FoCS Breadth:
Overview of Bioinformatics

Niema Moshiri
UC San Diego SPIS 2019

What is **Bioinformatics**?

What is Bioinformatics?

- “Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data” —Wikipedia

What is Bioinformatics?

- “Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data” —Wikipedia
- “The collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied to molecular genetics and genomics” —Webster Dictionary

What is Bioinformatics?

- “Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data” —Wikipedia
- “The collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied to molecular genetics and genomics” —Webster Dictionary
- “Bioinformatics is conceptualizing biology in terms of macromolecules and then applying ‘informatics’ techniques to understand and organize the information associated with these molecules, on a large-scale” —Nick Luscombe

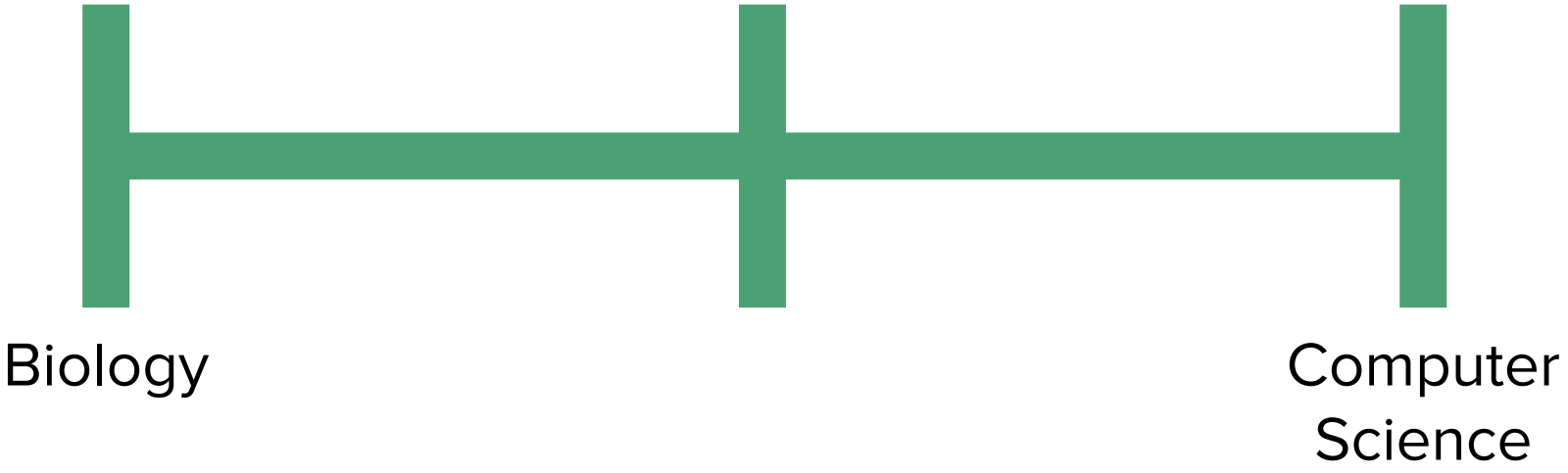
My Definition of Bioinformatics

My Definition of Bioinformatics

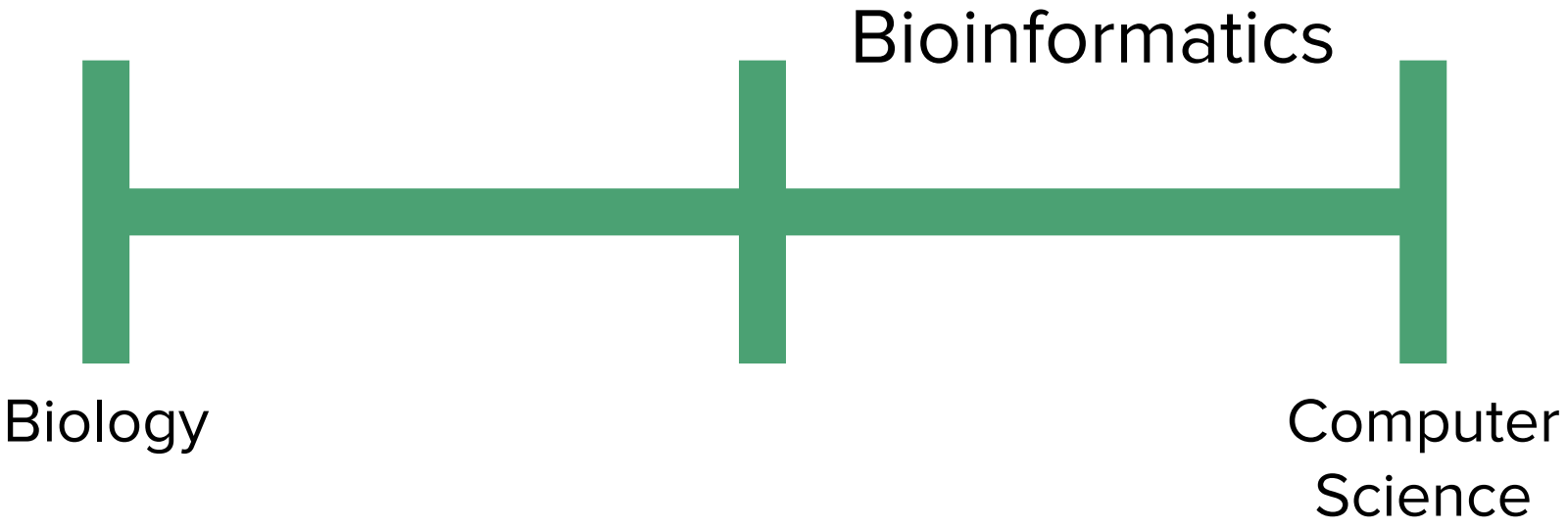
Biology

My Definition of Bioinformatics

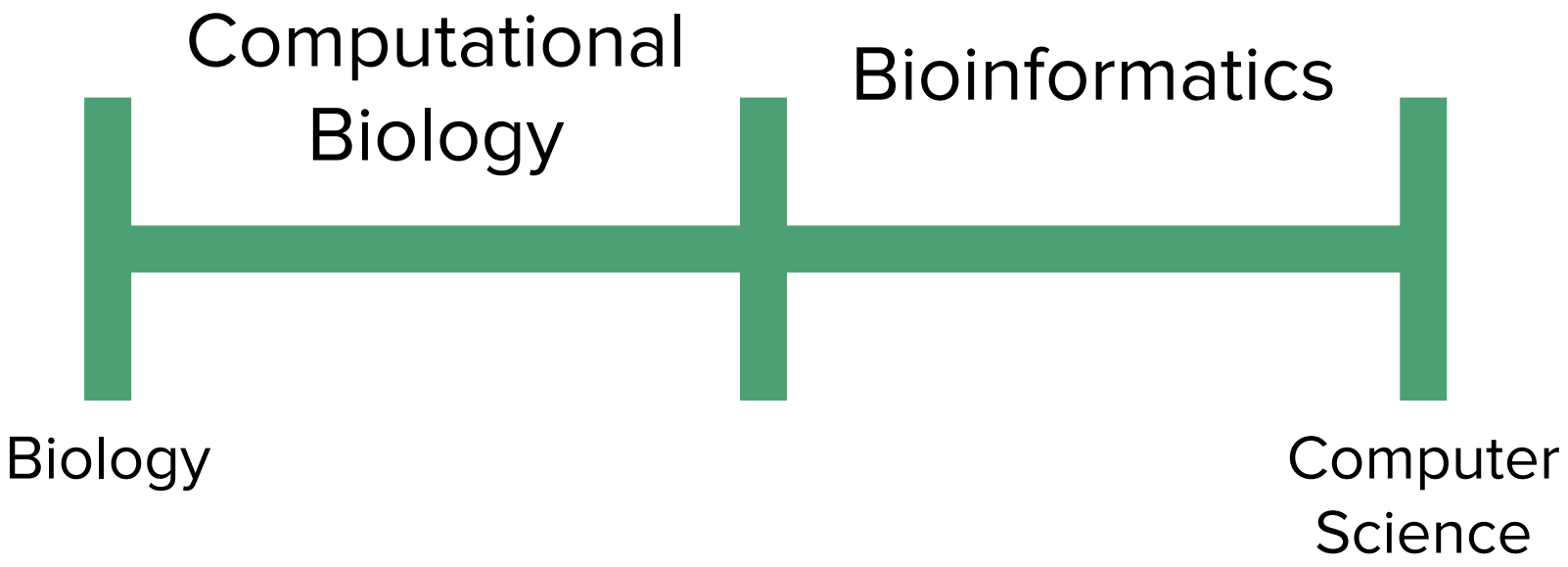
My Definition of Bioinformatics



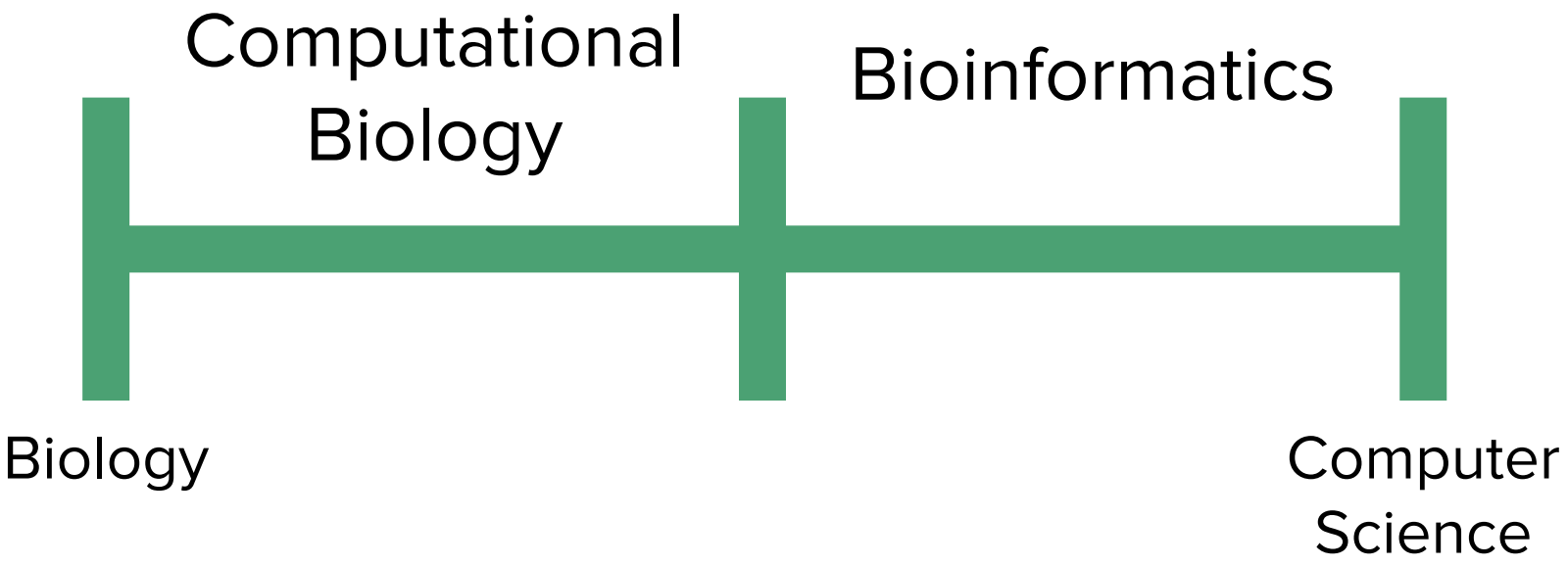
My Definition of Bioinformatics



My Definition of Bioinformatics



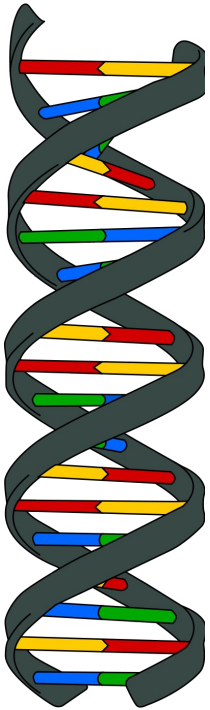
My Definition of Bioinformatics



Chemistry? Physics? Statistics?

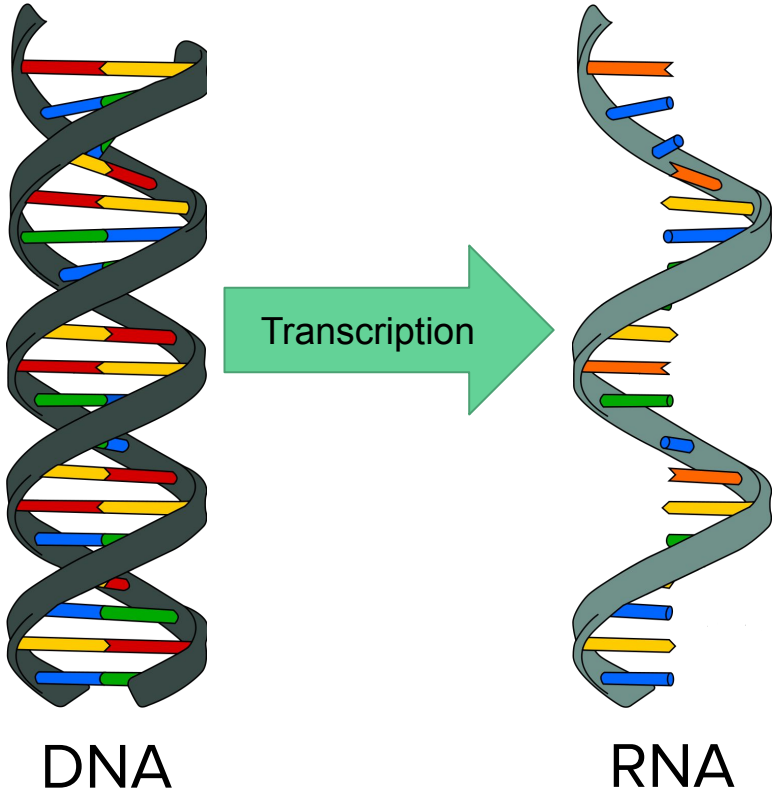
The Central Dogma

The Central Dogma of Biology

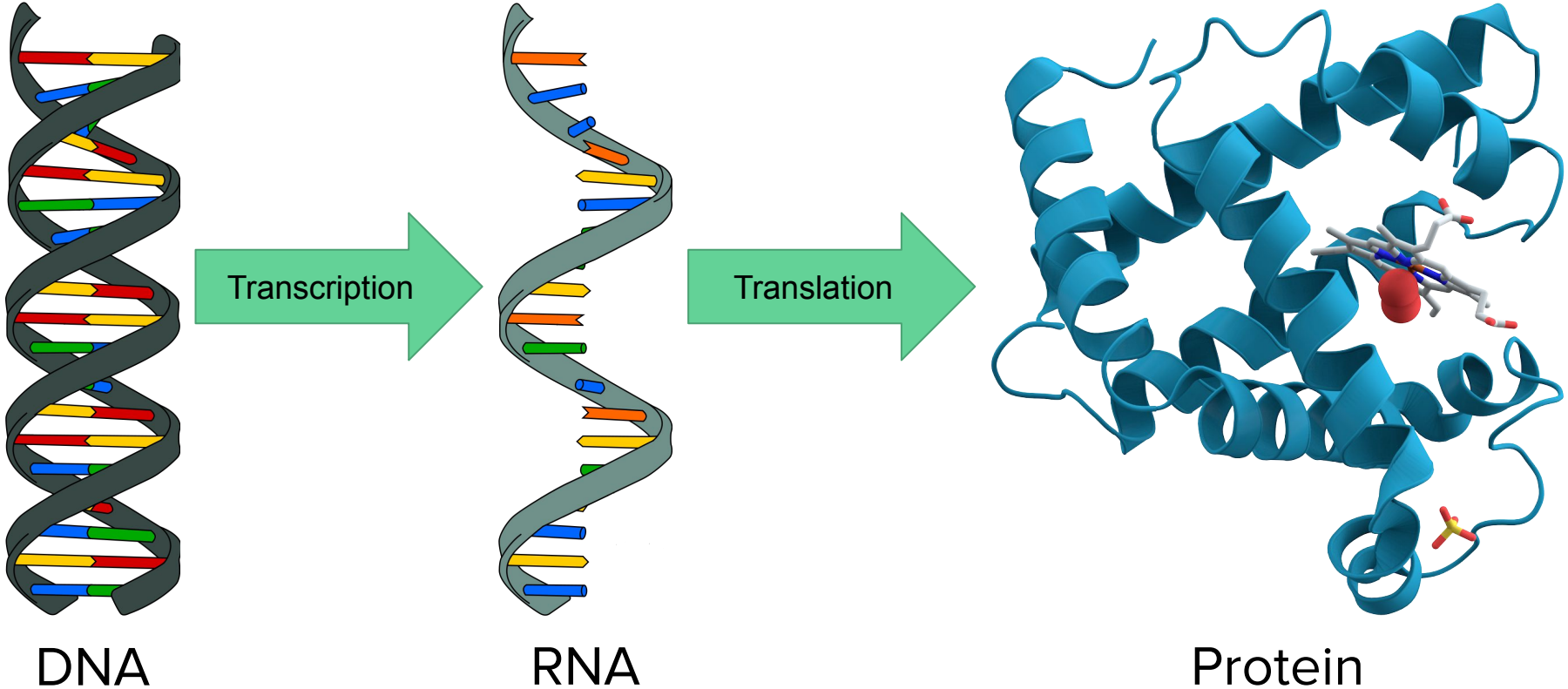


DNA

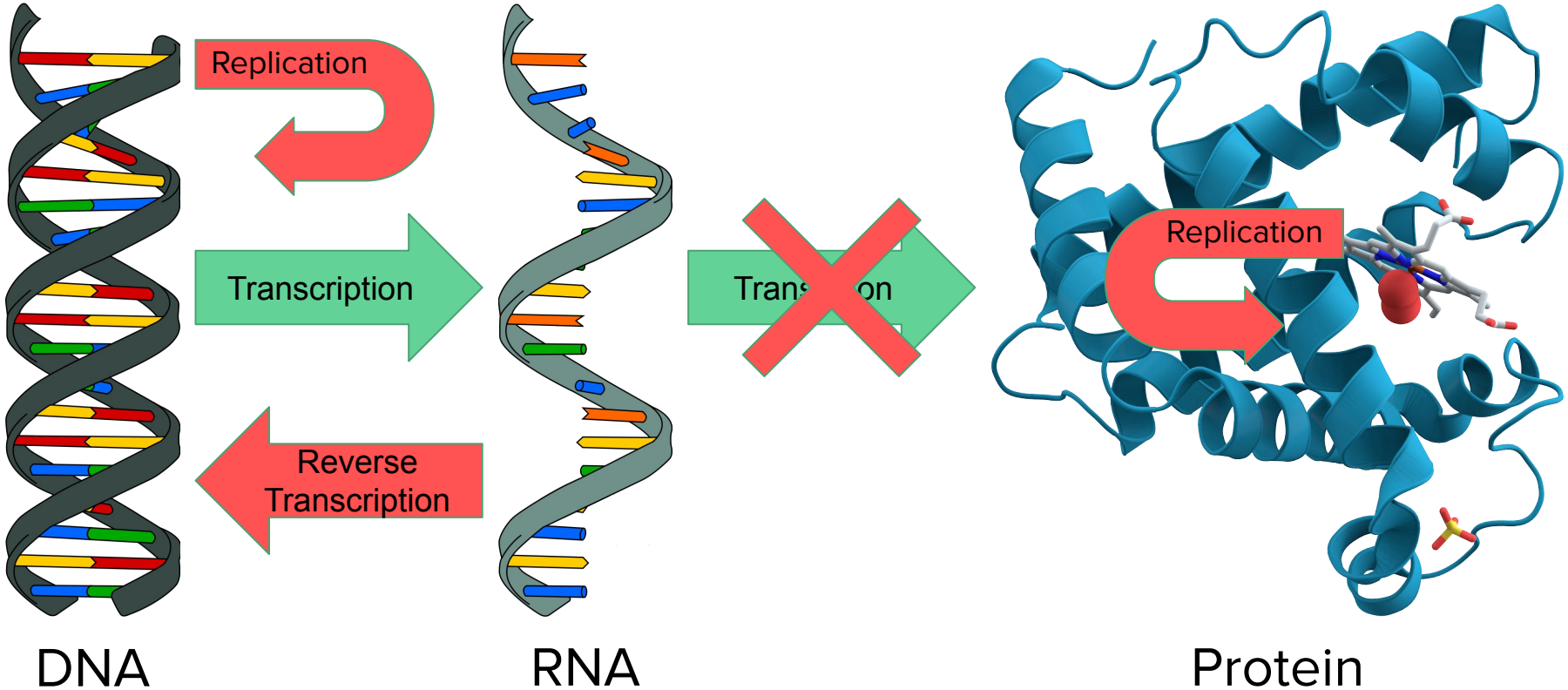
The Central Dogma of Biology



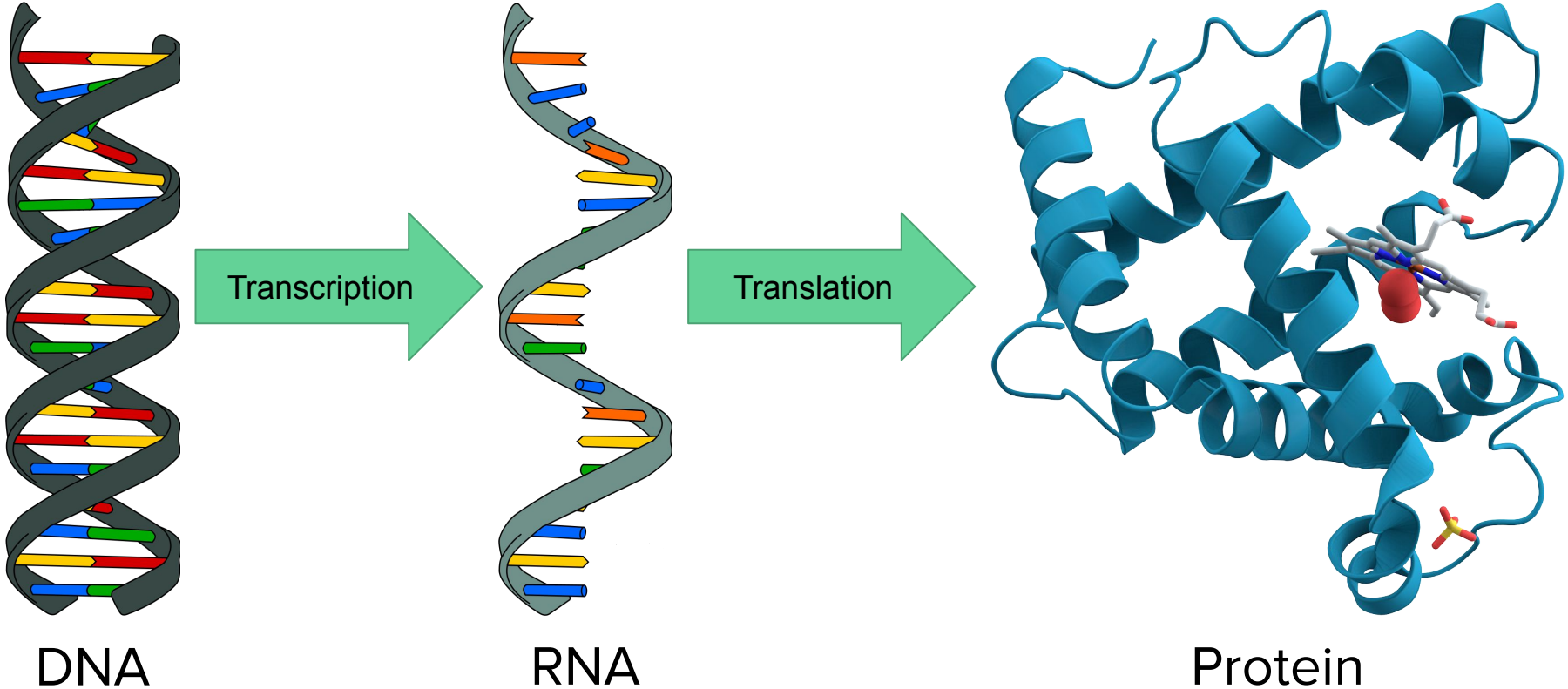
The Central Dogma of Biology



The Central Dogma of Biology



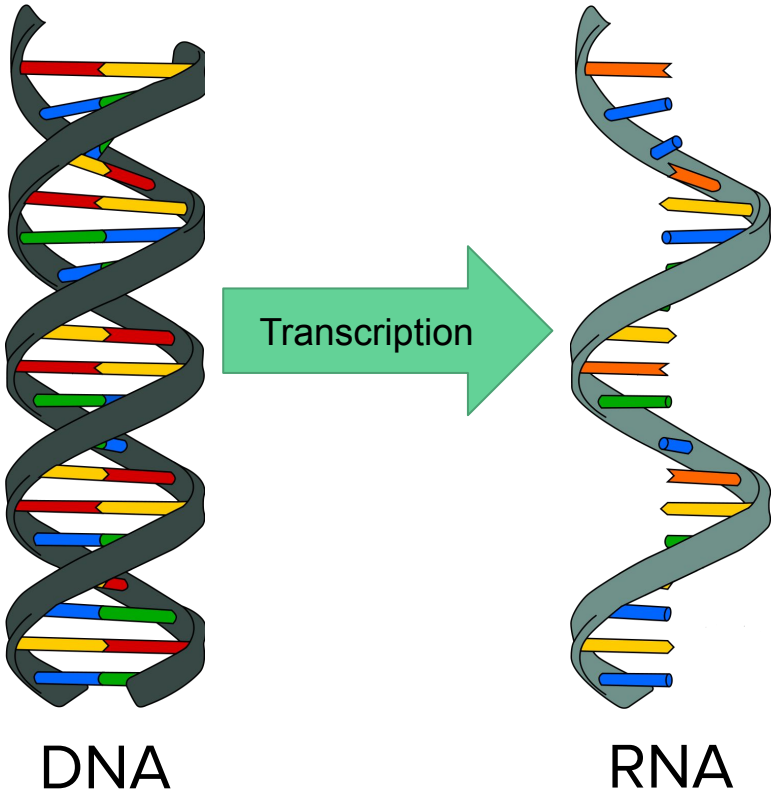
The Central Dogma of Biology



Transcription

Transcription

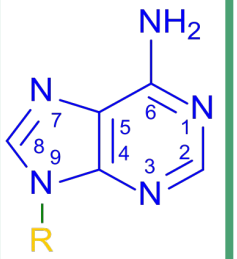
- DNA is **transcribed** to RNA



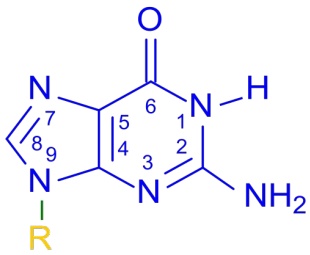
Transcription

- DNA is transcribed to RNA
  - DNA alphabet is {A, C, G, T}

Purines

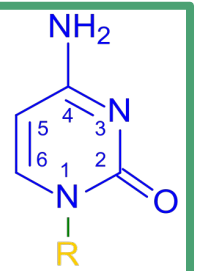


Adenine

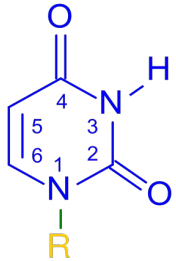


Guanine

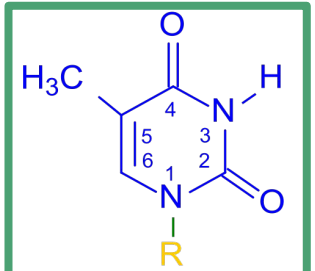
Pyrimidines



Cytosine



Uracil

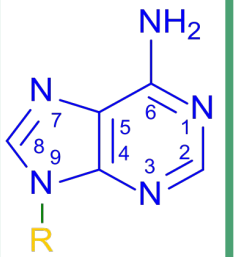


Thymine

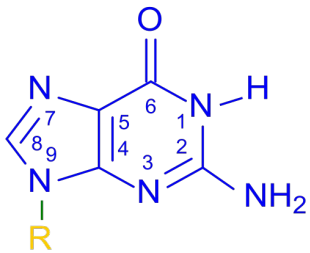
Transcription

- DNA is transcribed to RNA
  - DNA alphabet is {A, C, G, **T**}
  - RNA alphabet is {A, C, G, **U**}

Purines

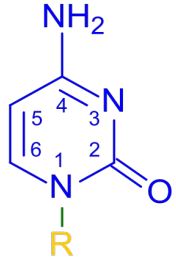


Adenine

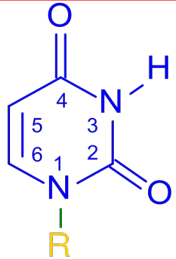


Guanine

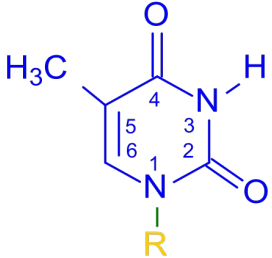
Pyrimidines



Cytosine



Uracil



Thymine

Transcription

- DNA is transcribed to RNA
  - DNA alphabet is {A, C, G, T}
  - RNA alphabet is {A, C, G, U}
- <u>Mechanism</u>

Transcription

- DNA is transcribed to RNA
  - DNA alphabet is {A, C, G, T}
  - RNA alphabet is {A, C, G, U}
- <u>Mechanism</u>
  - Transcription Factor (TF) binds to the gene's promoter

Transcription

- DNA is transcribed to RNA
  - DNA alphabet is {A, C, G, T}
  - RNA alphabet is {A, C, G, U}
- <u>Mechanism</u>
  - Transcription Factor (TF) binds to the gene's promoter
  - RNA Polymerase binds near the transcription start site

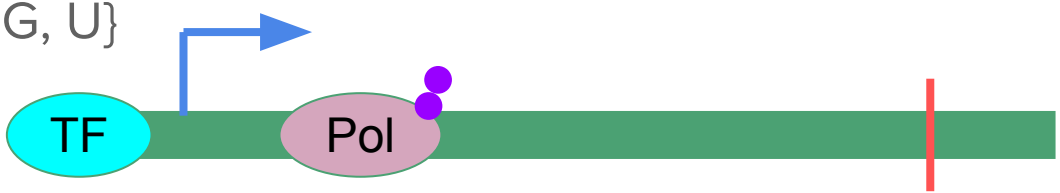
Transcription

- DNA is transcribed to RNA
  - DNA alphabet is {A, C, G, T}
  - RNA alphabet is {A, C, G, U}
- <u>Mechanism</u>

  - Transcription Factor (TF) binds to the gene's promoter
  - RNA Polymerase binds near the transcription start site
  - RNA Polymerase transcribes DNA to RNA...

Transcription

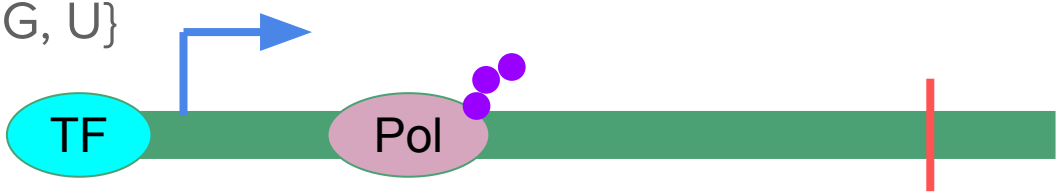
- DNA is transcribed to RNA
  - DNA alphabet is {A, C, G, T}
  - RNA alphabet is {A, C, G, U}
- <u>Mechanism</u>



  - Transcription Factor (TF) binds to the gene's promoter
  - RNA Polymerase binds near the transcription start site
  - RNA Polymerase transcribes DNA to RNA...

Transcription

- DNA is transcribed to RNA
  - DNA alphabet is {A, C, G, T}
  - RNA alphabet is {A, C, G, U}
- <u>Mechanism</u>



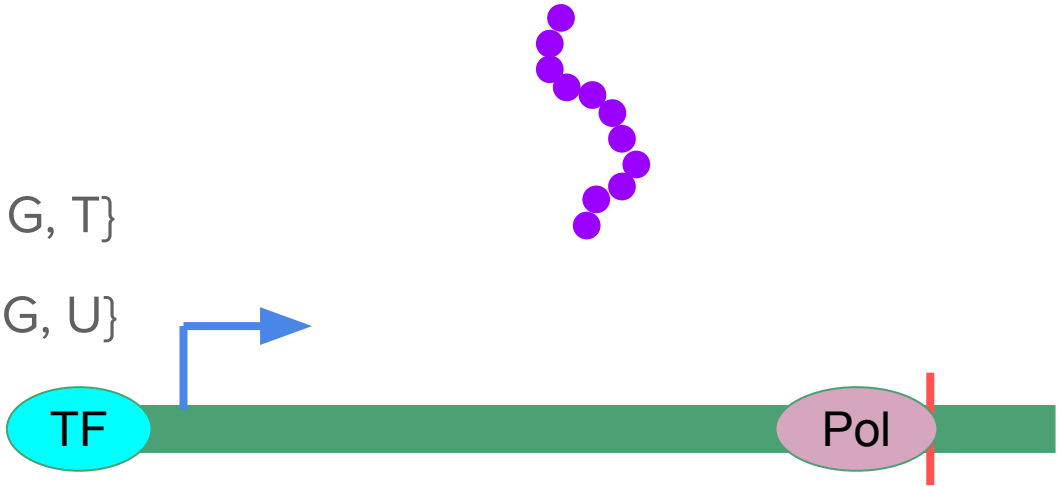
  - Transcription Factor (TF) binds to the gene's promoter
  - RNA Polymerase binds near the transcription start site
  - RNA Polymerase transcribes DNA to RNA...

Transcription

- DNA is transcribed to RNA
  - DNA alphabet is {A, C, G, T}
  - RNA alphabet is {A, C, G, U}
- <u>Mechanism</u>
  - Transcription Factor (TF) binds to the gene's promoter
  - RNA Polymerase binds near the transcription start site
  - RNA Polymerase transcribes DNA to RNA... until it hits the transcription end site

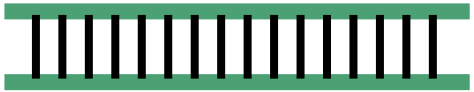
Transcription

- DNA is transcribed to RNA
  - DNA alphabet is {A, C, G, T}
  - RNA alphabet is {A, C, G, U}
- <u>Mechanism</u>
  - Transcription Factor (TF) binds to the gene's promoter
  - RNA Polymerase binds near the transcription start site
  - RNA Polymerase transcribes DNA to RNA... until it hits the transcription end site



Transcription: Summary

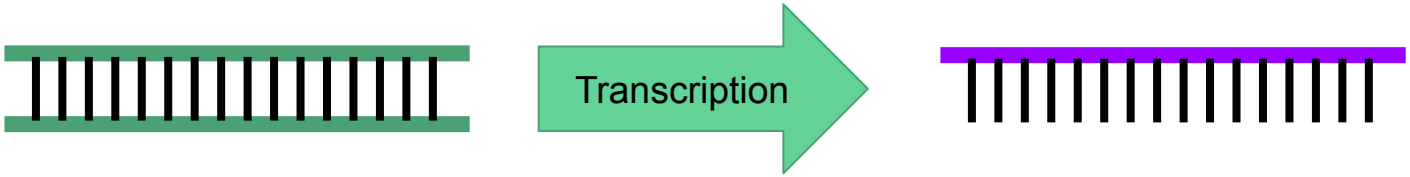
DNA: GAGCTGATGGCTACTACACATATTGCCAGTTGATGGGTT



Transcription: Summary

DNA: GAGCTGATGGCTACTACACATATTGCCAGTTGATGGGTT

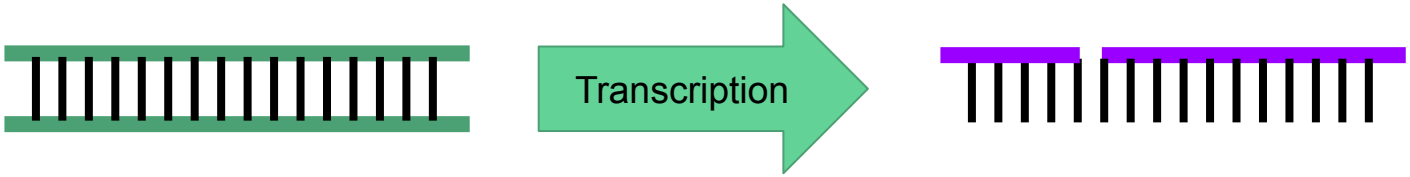
RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGUUGAUGGGUU



Transcription: Summary

DNA: GAGCTGATGGCTACTACACATATTGCCAGTTGATGGGTT

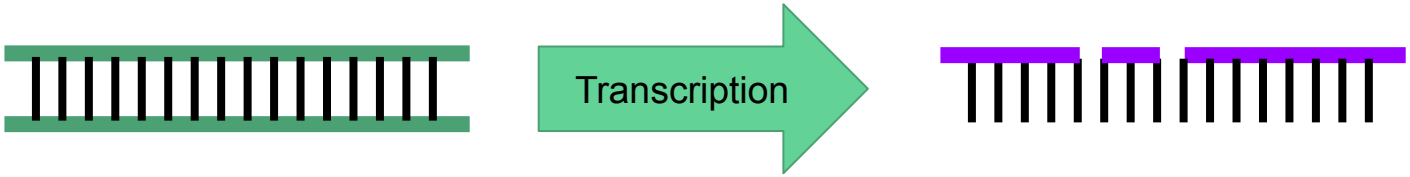
RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGUUGAUGGGUU



Transcription: Summary

DNA: GAGCTGATGGCTACTACACATATTGCCAGTTGATGGGTT

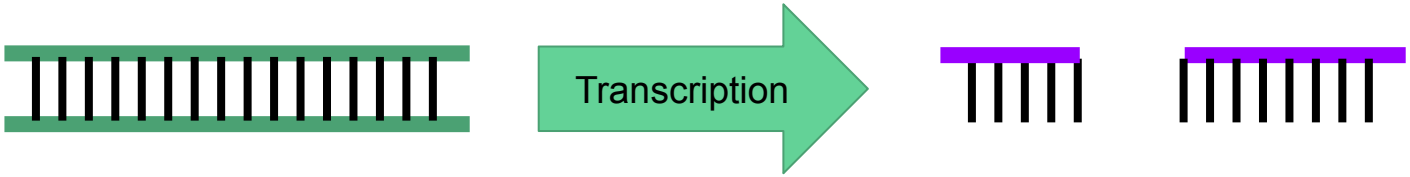
RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGUUGAUGGGUU



Transcription: Summary

DNA: GAGCTGATGGCTACTACACATATTGCCAGTTGATGGGTT

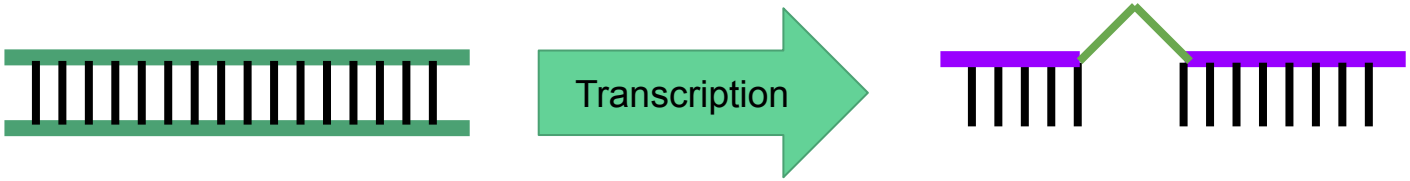
RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGUUGAUGGGUU



Transcription: Summary

DNA: GAGCTGATGGCTACTACACATATTGCCAGTTGATGGGTT

RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGUUGAUGGGUU



Transcription: Summary

DNA: GAGCTGATGGCTACTACACATATTGCCAGTTGATGGGTT

RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGUUGAUGGGUU

Transcription: Summary

DNA: GAGCTGATGGCTACTACACATATTGCCAGTTGATGGGTT

RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGUUGAUGGGUU

Transcription: Summary

DNA: GAGCTGATGGCTACTACACATATTGCCAGTTGATGGGTT

RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGUUGAUGGGUU

Transcription: Summary

DNA: GAGCTGATGGCTACTACACATATTGCCAGTTGATGGGTT

RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGUUGAUGGGUU

Transcription: Summary

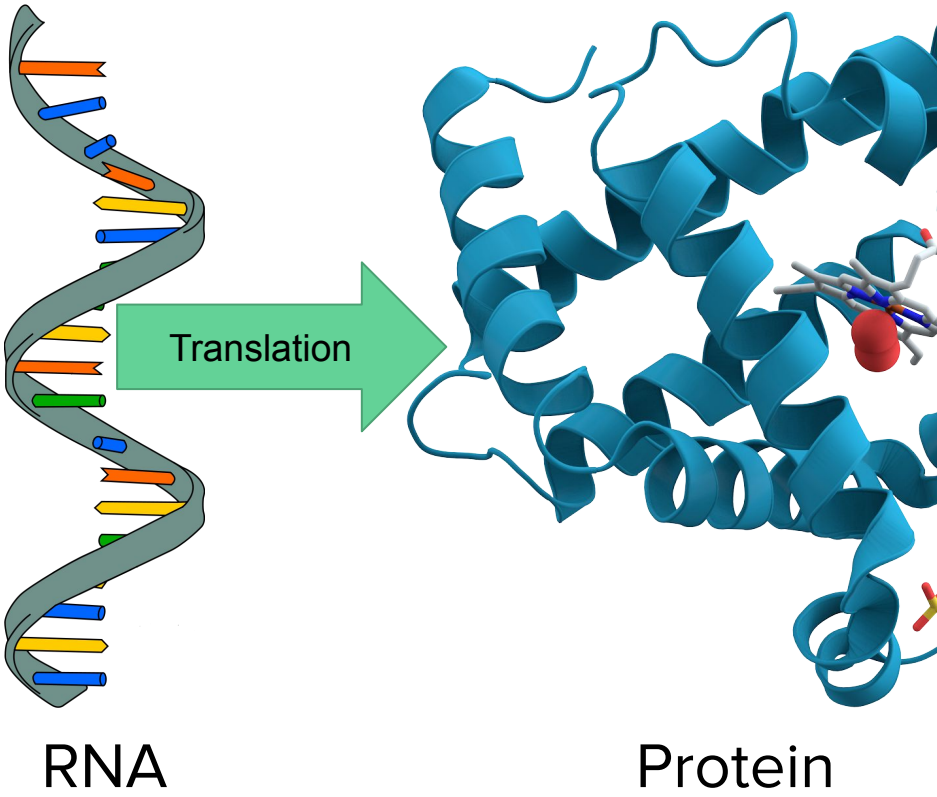
DNA: GAGCTGATGGCTACTACACATATTGCCAGTTGATGGGTT

RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGUUGAUGGGUU

Translation

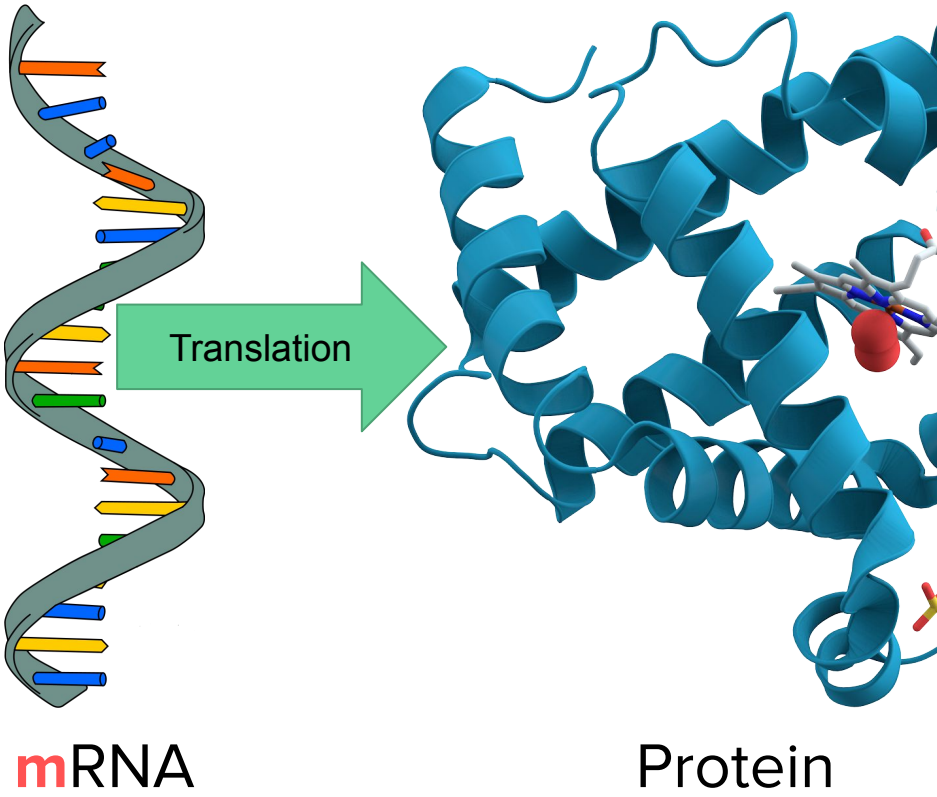
Translation

- RNA is **translated** to Protein



Translation

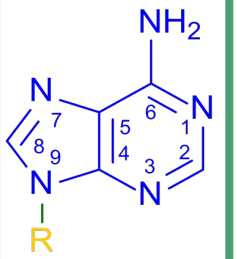
- mRNA is **translated** to Protein
  - “Messenger” RNA



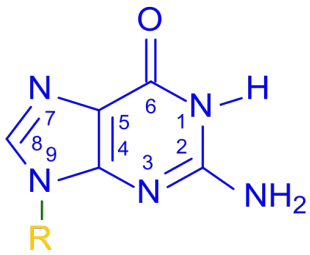
Translation

- mRNA is translated to Protein
  - “Messenger” RNA
  - RNA alphabet is {A, C, G, U}

Purines

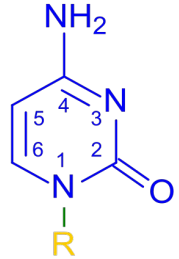


Adenine

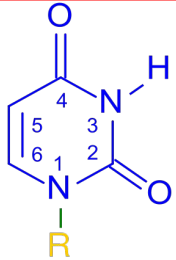


Guanine

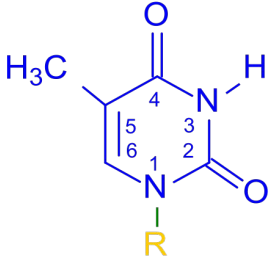
Pyrimidines



Cytosine



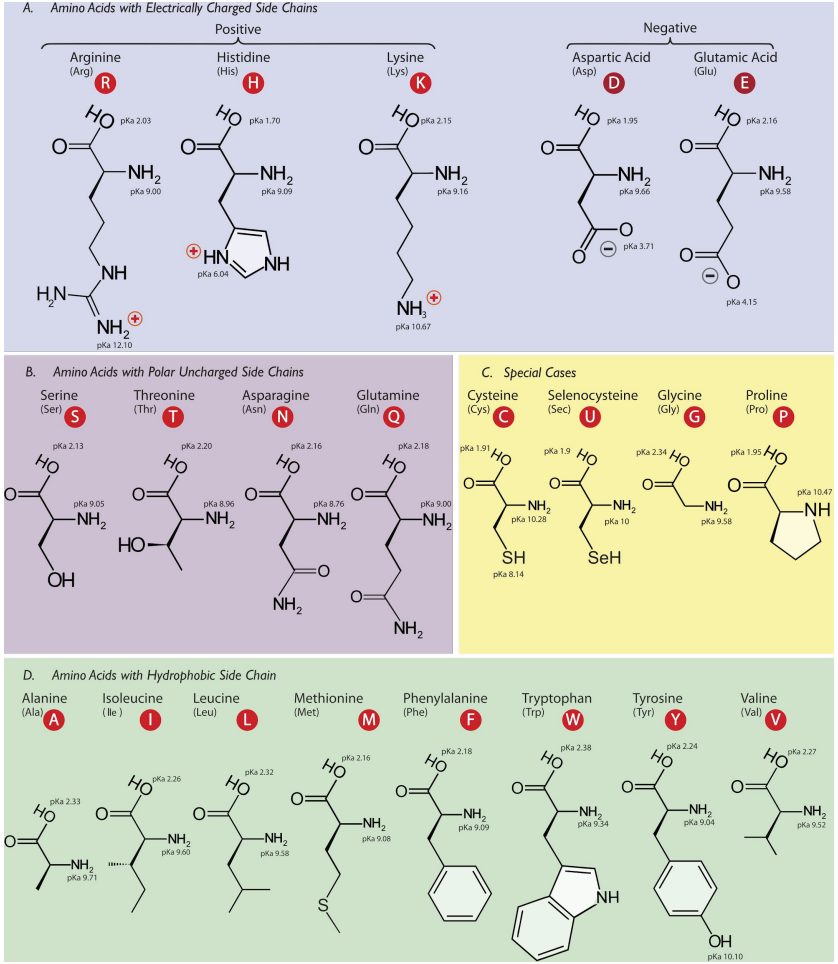
Uracil



Thymine

Translation

- mRNA is translated to Protein
  - “Messenger” RNA
  - RNA alphabet is {A, C, G, U}
  - Protein alphabet is 20 letters



Translation

- mRNA is translated to Protein
  - “Messenger” RNA
  - RNA alphabet is {A, C, G, U}
  - Protein alphabet is 20 letters
  - Each triplet (“codon”) of RNA maps to a specific amino acid

first letter	second letter: U	C	A	G	third letter
U	UUU, UUC Phe UUA, UUG Leu	UCU, UCC, UCA, UCG Ser	UAU, UAC Tyr UAA stop UAG stop	UGU, UGC Cys UGA stop UGG Trp	U C A G
C	CUU, CUC, CUA, CUG Leu	CCU, CCC, CCA, CCG Pro	CAU, CAC His CAA, CAG Gln	CGU, CGC, CGA, CGG Arg	U C A G
A	AUU, AUC, AUA Ile AUG Met	ACU, ACC, ACA, ACG Thr	AAU, AAC Asn AAA, AAG Lys	AGU, AGC Ser AGA, AGG Arg	U C A G
G	GUU, GUC, GUA, GUG Val	GCU, GCC, GCA, GCG Ala	GAU, GAC Asp GAA, GAG Glu	GGU, GGC, GGA, GGG Gly	U C A G

Translation: Mechanism

- Translation starts at an early-on AUG (not necessarily the first)

Translation: Mechanism

- Translation starts at an early-on AUG (not necessarily the first)
- Starting with AUG, each codon is “translated” to a specific amino acid

Translation: Mechanism

- Translation starts at an early-on AUG (not necessarily the first)
- Starting with AUG, each codon is “translated” to a specific amino acid
- Translation continues codon-by-codon until a STOP codon is reached

Translation: Summary

RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGUUGAUGGGUU

Protein:

Translation: Summary

```
    RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGUUGAUGGGUU
Protein: M
```

Translation: Summary

RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGUUGAUGGGUU

Protein: MA

Translation: Summary

RNA: GAGCUGAUGGCU**ACU**ACACAUAUUGCCAGUUGAUGGGUU

Protein: MA**T**

Translation: Summary

RNA: GAGCUGAUGGCUACU**ACA**CAUAUUGCCAGUUGAUGGGUU

Protein: MAT**T**

Translation: Summary

RNA: GAGCUGAUGGCUACUACA**CAU**AUUGCCAGUUGAUGGGUU

Protein: MATT**H**

Translation: Summary

RNA: GAGCUGAUGGCUACUACACAU**AUU**GCCAGUUGAUGGGUU

Protein: MATTH**I**

Translation: Summary

```
    RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGUUGAUGGGUU
Protein: MATTHIA
```

Translation: Summary

```
    RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGUUGAUGGGUU
Protein: MATTHIAS
```

Translation: Summary

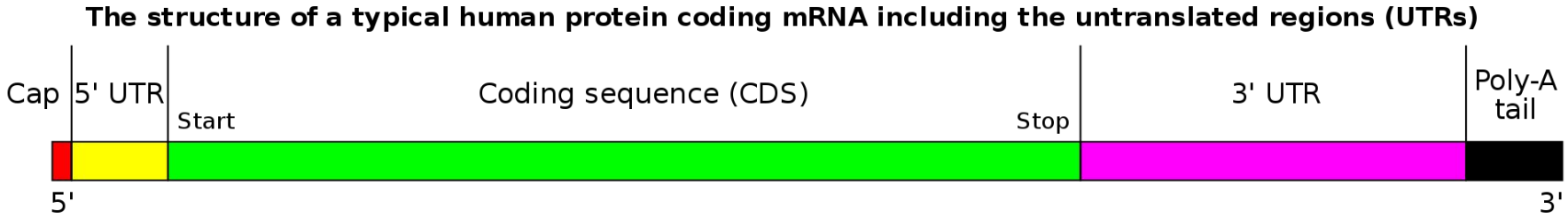
RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGU**UGA**UGGGUU

Protein: MATTHIAS

Translation: Summary

RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGUUGAUGGGUU

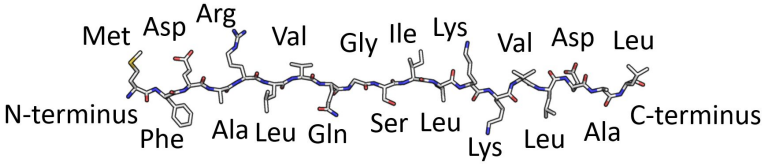
Protein: MATTHIAS



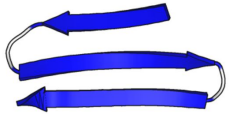
Protein Structure

- A protein's function is largely based on its structure

Primary



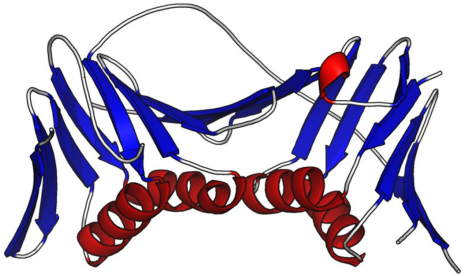
Secondary



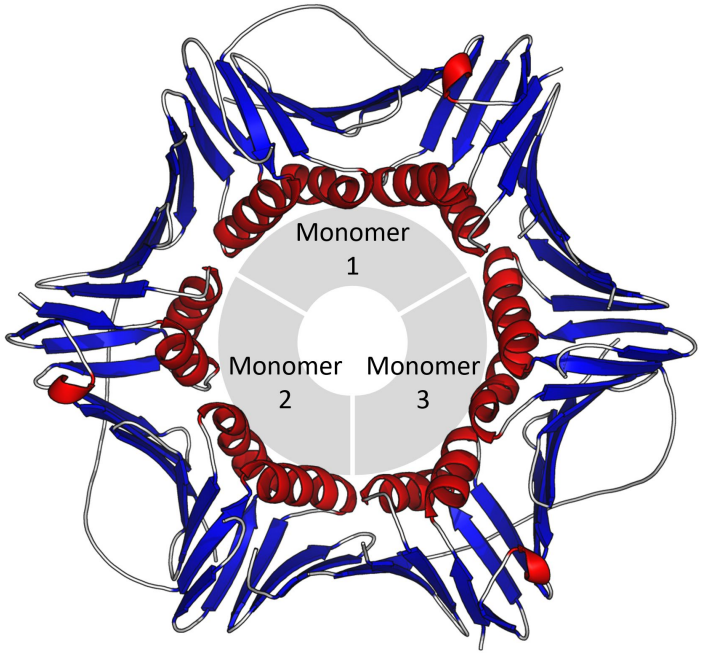
β-Sheet (3 strands)

α-helix

Tertiary



Quaternary



The Central Dogma: Summary

DNA: GAGCTGATGGCTACTACACATATTGCCAGTTGATGGGTT

Transcription

RNA: GAGCUGAUGGCUACUACACAUAUUGCCAGUUGAUGGGUU

Translation

Protein: MATTHIAS

Natural Selection

Natural Selection

- There is always natural variance (both “genotypic” and “phenotypic”) in a population of a given species

Natural Selection

- There is always natural variance (both “genotypic” and “phenotypic”) in a population of a given species
- Natural Selection: Traits that “improve the fitness” of an organism will cause that organism to be more likely to reproduce

Natural Selection

- There is always natural variance (both “genotypic” and “phenotypic”) in a population of a given species
- Natural Selection: Traits that “improve the fitness” of an organism will cause that organism to be more likely to reproduce
    - Traits that are “heritable” pass down to its offspring

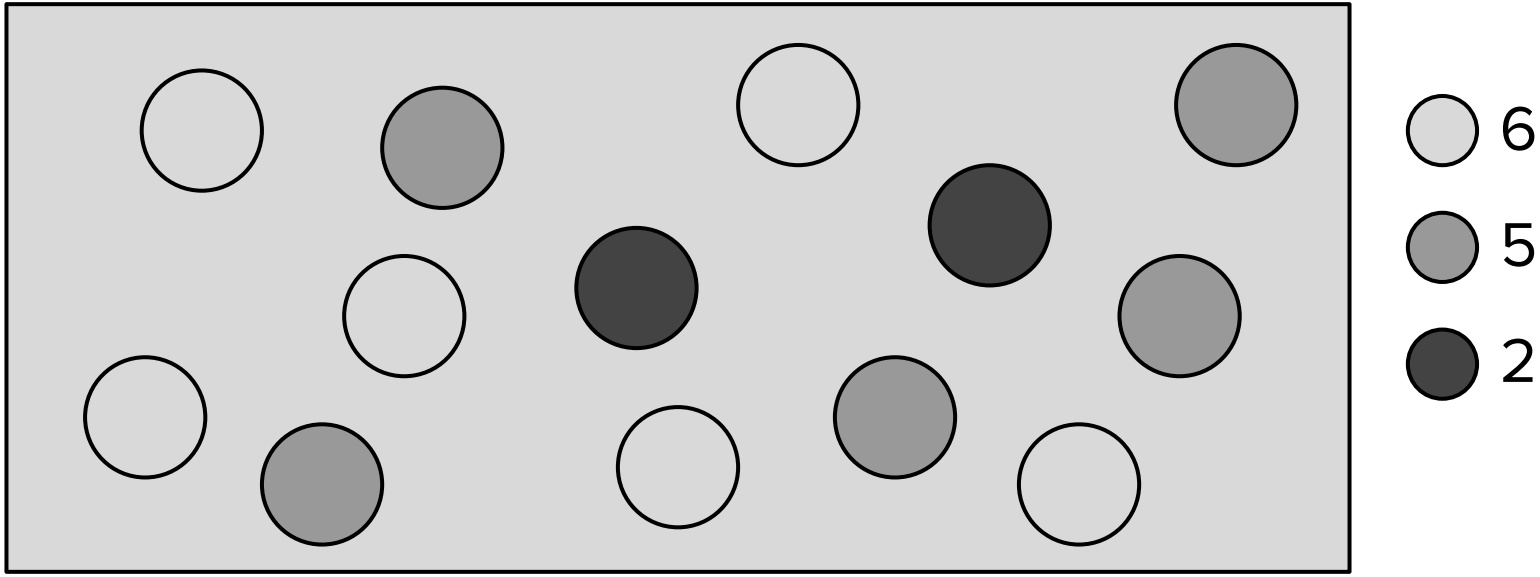
Natural Selection

- There is always natural variance (both “genotypic” and “phenotypic”) in a population of a given species
- Natural Selection: Traits that “improve the fitness” of an organism will cause that organism to be more likely to reproduce
  - Traits that are “heritable” pass down to its offspring
  - Individuals without this trait are less likely to reproduce

Natural Selection

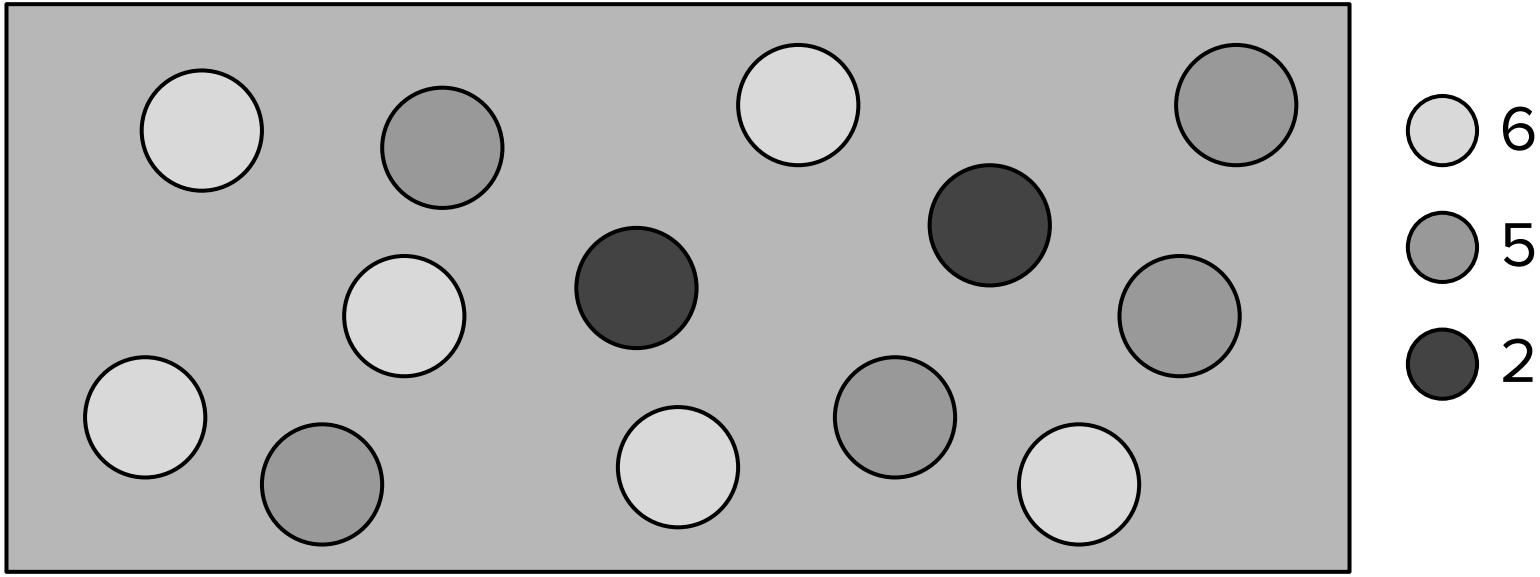
- There is always natural variance (both “genotypic” and “phenotypic”) in a population of a given species
- Natural Selection: Traits that “improve the fitness” of an organism will cause that organism to be more likely to reproduce
  - Traits that are “heritable” pass down to its offspring
  - Individuals without this trait are less likely to reproduce
  - In the next generation, a larger portion of the population will have the trait

Natural Selection: Example



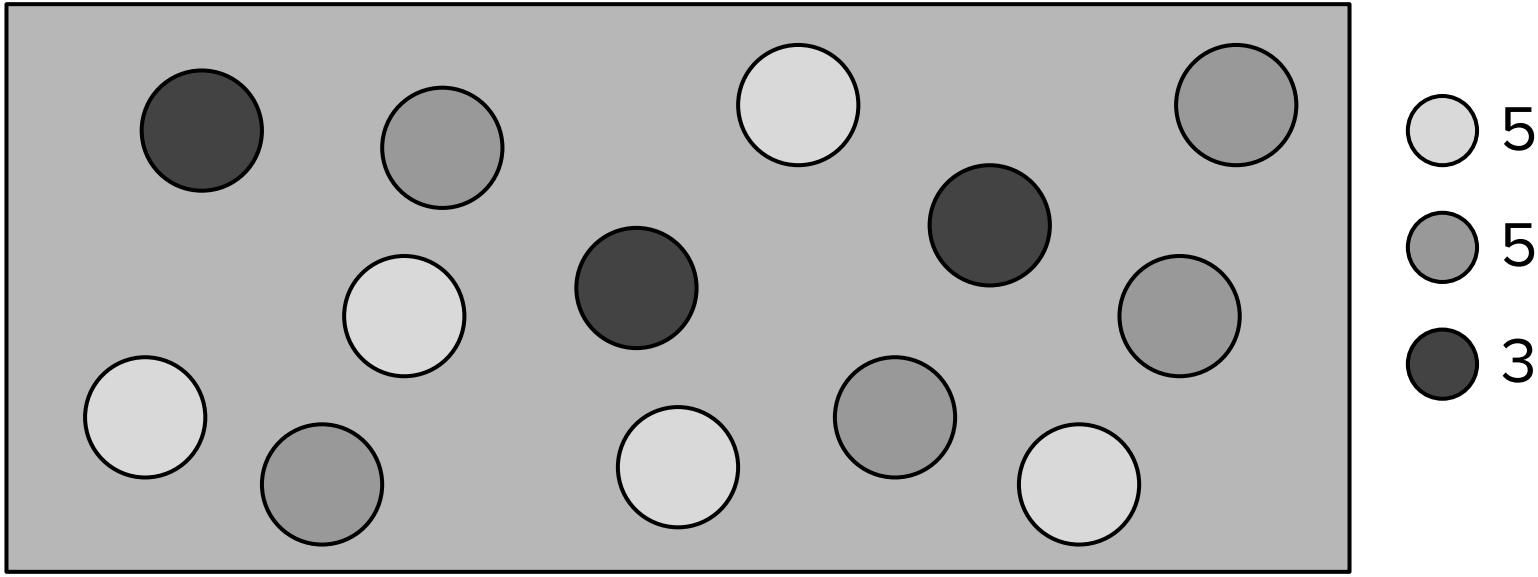
Generation 0

Natural Selection: Example



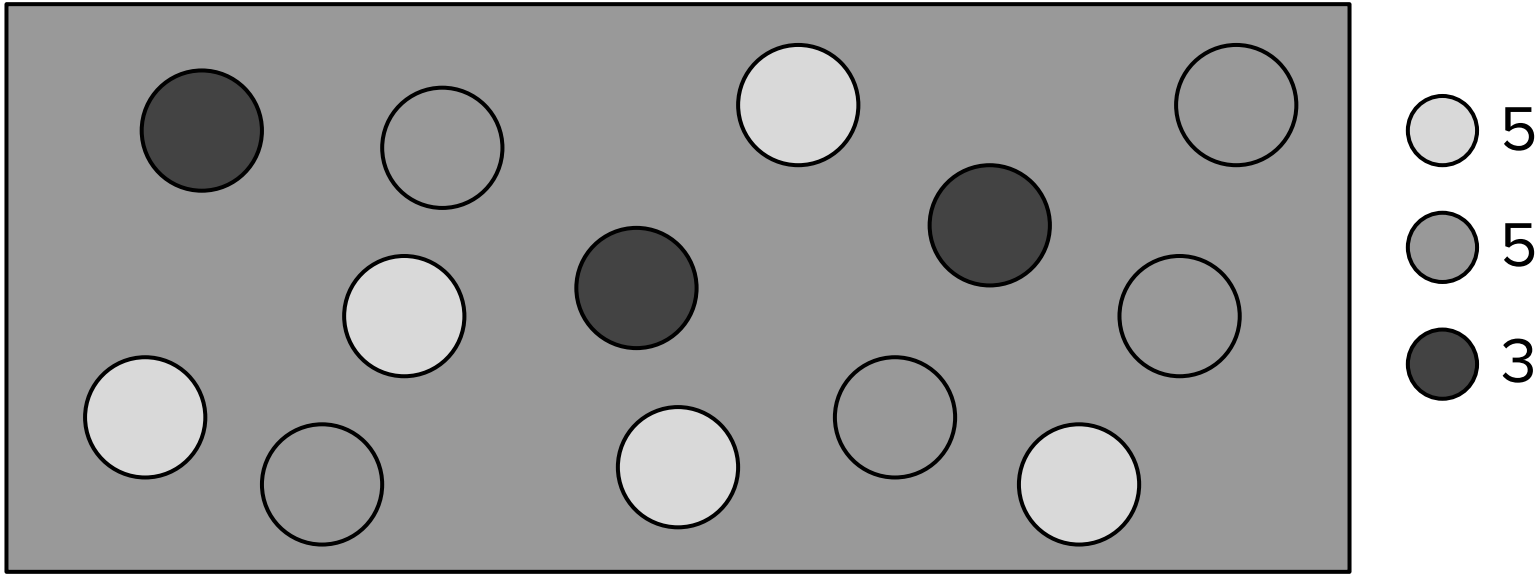
Generation 0

Natural Selection: Example



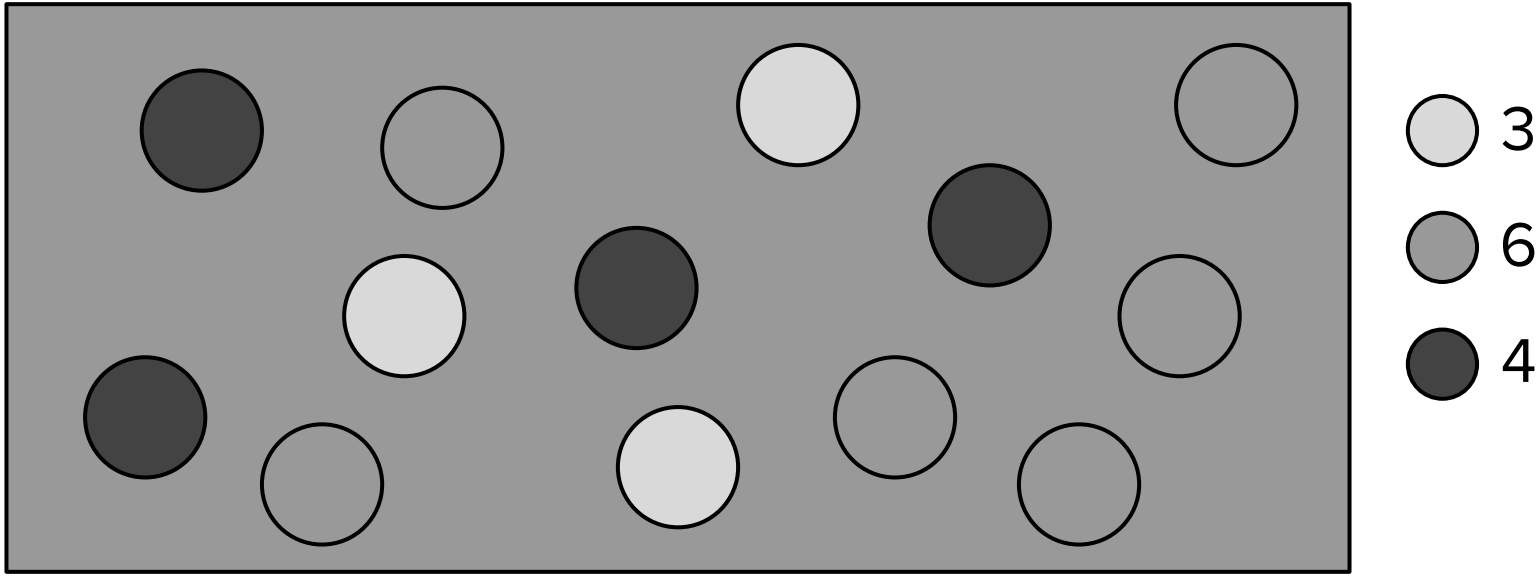
Generation 1

Natural Selection: Example



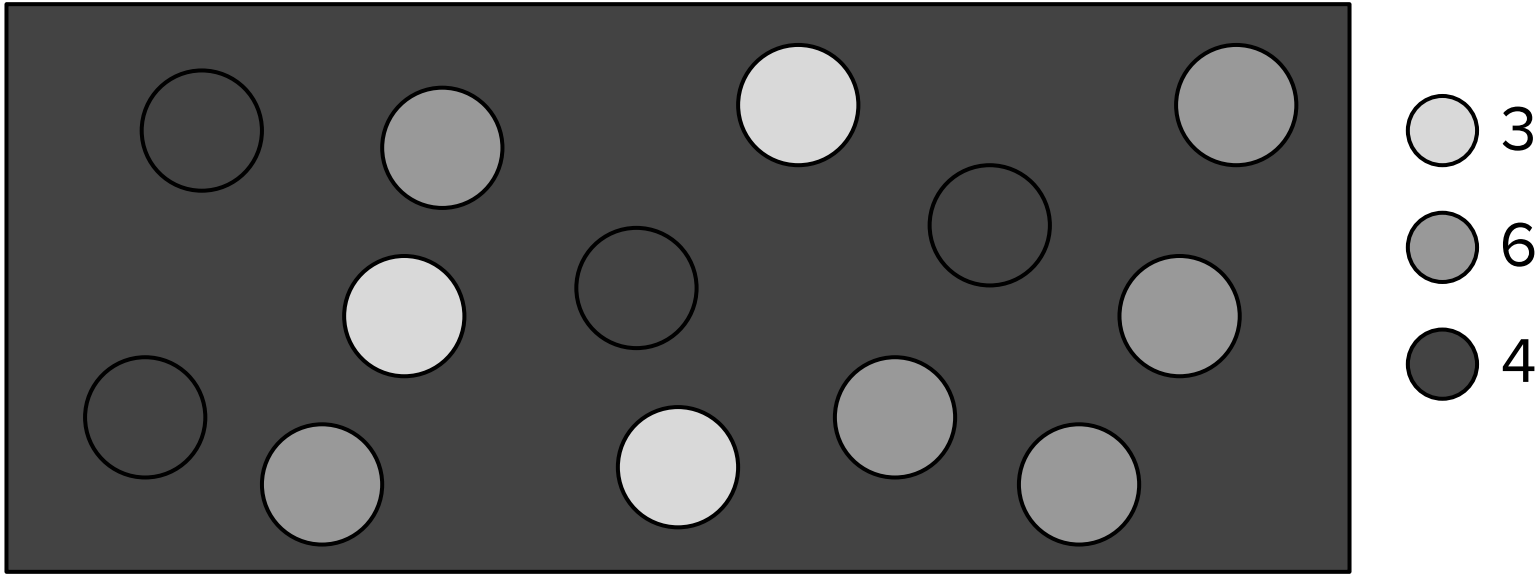
Generation 1

Natural Selection: Example



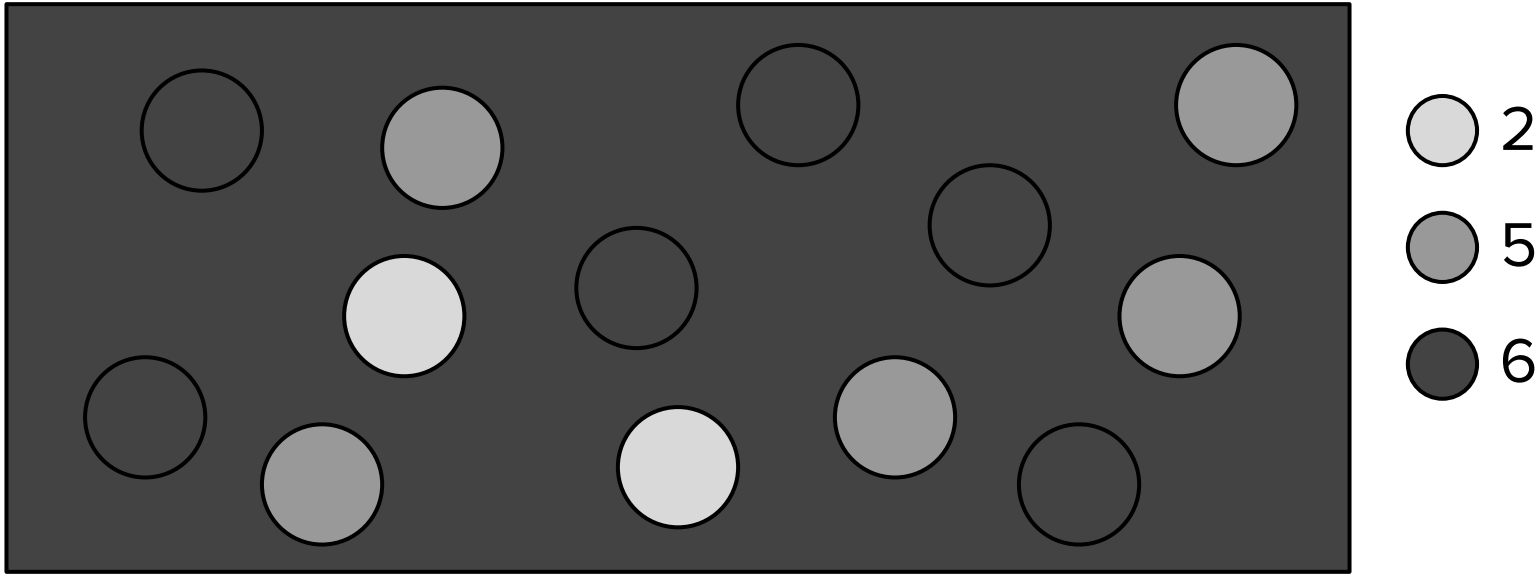
Generation 2

Natural Selection: Example



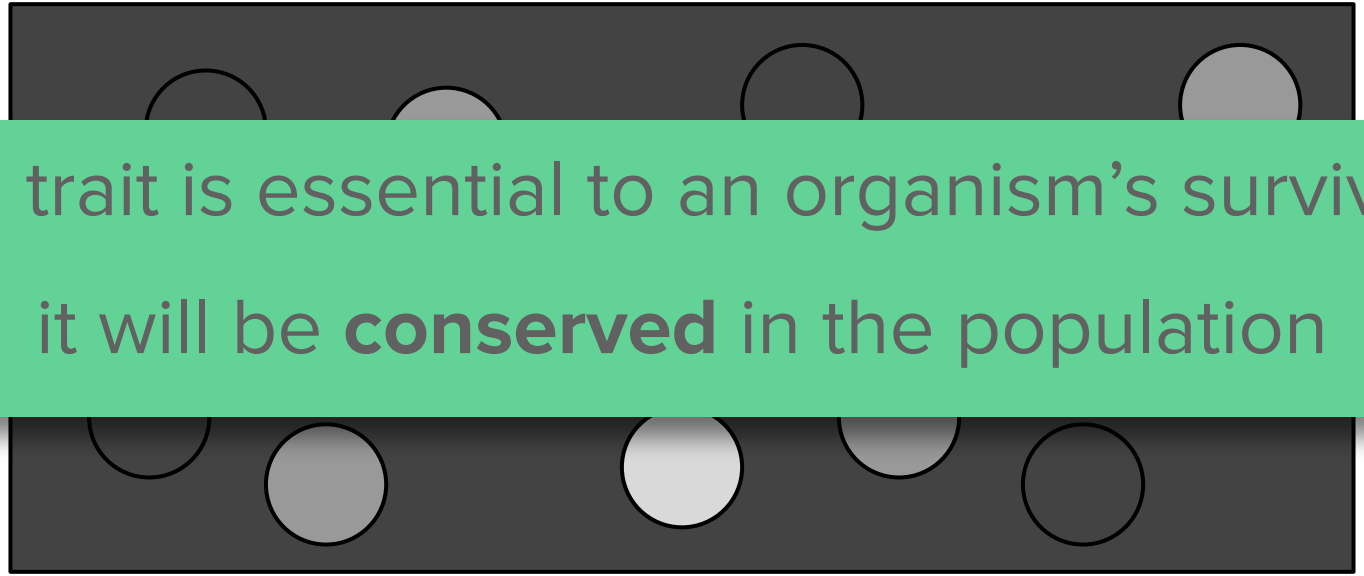
Generation 2

Natural Selection: Example



Generation 3

Natural Selection: Example



If a trait is essential to an organism's survival,

it will be **conserved** in the population

Generation 3

Sequence Alignment

Pairwise Sequence Alignment

- General Idea: If I have two strings *s* and *t*, if I were to stick gaps in either string, could I make them line up better?

Pairwise Sequence Alignment

- General Idea: If I have two strings *s* and *t*, if I were to stick gaps in either string, could I make them line up better?

AGTACGTACGT
ACGTACGTAAT

Pairwise Sequence Alignment

- General Idea: If I have two strings *s* and *t*, if I were to stick gaps in either string, could I make them line up better?

```
A-GTACGTACGT
ACGTACGTAA-T
```

Pairwise Sequence Alignment

- General Idea: If I have two strings *s* and *t*, if I were to stick gaps in either string, could I make them line up better?

```
A-GTACGTACGT
ACGTACGTAA-T
```

Pairwise Sequence Alignment

- General Idea: If I have two strings *s* and *t*, if I were to stick gaps in either string, could I make them line up better?

- Biological Motivation: Align an important gene in human and its "ortholog" (equivalent) in mouse to see which parts are conserved

Pairwise Sequence Alignment: Scoring Function

Given an **alignment**, a **gap penalty** $\boldsymbol{\sigma}$, and a **scoring matrix** $\boldsymbol{M}$, let the **score** of the alignment be defined as the **sum** of the scores of each position of the alignment, where a position is scored $\boldsymbol{\sigma}$ if either sequence has a **gap**, else $\boldsymbol{M(c,c')}$ where c is the symbol at the position in one sequence and c' is the symbol at the position in the other sequence

Pairwise Sequence Alignment: Scoring Function

```
A-GTACGTACGT
ACGTACGTAA-T
```

Score: 0

	A	C	G	T
A	+1	-1	-1	-1
C	-1	+1	-1	-1
G	-1	-1	+1	-1
T	-1	-1	-1	+1

$\sigma = -1$

Pairwise Sequence Alignment: Scoring Function

```
A-GTACGTACGT
ACGTACGTAA-T
```

Score: 1

	A	C	G	T
A	+1	-1	-1	-1
C	-1	+1	-1	-1
G	-1	-1	+1	-1
T	-1	-1	-1	+1

$\sigma = -1$

Pairwise Sequence Alignment: Scoring Function

```
A-GTACGTACGT
ACGTACGTAA-T
```

Score: 0

	A	C	G	T
A	+1	-1	-1	-1
C	-1	+1	-1	-1
G	-1	-1	+1	-1
T	-1	-1	-1	+1

$\sigma = -1$

Pairwise Sequence Alignment: Scoring Function

```
A-GTACGTACGT
ACGTACGTAA-T
```

Score: 1

	A	C	G	T
A	+1	-1	-1	-1
C	-1	+1	-1	-1
G	-1	-1	+1	-1
T	-1	-1	-1	+1

$\sigma = -1$

Pairwise Sequence Alignment: Scoring Function

```
A-GTACGTACGT
ACGTACGTAA-T
```

Score: 2

	A	C	G	T
A	+1	-1	-1	-1
C	-1	+1	-1	-1
G	-1	-1	+1	-1
T	-1	-1	-1	+1

$\sigma = -1$

Pairwise Sequence Alignment: Scoring Function

```
A-GTACGTACGT
ACGTACGTAA-T
```

Score: 3

	A	C	G	T
A	+1	-1	-1	-1
C	-1	+1	-1	-1
G	-1	-1	+1	-1
T	-1	-1	-1	+1

$\sigma = -1$

Pairwise Sequence Alignment: Scoring Function

```
A-GTACGTACGT
ACGTACGTAA-T
```

Score: 4

	A	C	G	T
A	+1	-1	-1	-1
C	-1	+1	-1	-1
G	-1	-1	+1	-1
T	-1	-1	-1	+1

$\sigma = -1$

Pairwise Sequence Alignment: Scoring Function

```
A-GTACGTACGT
ACGTACGTAA-T
```

Score: 5

	A	C	G	T
A	+1	-1	-1	-1
C	-1	+1	-1	-1
G	-1	-1	+1	-1
T	-1	-1	-1	+1

$\sigma = -1$

Pairwise Sequence Alignment: Scoring Function

```
A-GTACGTACGT
ACGTACGTAA-T
```

Score: 6

	A	C	G	T
A	+1	-1	-1	-1
C	-1	+1	-1	-1
G	-1	-1	+1	-1
T	-1	-1	-1	+1

$\sigma = -1$

Pairwise Sequence Alignment: Scoring Function

```
A-GTACGTACGT
ACGTACGTAA-T
```

Score: 7

	A	C	G	T
A	+1	-1	-1	-1
C	-1	+1	-1	-1
G	-1	-1	+1	-1
T	-1	-1	-1	+1

$\sigma = -1$

Pairwise Sequence Alignment: Scoring Function

```
A-GTACGTACGT
ACGTACGTAA-T
```

Score: 6

	A	C	G	T
A	+1	-1	-1	-1
C	-1	+1	-1	-1
G	-1	-1	+1	-1
T	-1	-1	-1	+1

$\sigma = -1$

Pairwise Sequence Alignment: Scoring Function

```
A-GTACGTACGT
ACGTACGTAA-T
```

Score: 5

	A	C	G	T
A	+1	-1	-1	-1
C	-1	+1	-1	-1
G	-1	-1	+1	-1
T	-1	-1	-1	+1

$\sigma = -1$

Pairwise Sequence Alignment: Scoring Function

```
A-GTACGTACGT
ACGTACGTAA-T
```

Score: 6

	A	C	G	T
A	+1	-1	-1	-1
C	-1	+1	-1	-1
G	-1	-1	+1	-1
T	-1	-1	-1	+1

$\sigma = -1$

Pairwise Sequence Alignment: Scoring Function

```
A-GTACGTACGT
ACGTACGTAA-T
```

Score: 6

	A	C	G	T
A	+1	-1	-1	-1
C	-1	+1	-1	-1
G	-1	-1	+1	-1
T	-1	-1	-1	+1

$\sigma = -1$

Pairwise Sequence Alignment: Scoring Function

A-GTACGTACGT

We want to **maximize** this scoring function

Score: 6

	A	C	G	T
A	+1	-1	-1	-1
G	-1	-1	+1	-1
T	-1	-1	-1	+1

$\sigma = -1$

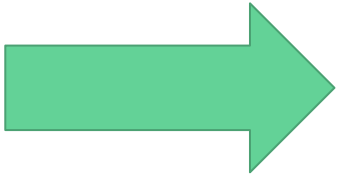
The Global Alignment Problem

Given two strings ***s*** and ***t***, a gap penalty ***σ***, and a scoring matrix ***M***, return a **maximum-scoring** alignment of *s* and *t*

The Global Alignment Problem

Given two strings ***s*** and ***t***, a gap penalty ***σ***, and a scoring matrix ***M***, return a **maximum-scoring** alignment of *s* and *t*

```
AGTACGTACGT
ACGTACGTAAT
```



```
A-GTACGTACGT
ACGTACGTAA-T
```

The Local Alignment Problem

Given two strings ***s*** and ***t***, a gap penalty ***σ***, and a scoring matrix ***M***,
return a **maximum-scoring** alignment of
a *substring* of *s* and a *substring* of *t*

The Local Alignment Problem

Given two strings ***s*** and ***t***, a gap penalty ***σ***, and a scoring matrix ***M***,
return a **maximum-scoring** alignment of
a *substring* of *s* and a *substring* of *t*

AGTACGTACGT
ACGTACGTAAT

GTACGTA
GTACGTA

The Multiple Sequence Alignment Problem

Given **multiple strings**, a gap penalty $\boldsymbol{\sigma}$, and a scoring matrix $\boldsymbol{M}$, return a **maximum-scoring** alignment of the strings

The Multiple Sequence Alignment Problem

Given **multiple strings**, a gap penalty ***σ***, and a scoring matrix ***M***, return a **maximum-scoring** alignment of the strings

```
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
    RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE
  RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD
Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD
  RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE
  RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK
  RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS
  RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE
  RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN
  RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE
  RLA0_METAC ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ETIP
  RLA0_METMA ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ESIP
```

Variant Calling

Variant Calling

- Any two humans have genomes that are roughly 99.9% identical

Variant Calling

- Any two humans have genomes that are roughly 99.9% identical

- Single Nucleotide Variants (SNVs)

ACATACGTACGT
ACGTACGTACGT
ACGTACGTACGT
ACATACGTTCGT
ACGTACGTACGT
ACGTACGTACGT
ACATACGTACGT
ACGTACGTACGT
ACGTACGTTCGT

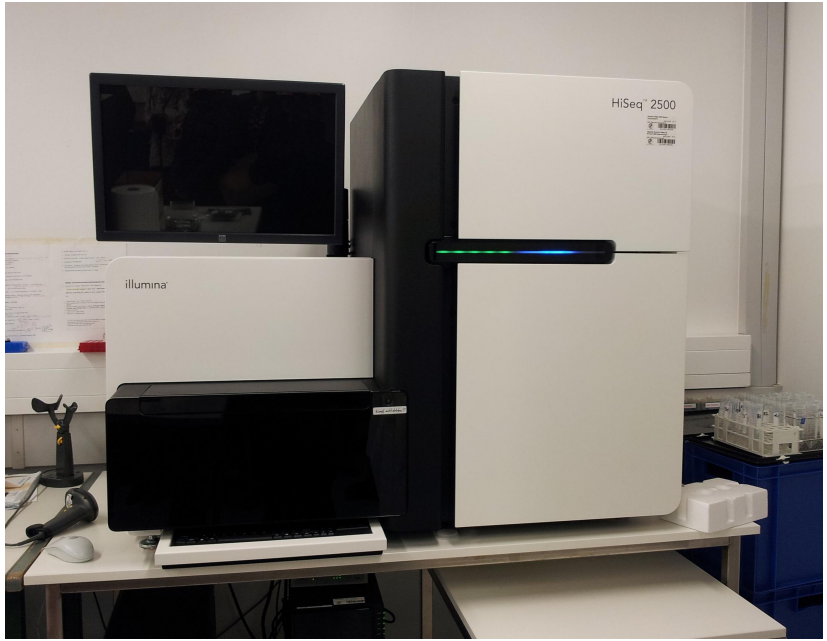
Variant Calling

- Any two humans have genomes that are roughly 99.9% identical
- Single Nucleotide Variants (SNVs)
- Structural Variants (SVs)

ACAGCAGCAGCAGTT
ACAGCAGTT
ACAGTT
ACAGCAGCAGTT

SNV Calling: General Approach

- Sequence the DNA of the individual



SNV Calling: General Approach

- Sequence the DNA of the individual
- Align the reads to the reference genome

SNV Calling: General Approach

- Sequence the DNA of the individual
- Align the reads to the reference genome
- For each site in the genome, predict the genotype based on the reads

```
          ACTTACGT
             GTACGTAC
        TACGTACG
           CTTACGTA
     CGTACTTA
REF: ...ACGTACGTACGTACGTACGTACGT...
```

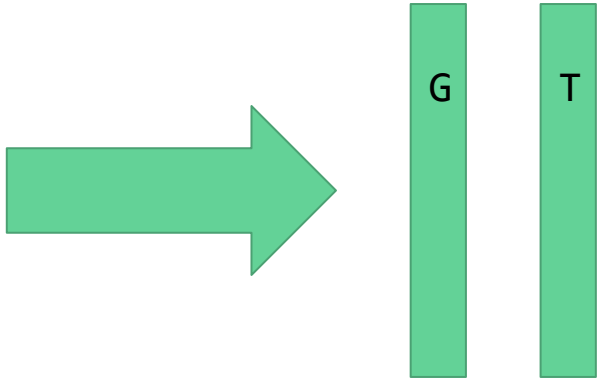
SNV Calling: General Approach

- Sequence the DNA of the individual
- Align the reads to the reference genome
- For each site in the genome, predict the genotype based on the reads

```
ACTTACGT
   GTACGTAC
TACGTACG
  CTTACGTA
CGTACTTA
REF: ...ACGTACGTACGTACGTACGTACGT...
```

50% G

50% T



SNV Calling: Challenges

- Some regions of the genome are difficult to sequence

SNV Calling: Challenges

- Some regions of the genome are difficult to sequence
- Sequencing technologies have sequencing error

SNV Calling: Challenges

- Some regions of the genome are difficult to sequence
- Sequencing technologies have sequencing error
- Sequencing technologies have sampling error

Population Genetics

- Once we’ve called SNVs and SVs in enough people, what can we do?

Population Genetics

- Once we've called SNVs and SVs in enough people, what can we do?
    - Genome-Wide Association Studies (GWAS)

Population Genetics

- Once we've called SNVs and SVs in enough people, what can we do?
  - Genome-Wide Association Studies (GWAS)
  - Genetic Ancestry/Admixture

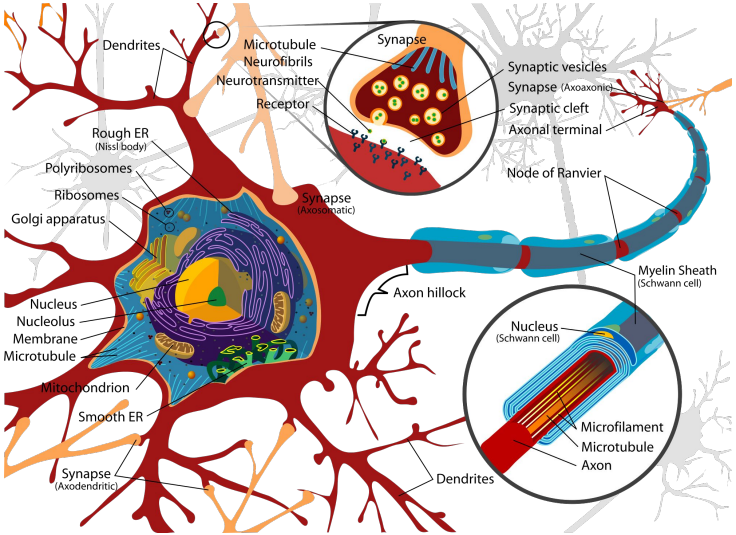
Population Genetics

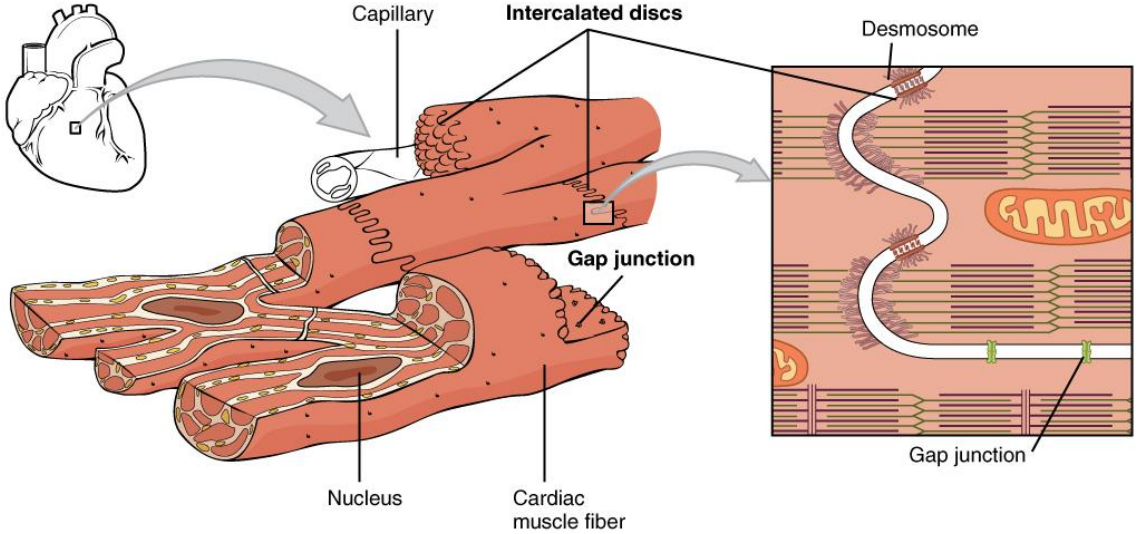
- Once we've called SNVs and SVs in enough people, what can we do?
  - Genome-Wide Association Studies (GWAS)
  - Genetic Ancestry/Admixture
  - Genetic Counseling

Differential Expression Analysis

Differential Expression Analysis: RNA-Seq

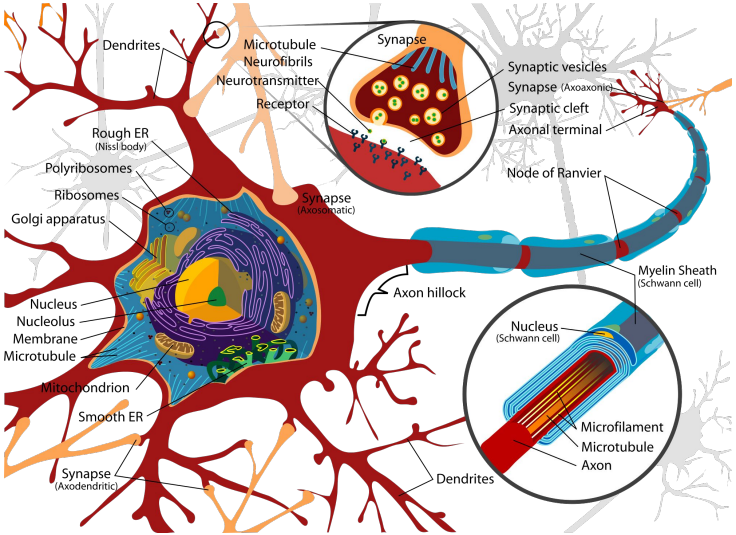
- All cells in the body have (roughly) identical genomes

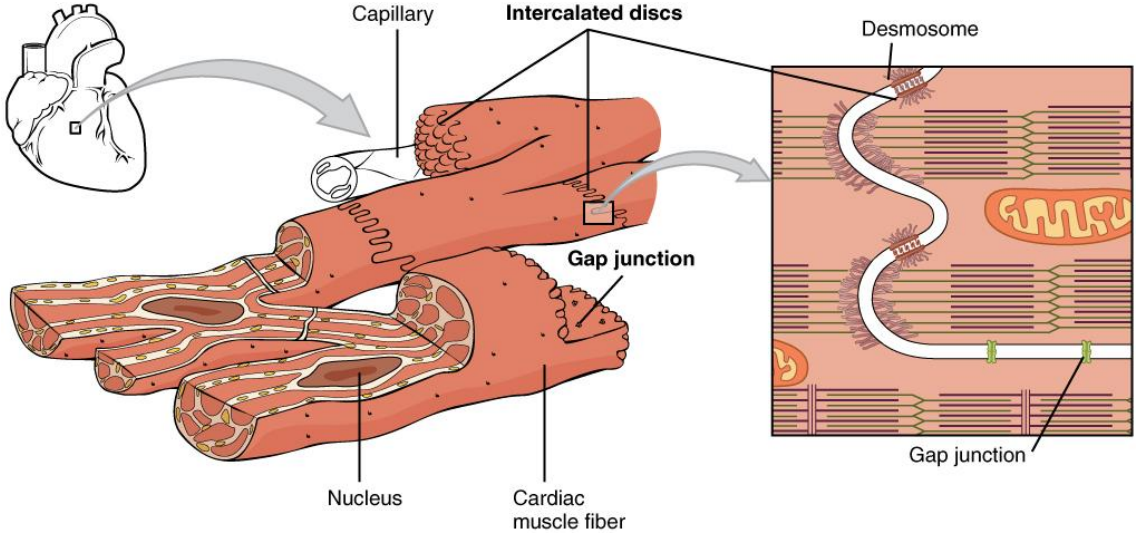




Differential Expression Analysis: RNA-Seq

- All cells in the body have (roughly) identical genomes
  - Differences in how they look/function are caused by “differential expression” of genes



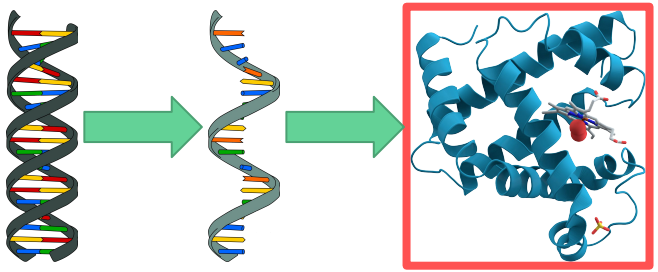


Differential Expression Analysis: RNA-Seq

- All cells in the body have (roughly) identical genomes
  - Differences in how they look/function are caused by “differential expression” of genes
- Biological Question: Given two different samples, what genes are differentially expressed across them?

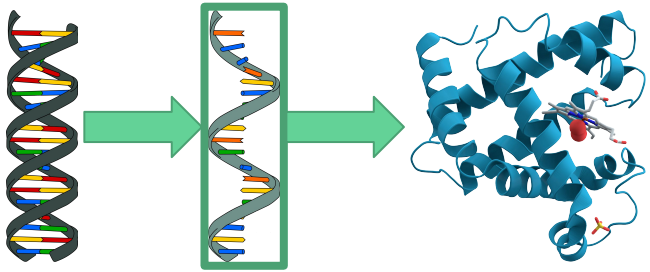
Differential Expression Analysis: RNA-Seq

- All cells in the body have (roughly) identical genomes
  - Differences in how they look/function are caused by “differential expression” of genes
- <u>Biological Question</u>: Given two different samples, what genes are differentially expressed across them?
  - We want to measure protein levels, but we can’t in high-throughput



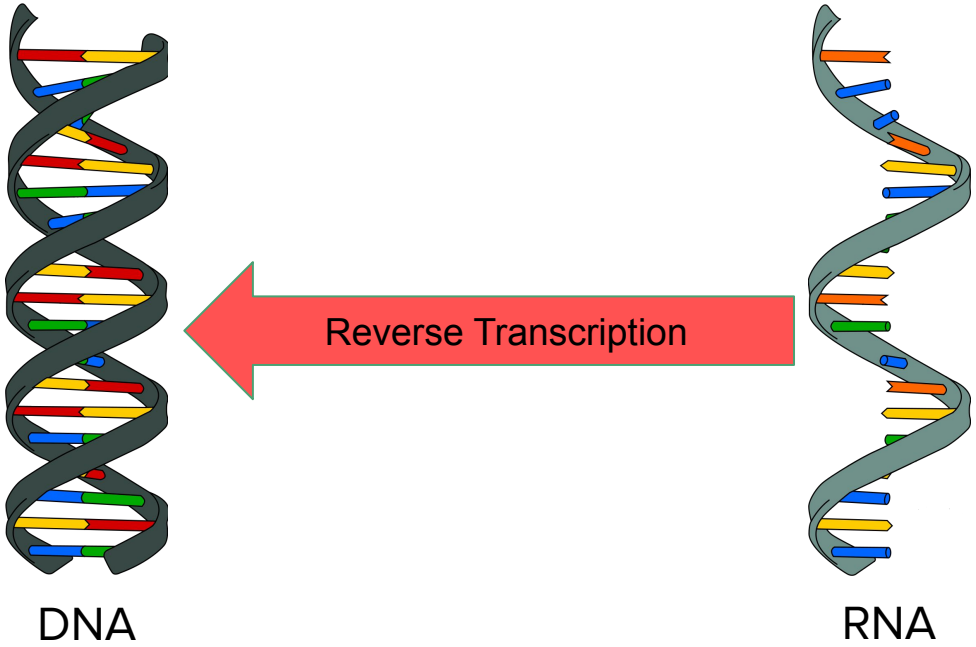
Differential Expression Analysis: RNA-Seq

- All cells in the body have (roughly) identical genomes
  - Differences in how they look/function are caused by “differential expression” of genes
- <u>Biological Question</u>: Given two different samples, what genes are differentially expressed across them?
  - We want to measure protein levels, but we can’t in high-throughput
  - Instead, we measure RNA levels



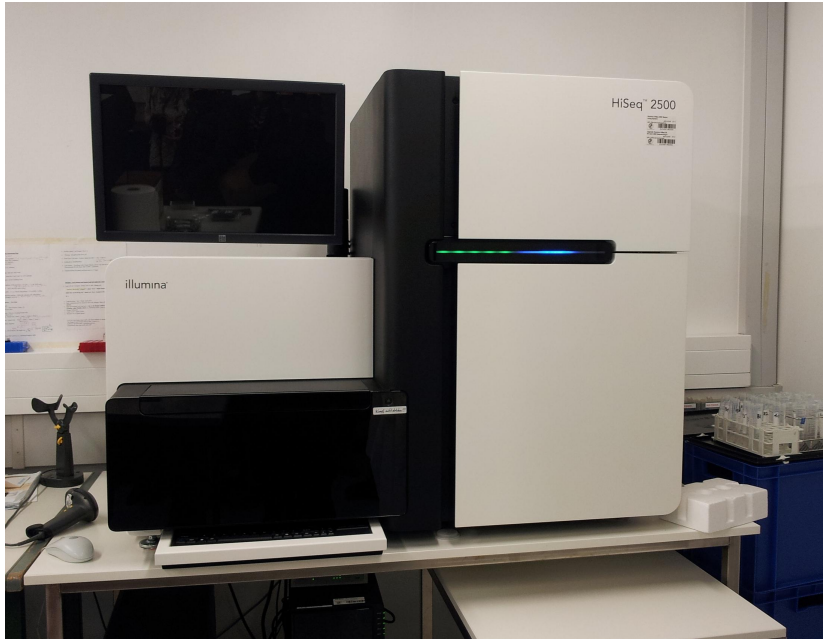
SNV Calling: General Approach

- Reverse Transcribe RNA to DNA



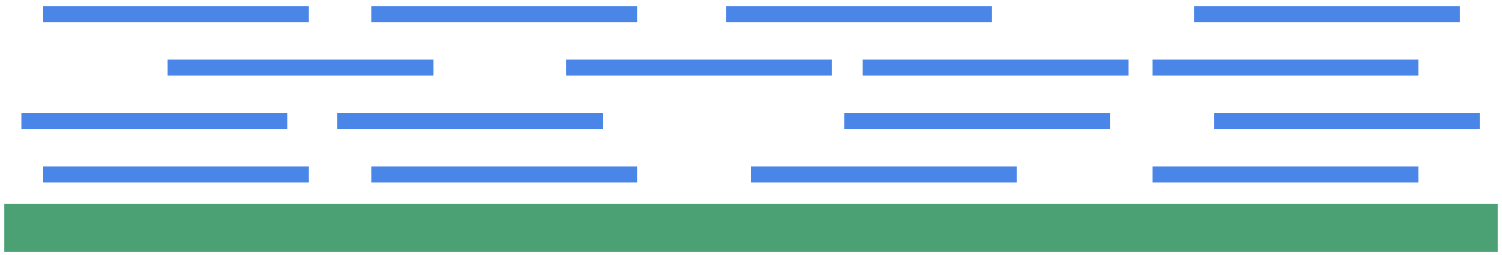
SNV Calling: General Approach

- Reverse Transcribe RNA to DNA
- Sequence the resulting DNA



SNV Calling: General Approach

- Reverse Transcribe RNA to DNA
- Sequence the resulting DNA
- Align the reads to the reference genome



SNV Calling: General Approach

Gene	Sample 1 Count	Sample 2 Count
A	###	###
B	###	###
C	###	###

- Reverse Transcribe RNA to DNA
- Sequence the resulting DNA
- Align the reads to the reference genome
- Count the number of reads that mapped to each gene

SNV Calling: General Approach

Gene	Sample 1 FPKM	Sample 2 FPKM
A	###	###
B	###	###
C	###	###

- Reverse Transcribe RNA to DNA
- Sequence the resulting DNA
- Align the reads to the reference genome
- Count the number of reads that mapped to each gene
- Normalize by gene length and by sequencing depth

SNV Calling: General Approach

Gene	Sample 1 FPKM	Sample 2 FPKM	Log-2 Ratio	*p*
A	###	###	###	###
B	###	###	###	###
C	###	###	###	###

- Reverse Transcribe RNA to DNA
- Sequence the resulting DNA
- Align the reads to the reference genome
- Count the number of reads that mapped to each gene
- Normalize by gene length and by sequencing depth
- Perform differential expression statistical tests for each gene

Genome Assembly

Genome Assembly

- What is the genome sequence of a given organism?

...ATACAGTGGAACACCATCTG...

Genome Assembly

- What is the genome sequence of a given organism?
- We are able to sequence small fragments of an organism's genome

ATACAG
CAGTGG
GGAACA
CACCAT
CCATCT

Genome Assembly

- What is the genome sequence of a given organism?
- We are able to sequence small fragments of an organism's genome
- How do we tie these small fragments together into a single string?

```
ATACAG
    CAGTGG
          GGAACA
                CACCAT
                   CCATCT
...ATACAGTGGAACACCATCTG...
```

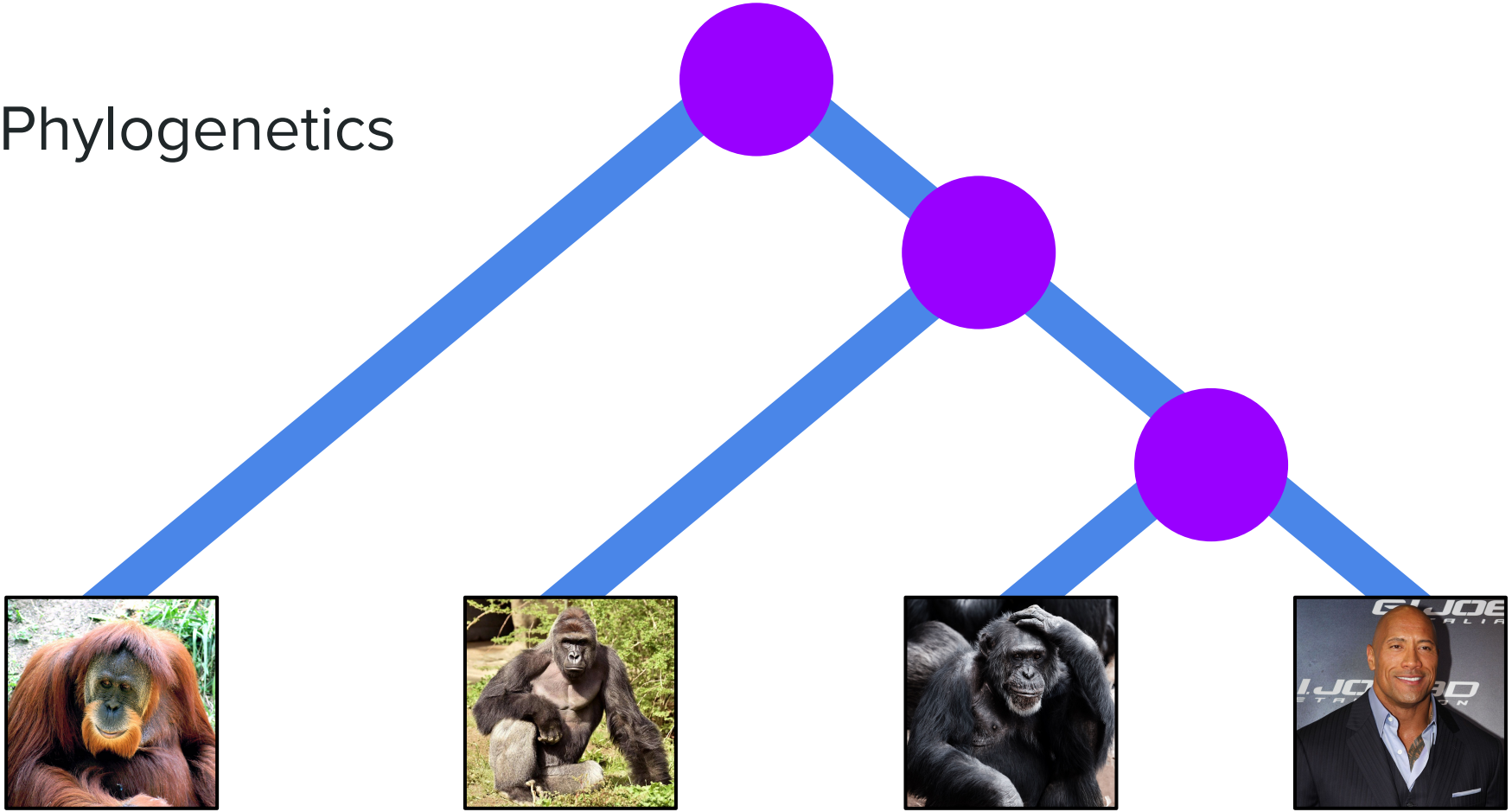
Genome Assembly

- What is the genome sequence of a given organism?
- We are able to sequence small fragments of an organism's genome
- How do we tie these small fragments together into a single string?

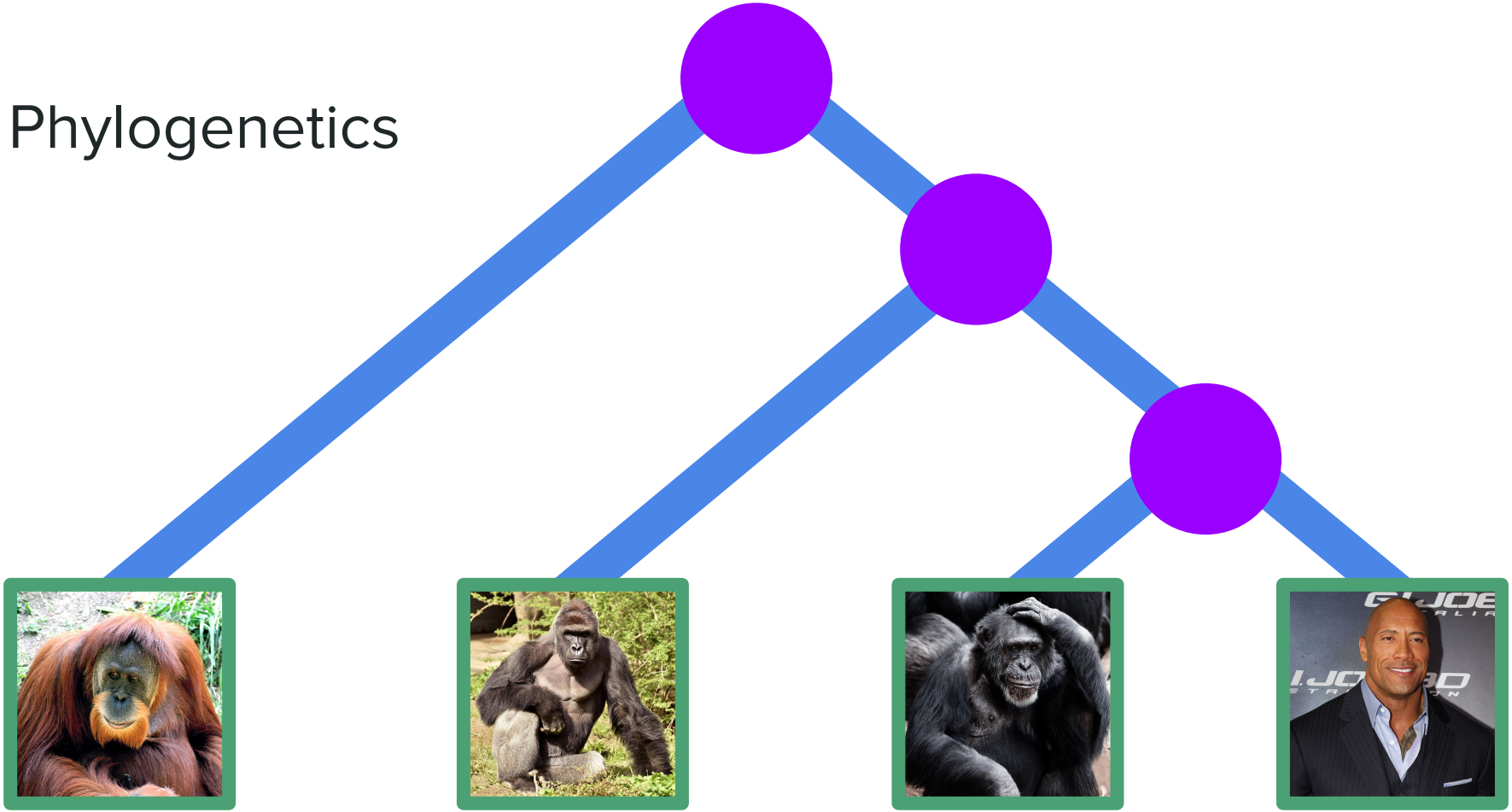
- <u>Computational Problem</u>: Given a list of strings *reads*, find the shortest superstring of *reads*

Phylogenetics

Phylogenetics

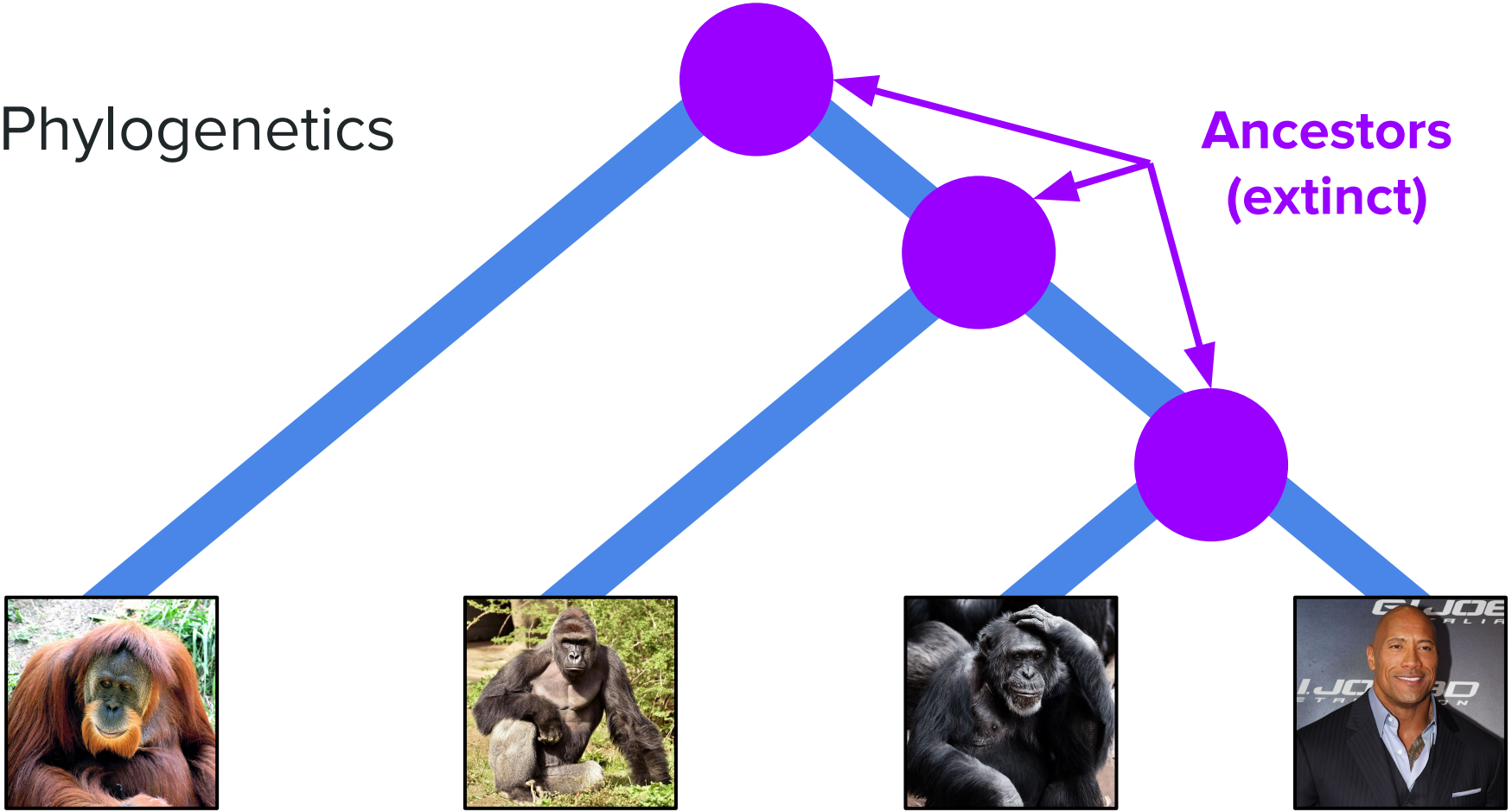


Phylogenetics

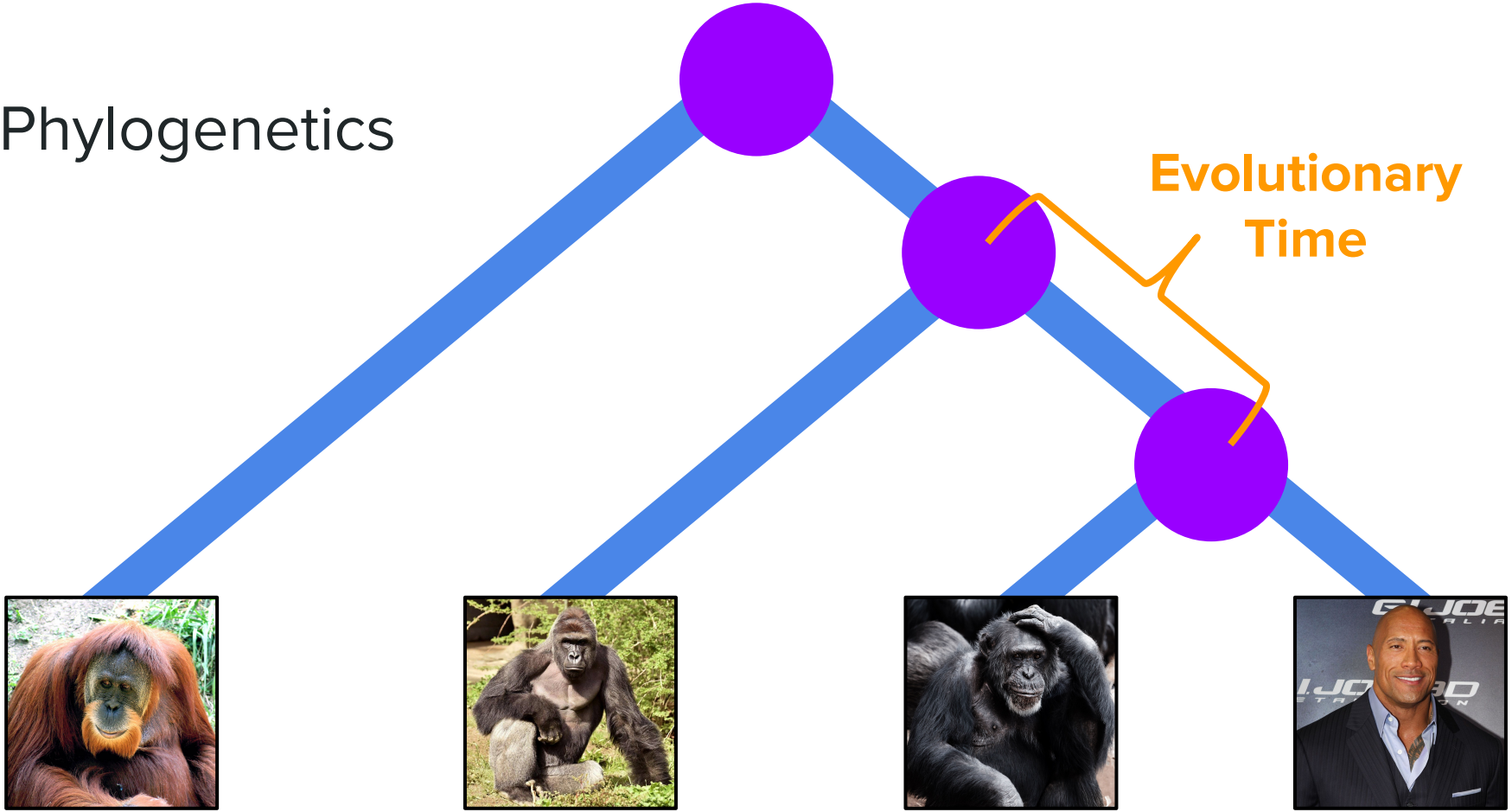


Present-Day Species

Phylogenetics



Phylogenetics



Models of Evolution

Models of Evolution

- Models of Tree Evolution: Describe a probability distribution over the shapes of the phylogenetic trees

Models of Evolution

- Models of Tree Evolution: Describe a probability distribution over the shapes of the phylogenetic trees
  - Are some tree topologies more likely to be observed?

Models of Evolution

- Models of Tree Evolution: Describe a probability distribution over the shapes of the phylogenetic trees
  - Are some tree topologies more likely to be observed?

- Models of Sequence Evolution: Describe a probability distribution over the observed sequences

Models of Evolution

- Models of Tree Evolution: Describe a probability distribution over the shapes of the phylogenetic trees
  - Are some tree topologies more likely to be observed?

- Models of Sequence Evolution: Describe a probability distribution over the observed sequences
  - Are some sequences more likely to be observed (e.g. fitness)?

Phylogenetic Inference

- Can we somehow reconstruct the evolutionary history of species based solely on their sequences?

Phylogenetic Inference

- Can we somehow reconstruct the evolutionary history of species based solely on their sequences?
  - Raw Sequences ➔ Multiple Sequence Alignment ➔ Tree

Phylogenetic Inference

- Can we somehow reconstruct the evolutionary history of species based solely on their sequences?
  - Raw Sequences ➔ Multiple Sequence Alignment ➔ Tree
- <u>Maximum Likelihood</u>: Given a **multiple sequence alignment** and a **model** of (sequence evolution), find the tree that maximizes the "likelihood function" (i.e., probability of observing the alignment given the tree)

Summary

Summary

- We started with some basic molecular biology review

Summary

- We started with some basic molecular biology review
- We then introduced multiple important biological problems and discussed their bioinformatics computational problem formulation

Summary

- We started with some basic molecular biology review
- We then introduced multiple important biological problems and discussed their bioinformatics computational problem formulation
- **Bioinformatics = BIG data!**

Summary

- We started with some basic molecular biology review
- We then introduced multiple important biological problems and discussed their bioinformatics computational problem formulation
- Bioinformatics = BIG data!
    - **We need efficient algorithms**

Summary

- We started with some basic molecular biology review
- We then introduced multiple important biological problems and discussed their bioinformatics computational problem formulation
- Bioinformatics = BIG data!
  - We need efficient algorithms
  - **We need optimized implementations of these algorithms**