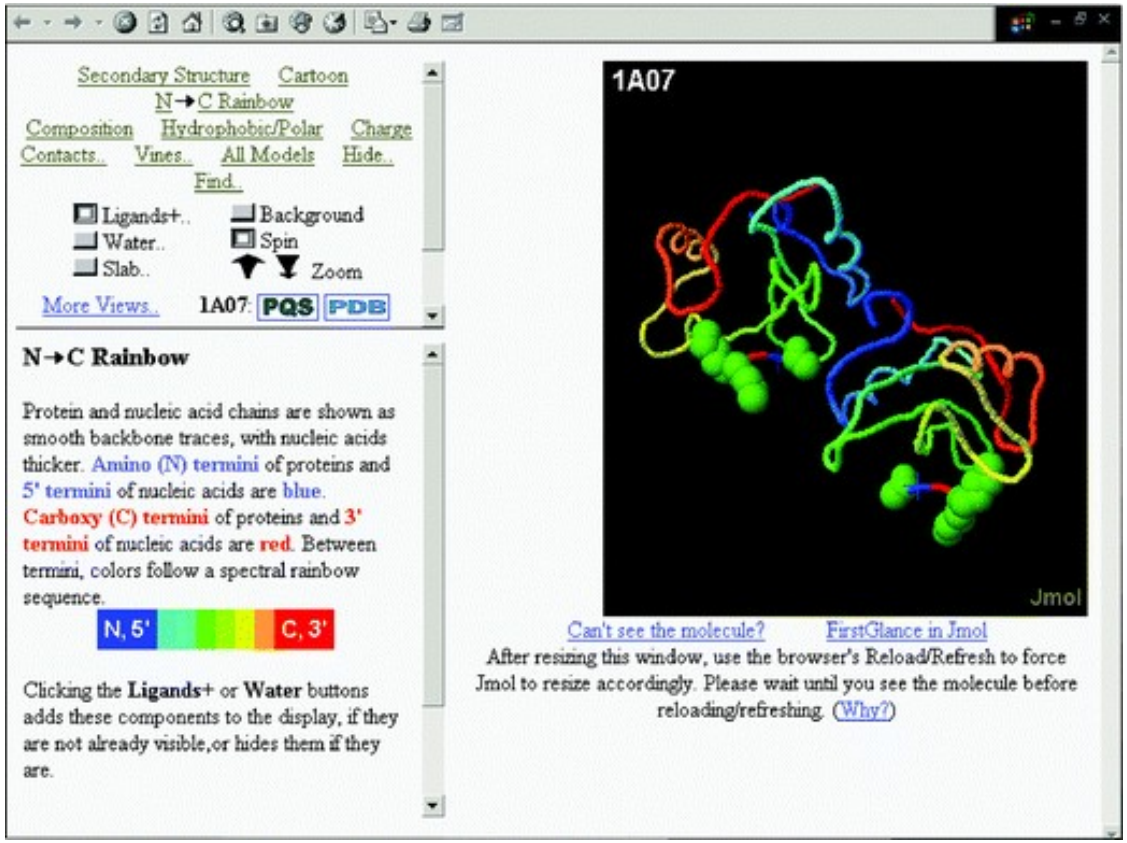
Coils

Turns



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



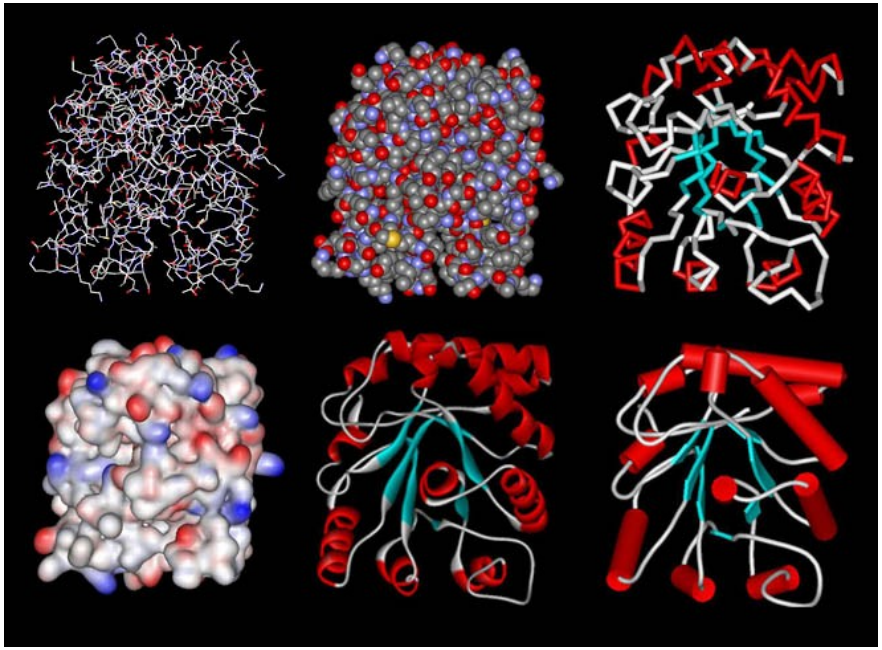
Protein Folding - Applications

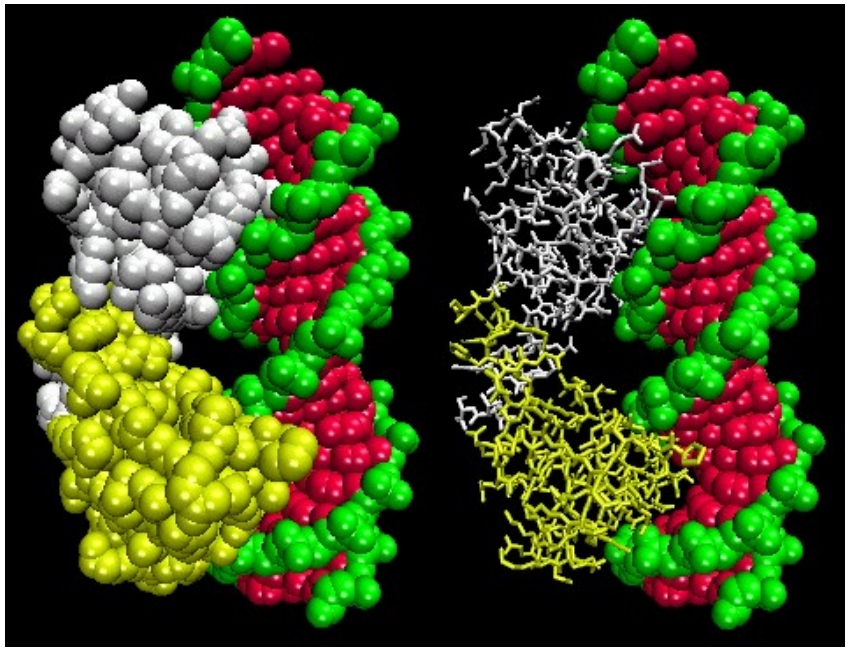
- **Protein must fold correctly to function**

- Misfolded proteins
  - Accumulation – Huntington's and Parkinson's disease
  - Tagged for degradation – emphysema, cystic fibrosis
    - Pharmaceutical chaperones – fold mutated proteins to render them functional
- Proteins must be correctly folded into specific, three-dimensional shapes in order to function correctly.
- Unfolded or misfolded proteins contribute to the pathology of many diseases.

Protein DataBase (PDB)

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



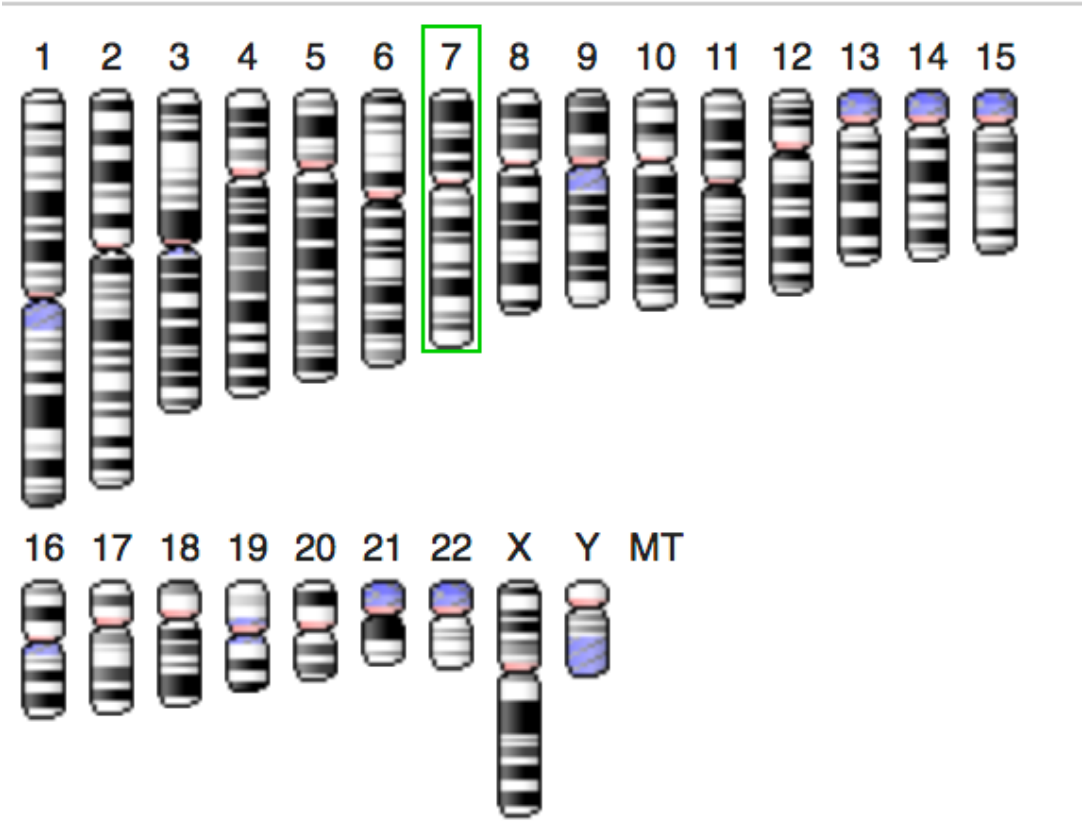


In Closing ...

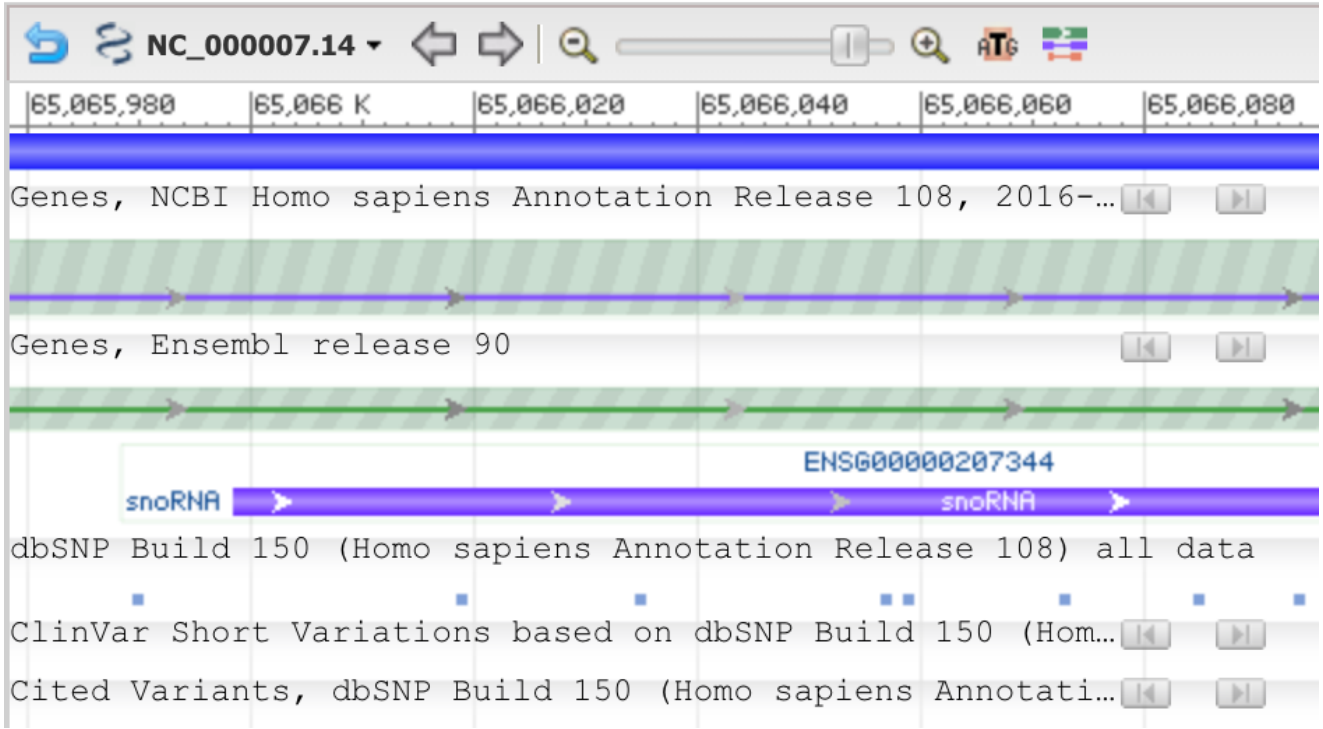
Bioinformatics is diverse and exciting!

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



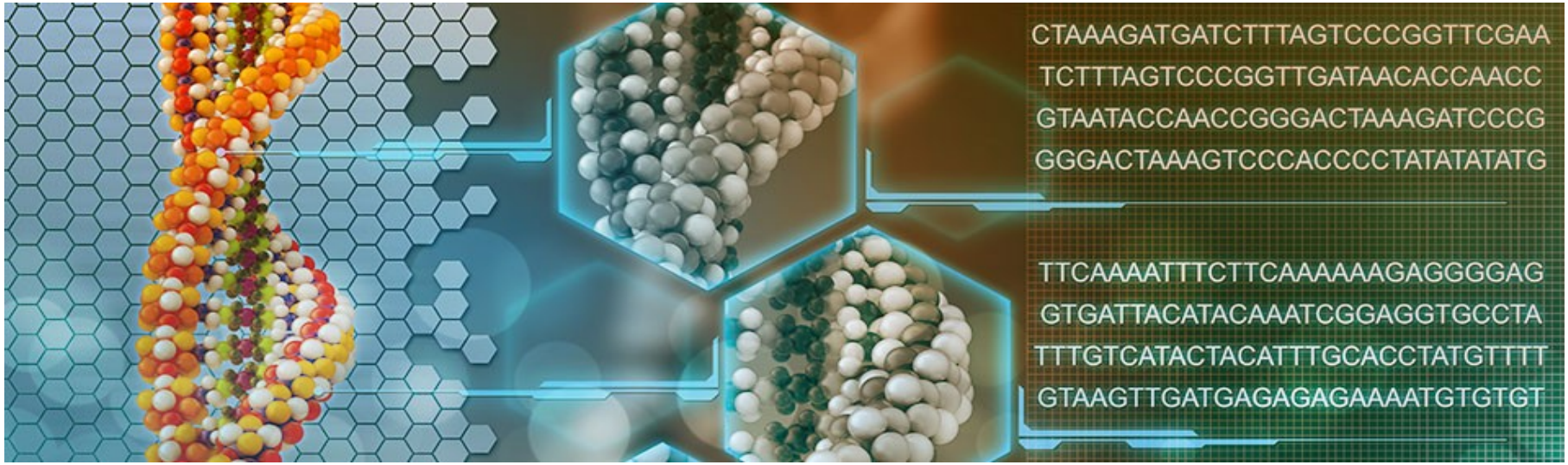
- An understanding of the computational foundations of bioinformatics (programming, databases, etc.).
- An understanding of how genetic information is obtained and processed.
- The ability to use basic bioinformatics software tools to study genetic information.

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!



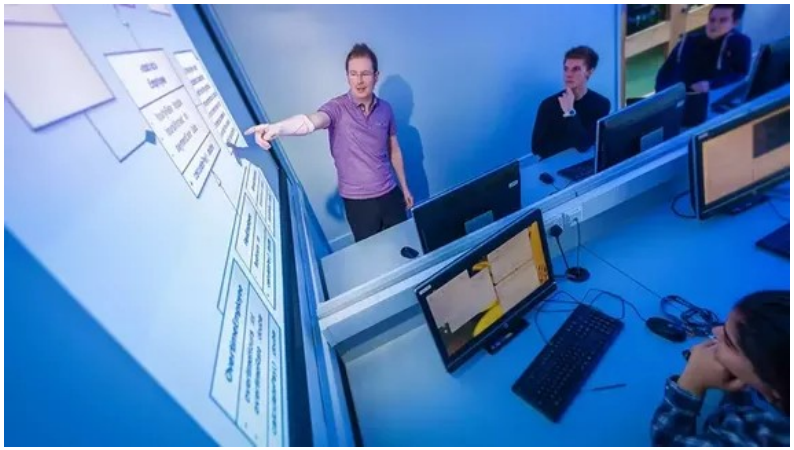
Skills in Careers

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

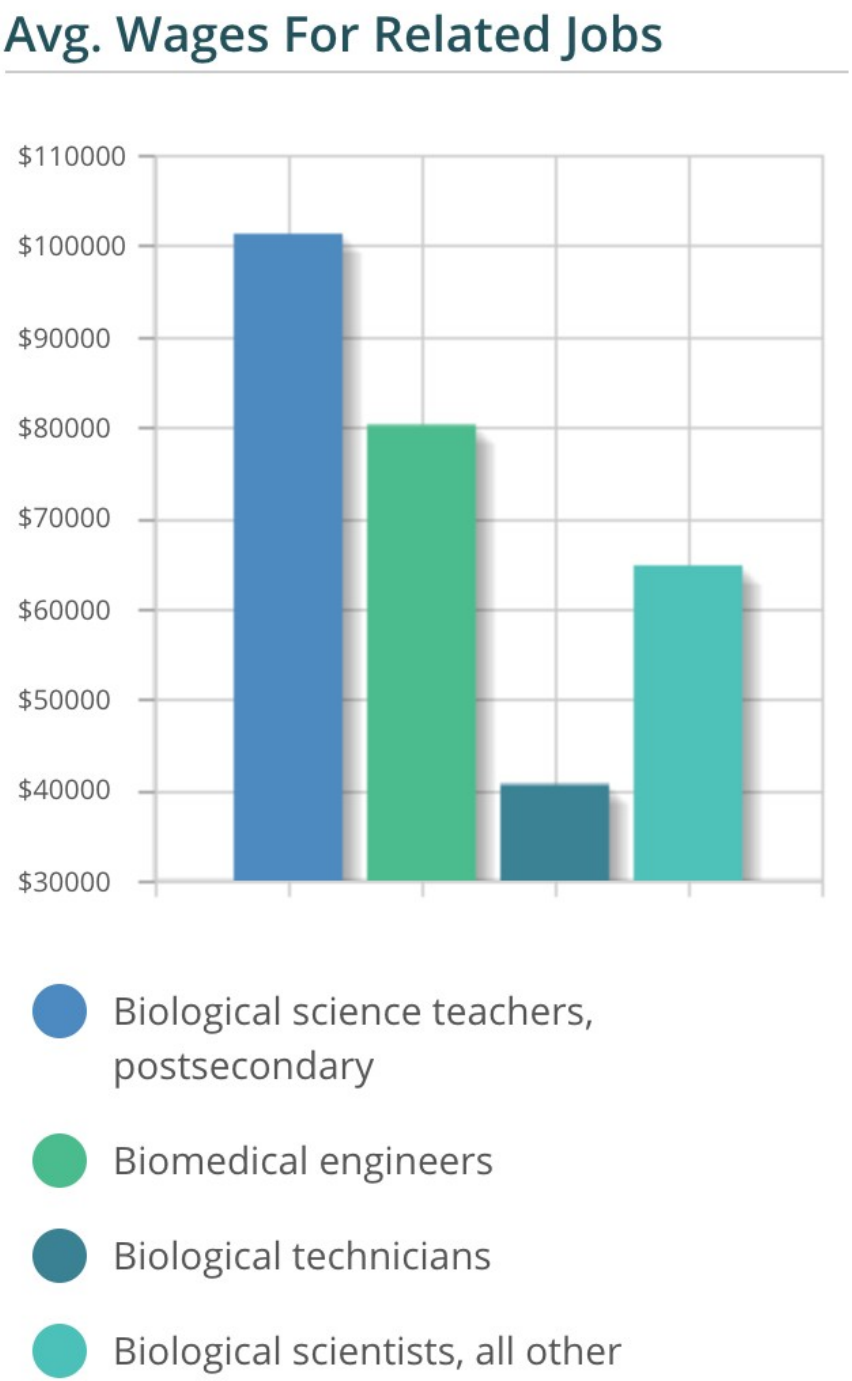
- Computer scientists
  - BioMedical skills
  - Understanding of biological systems and mechanisms
  - Early detection of disease by data
  - Modeling of therapeutic remedies
  - Others

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

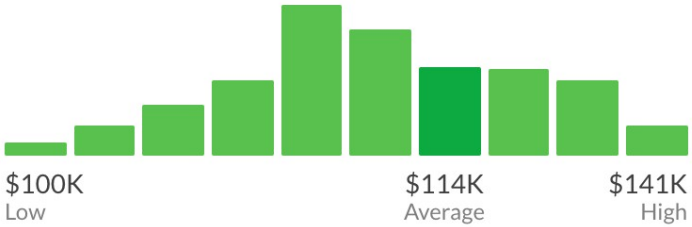
287 Salaries Updated Nov 27, 2017

About This Data

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

Careers in Bioinformatics

- 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 Internship and Career resources

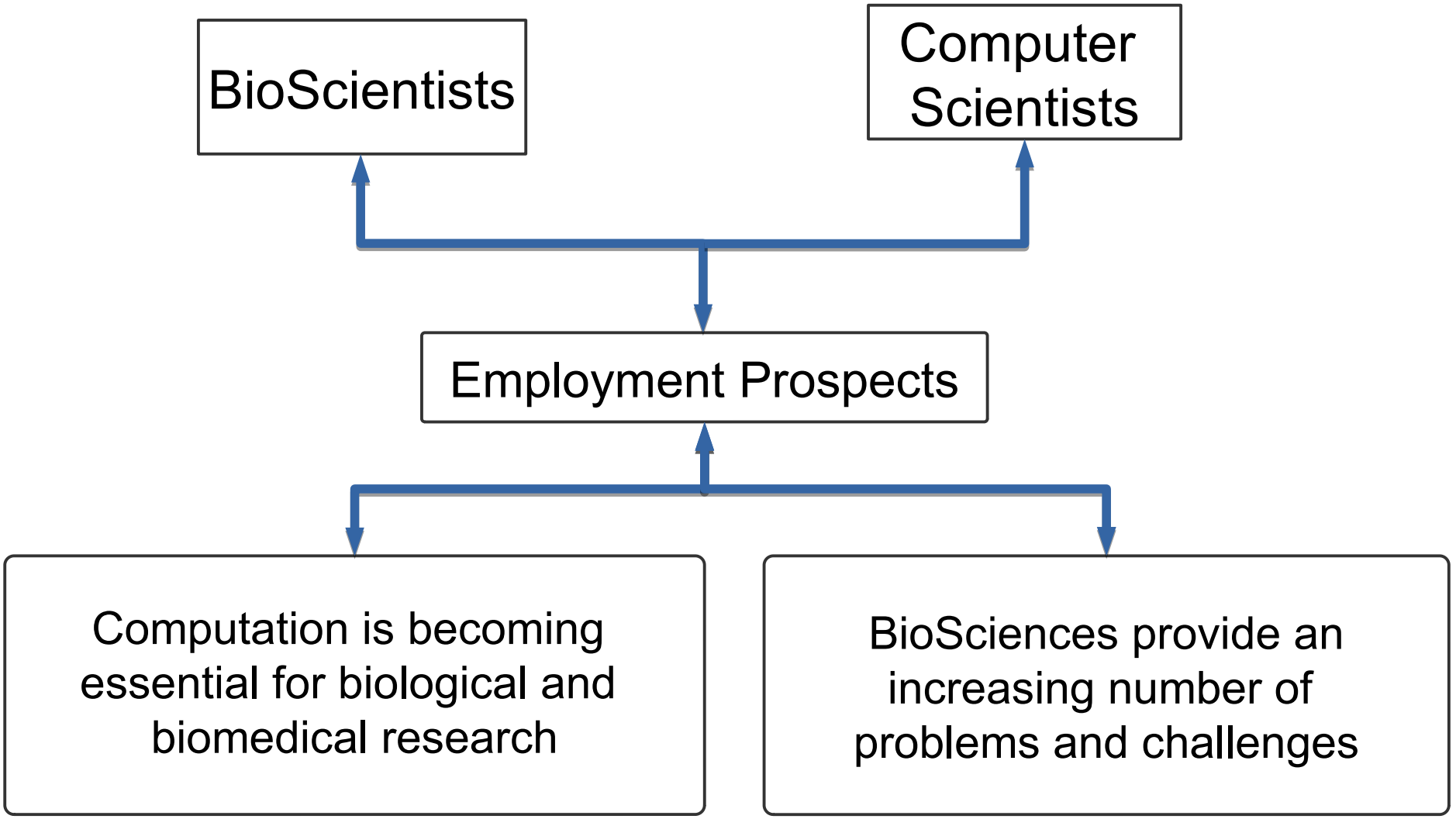
Resources

- 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/

Careers

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

The Value of the Bioinformatics Skills



In Bioinformatics, there is ...

ALLEGHENY COLLEGE