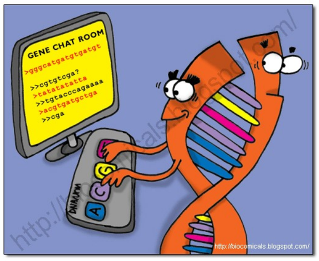
UCONN
UNIVERSITY OF CONNECTICUT

Introduction to Computational Biology & Bioinformatics – Part I

ENGR 1166 Biomedical Engineering

Computational biology

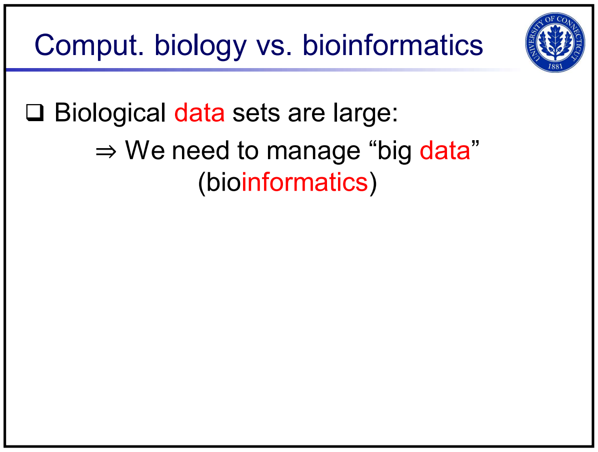
A marriage between biology and computer



http://biocomicals.blogspot.com/

Comput. biology vs. bioinformatics

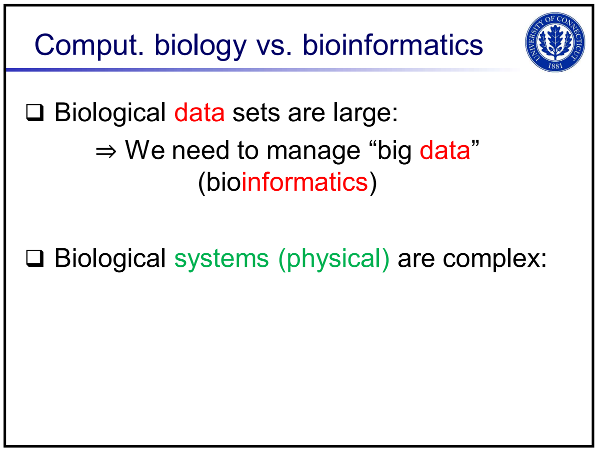
❑ Biological data sets are large:



Comput. biology vs. bioinformatics

- ❑ Biological data sets are large:

  ⇒ We need to manage “big data”
  (bioinformatics)

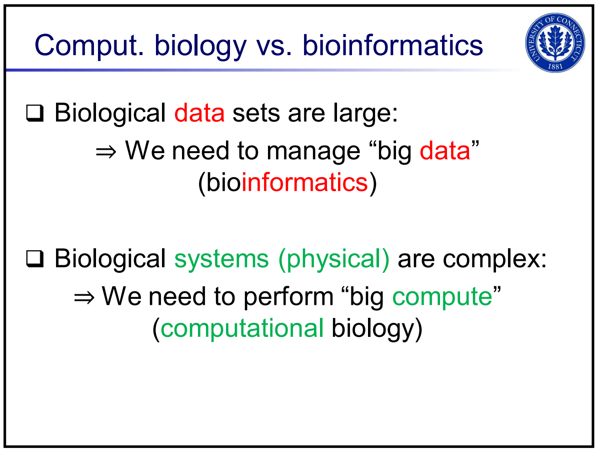


Comput. biology vs. bioinformatics

- ❑ Biological data sets are large:

  ⇒ We need to manage “big data”
  (bioinformatics)

- ❑ Biological systems (physical) are complex:



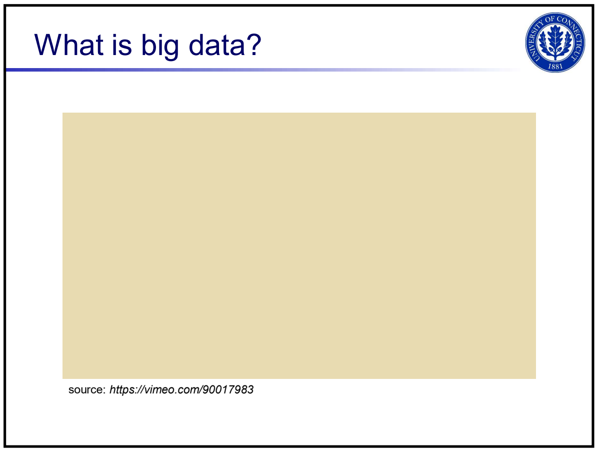
Comput. biology vs. bioinformatics

- ❑ Biological data sets are large:

  ⇒ We need to manage “big data”
  (bioinformatics)

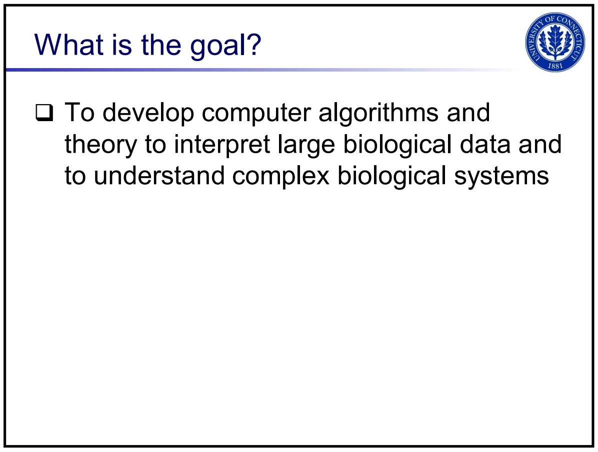
- ❑ Biological systems (physical) are complex:

  ⇒ We need to perform “big compute”
  (computational biology)



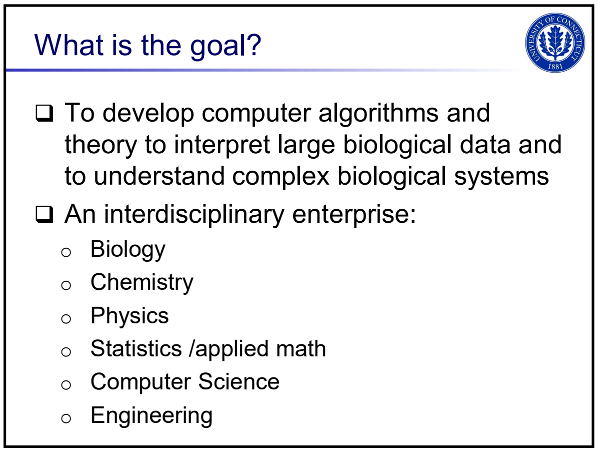
What is big data?

source: *https://vimeo.com/90017983*



What is the goal?

- ❑ To develop computer algorithms and theory to interpret large biological data and to understand complex biological systems



What is the goal?

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

Why do we need this expertise?

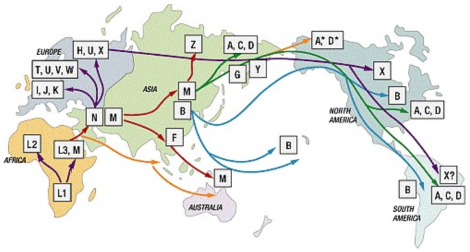
- ❑ Rapid explosion in our ability to acquire biological data
  - ○ ***How can we find robust patterns in these data?***

Why do we need this expertise?

- ❑ Rapid explosion in our ability to acquire biological data
  - ○ ***How can we find robust patterns in these data?***
- ❑ Recognition that biological phenomena are enormously complex and biological problems benefit from interdisciplinary approaches
  - ○ ***How can we understand, predict, and manipulate these systems?***

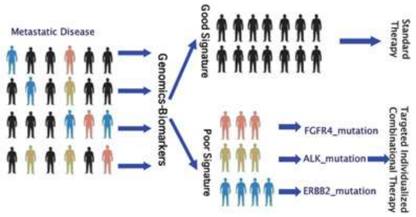
Applications

- ❑ **Origins of Humanity:** DNA sequencing is used to trace the most recent common maternal ancestor of all living humans



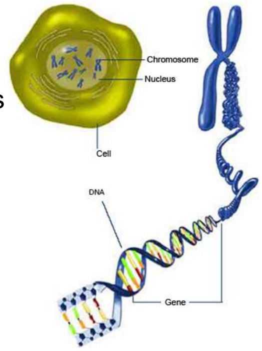
Applications

- □ **Personalized medicine:** Data about the subject’s own genome and mathematical models are used to tailor therapy to each individual



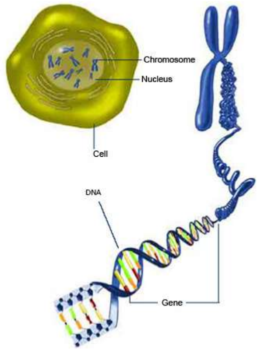
Key concepts

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



Key concepts

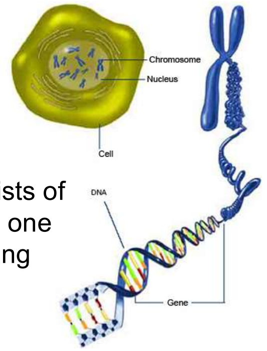
A **chromosome** is a continuous strands of DNA wrapped around a protein scaffold



Key concepts

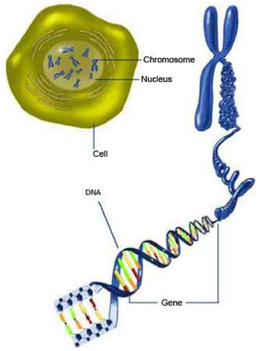
A **chromosome** is a continuous strands of DNA wrapped around a protein scaffold

The human genome consists of 23 pairs of chromosomes, one member of each pair coming from each parent



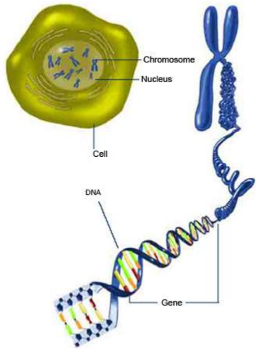
Key concepts

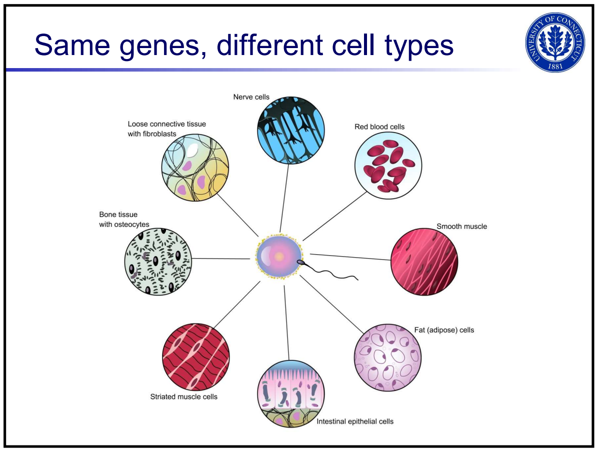
A **gene** is a segment of DNA that has information for making a specific type of protein



Key concepts

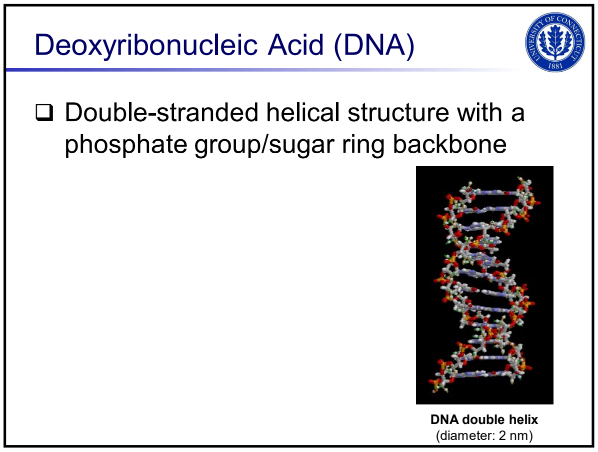
The **ribosome** is a large molecular machine that serves as the site where biological proteins are synthesized





Same genes, different cell types

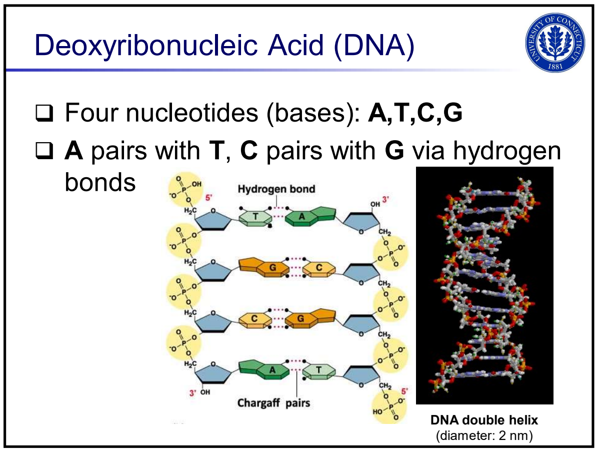
Nerve cells
Loose connective tissue with fibroblasts
Red blood cells
Bone tissue with osteocytes
Smooth muscle
Fat (adipose) cells
Striated muscle cells
Intestinal epithelial cells



Deoxyribonucleic Acid (DNA)

- ❑ Double-stranded helical structure with a phosphate group/sugar ring backbone

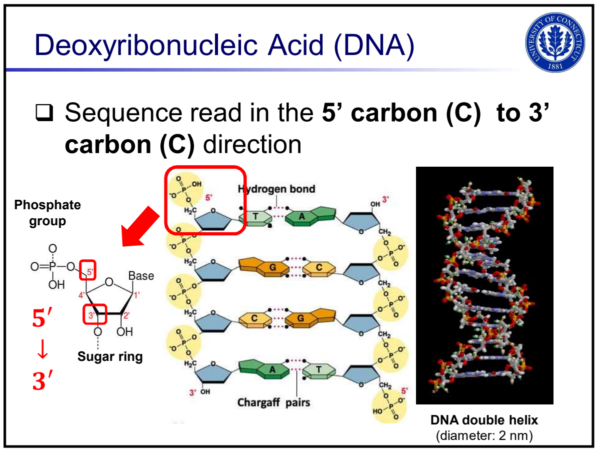
DNA double helix
(diameter: 2 nm)



Deoxyribonucleic Acid (DNA)

- ❑ Four nucleotides (bases): **A,T,C,G**
- ❑ **A** pairs with **T**, **C** pairs with **G** via hydrogen bonds

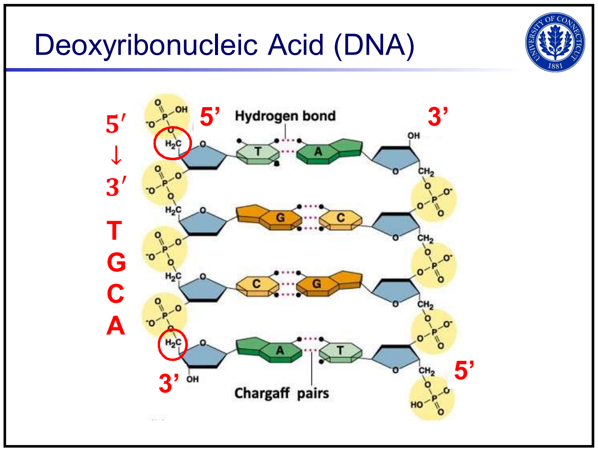
DNA double helix
(diameter: 2 nm)



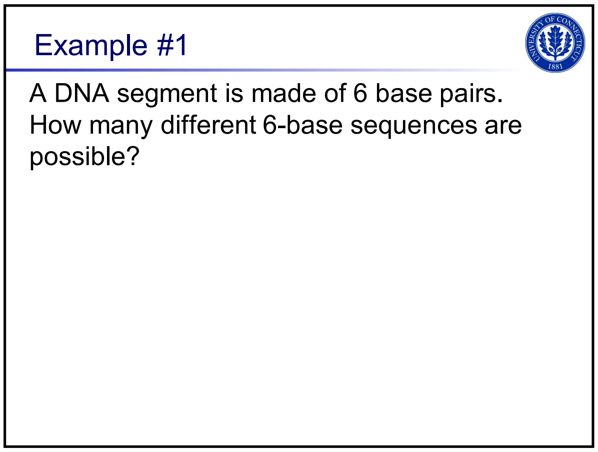
Deoxyribonucleic Acid (DNA)

- ❑ Sequence read in the **5’ carbon (C) to 3’ carbon (C)** direction

DNA double helix
(diameter: 2 nm)

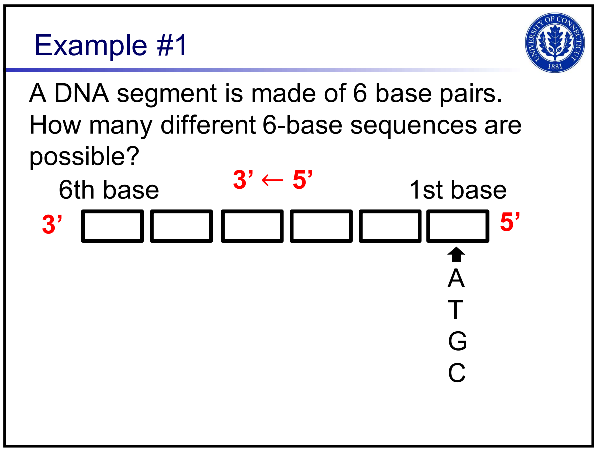


Deoxyribonucleic Acid (DNA)



Example #1

A DNA segment is made of 6 base pairs. How many different 6-base sequences are possible?



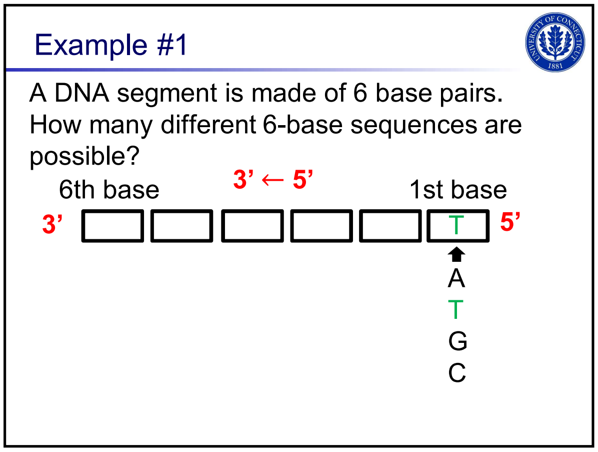
Example #1

A DNA segment is made of 6 base pairs. How many different 6-base sequences are possible?

6th base 3' ← 5' 1st base

3' □ □ □ □ □ □ 5'

A
T
G
C



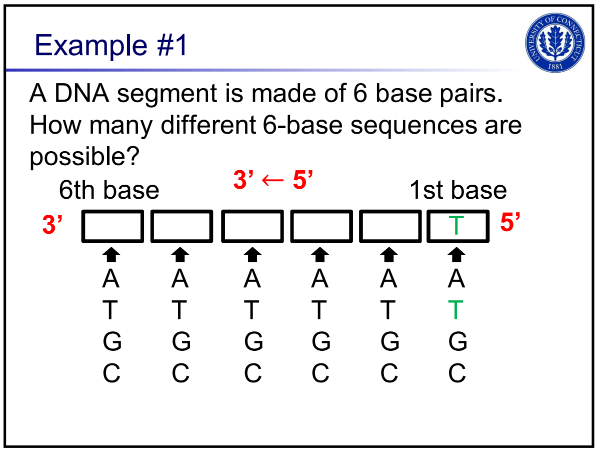
Example #1

A DNA segment is made of 6 base pairs. How many different 6-base sequences are possible?

6th base 3' ← 5' 1st base

3' □ □ □ □ □ T 5'

A
T
G
C



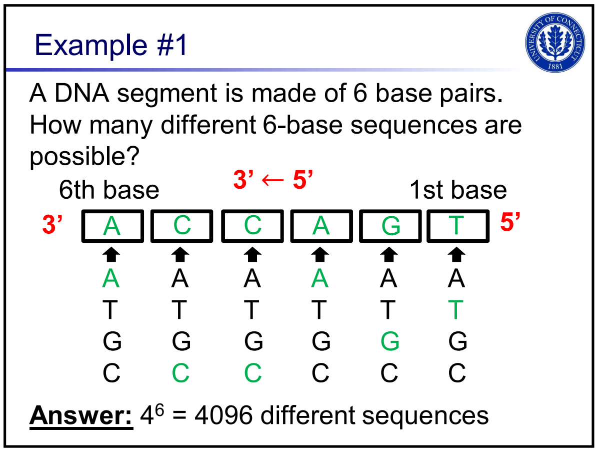
Example #1

A DNA segment is made of 6 base pairs. How many different 6-base sequences are possible?

6th base 3' ← 5' 1st base

3' □ □ □ □ □ T 5'

A A A A A A
T T T T T T
G G G G G G
C C C C C C



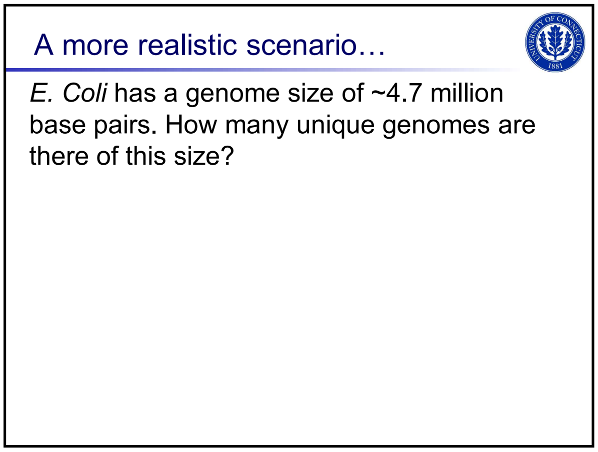
Example #1

A DNA segment is made of 6 base pairs. How many different 6-base sequences are possible?

6th base   3’ ← 5’   1st base

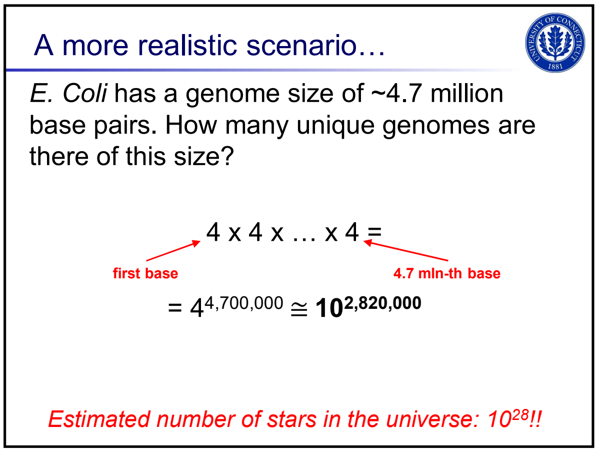
3’	A	C	C	A	G	T	5’
	⬆	⬆	⬆	⬆	⬆	⬆	
	A	A	A	A	A	A	
	T	T	T	T	T	T	
	G	G	G	G	G	G	
	C	C	C	C	C	C	

Answer: 4^6 = 4096 different sequences



A more realistic scenario…

E. Coli has a genome size of ~4.7 million base pairs. How many unique genomes are there of this size?



A more realistic scenario…

E. Coli has a genome size of ~4.7 million base pairs. How many unique genomes are there of this size?

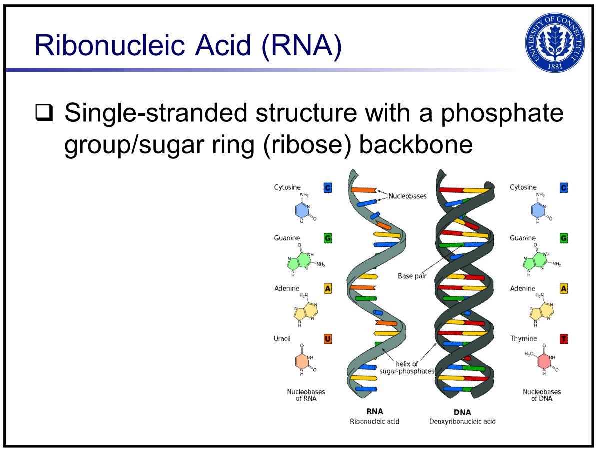
$$4 \times 4 \times \ldots \times 4 =$$

first base

4.7 mln-th base

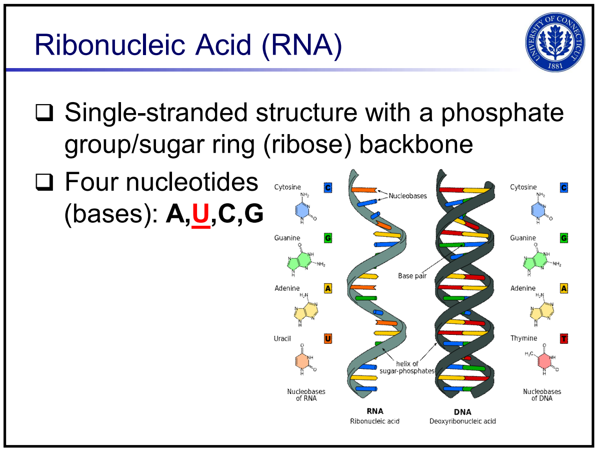
$$= 4^{4,700,000} \cong \mathbf{10}^{\mathbf{2,820,000}}$$

Estimated number of stars in the universe: 10^{28}!!



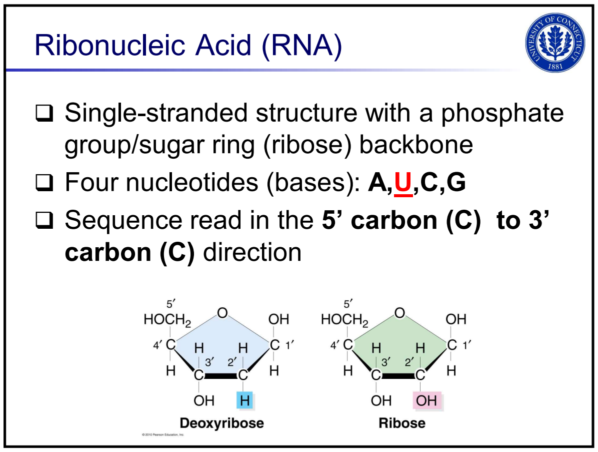
Ribonucleic Acid (RNA)

- ❑ Single-stranded structure with a phosphate group/sugar ring (ribose) backbone



Ribonucleic Acid (RNA)

- ❑ Single-stranded structure with a phosphate group/sugar ring (ribose) backbone
- ❑ Four nucleotides (bases): **A,U,C,G**



Ribonucleic Acid (RNA)

- ❑ Single-stranded structure with a phosphate group/sugar ring (ribose) backbone
- ❑ Four nucleotides (bases): **A,U,C,G**
- ❑ Sequence read in the **5' carbon (C) to 3' carbon (C)** direction

Three types of RNA

- ❑ **tRNA** (transfer RNA)

  It works as an adapter that brings amino acids to mRNA being translated

- ❑ **mRNA** (messenger RNA)

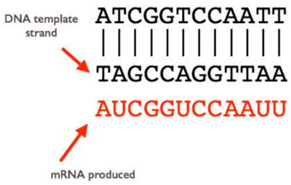
  It determines the eventual translated protein

- ❑ **rRNA** (ribosomal RNA)

  It is a component of the ribosome, which is a macromolecular machine that performs the translation from mRNA into proteins

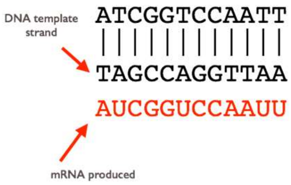
Transcription (DNA → RNA)

It is the first step of gene expression in which a particular segment of DNA (template strand) is copied into RNA (i.e., mRNA, rRNA, or tRNA) by the enzyme RNA polymerase



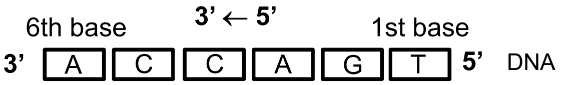
Transcription (DNA → RNA)

The resultant RNA sequence is a copy of the **complement** of the template strand, with **T** replaced by **U**



Example #2

A DNA template strand is shown below. What is the sequence of mRNA transcribed from the strand?

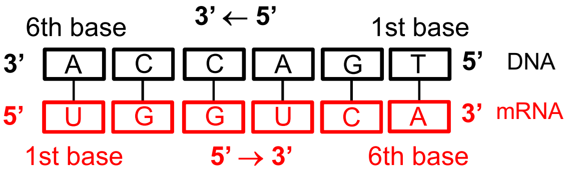


6th base **3’ ← 5’** 1st base

3’ A C C A G T **5’** DNA

Example #2

A DNA template strand is shown below. What is the sequence of mRNA transcribed from the strand?



6th base **3’ ← 5’** 1st base

3’ A C C A G T **5’** DNA

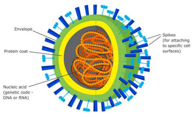
5’ U G G U C A **3’** mRNA

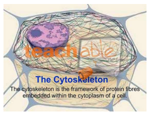
1st base **5’ → 3’** 6th base

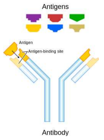
Answer: UGGUCA

Proteins

- ❑ Proteins are a special class of molecules operating within the cell

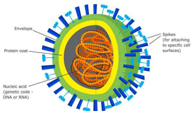

Envelope
Protein coat
Nucleic acid (genetic code - DNA or RNA)
Spikes (for attaching to specific cell surfaces)

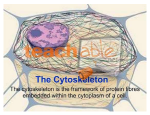

teachable.net
The Cytoskeleton
The cytoskeleton is the framework of protein fibres embedded within the cytoplasm of a cell.

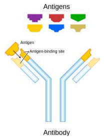


Proteins

- Depending on the task, there are several types of proteins, e.g.:
  - catalysts for chemical reactions (enzymes)
  - transportation and storage (hemoglobin)
  - regulation (hormones)
  - Defense from pathogens/invaders (antibodies)


Envelope
Protein coat
Nucleic acid (genetic code - DNA or RNA)
Spikes (for attaching to specific cell surfaces)

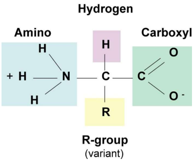

teachable.net
The Cytoskeleton
The cytoskeleton is the framework of protein fibres embedded within the cytoplasm of a cell.

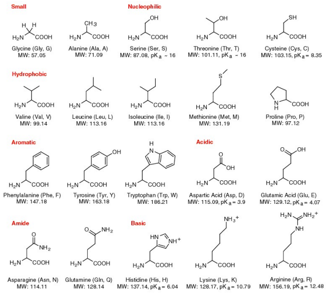


Proteins

- A proteins is made of 20 different types of basic units called **amino acids**

Amino Acid Structure

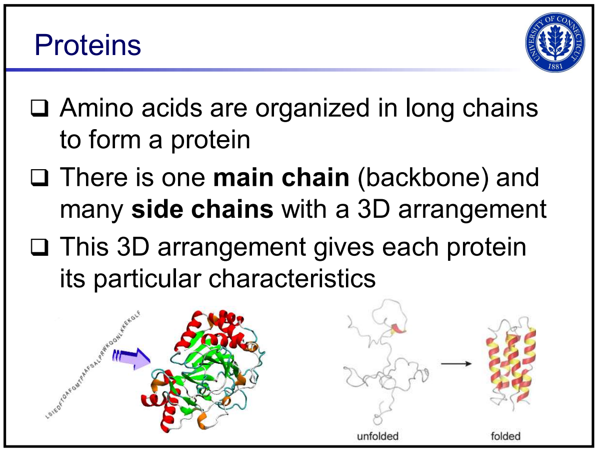



Small
Nucleophilic
Glycine (Gly, G) MW: 57.05
Alanine (Ala, A) MW: 71.09
Serine (Ser, S) MW: 87.08, pKa ~ 16
Threonine (Thr, T) MW: 101.11, pKa ~ 16
Cysteine (Cys, C) MW: 103.15, pKa = 8.35
Hydrophobic
Valine (Val, V) MW: 99.14
Leucine (Leu, L) MW: 113.16
Isoleucine (Ile, I) MW: 113.16
Methionine (Met, M) MW: 131.19
Proline (Pro, P) MW: 97.12
Aromatic
Acidic
Phenylalanine (Phe, F) MW: 147.18
Tyrosine (Tyr, Y) MW: 163.18
Tryptophan (Trp, W) MW: 186.21
Aspartic Acid (Asp, D) MW: 115.09, pKa = 3.9
Glutamic Acid (Glu, E) MW: 129.12, pKa = 4.07
Amide
Basic
Asparagine (Asn, N) MW: 114.11
Glutamine (Gln, Q) MW: 128.14
Histidine (His, H) MW: 137.14, pKa = 6.04
Lysine (Lys, K) MW: 128.17, pKa = 10.79
Arginine (Arg, R) MW: 156.19, pKa = 12.48

20 different R-groups

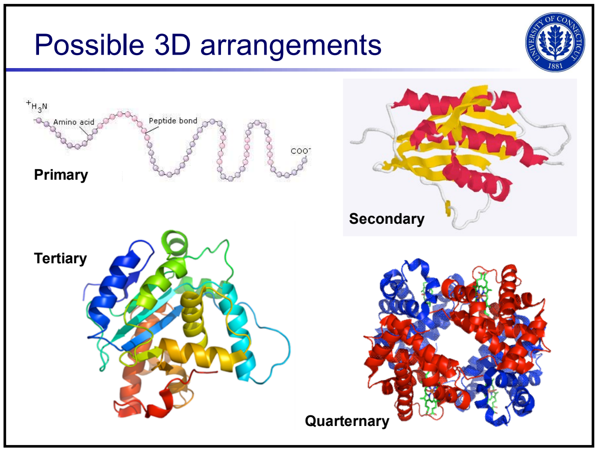
A list of amino acids

Alanine	**Ala**	**A**
Arginine	**Arg**	**R**
Aspartic Acid	**Asp**	**D**
Asparagine	**Asn**	**N**
Cysteine	**Cys**	**C**
Glutamic Acid	**Glu**	**E**
Glutamine	**Gln**	**Q**
Glycine	**Gly**	**G**
Histidine	**His**	**H**
Isoleucine	**Ile**	**I**
Leucine	**Leu**	**L**
Lysine	**Lys**	**K**
Methionine	**Met**	**M**
Phenylalanine	**Phe**	**F**
Proline	**Pro**	**P**
Serine	**Ser**	**S**
Threonine	**Thr**	**T**
Tryptophan	**Trp**	**W**
Tyrosine	**Tyr**	**Y**
Valine	**Val**	**V**

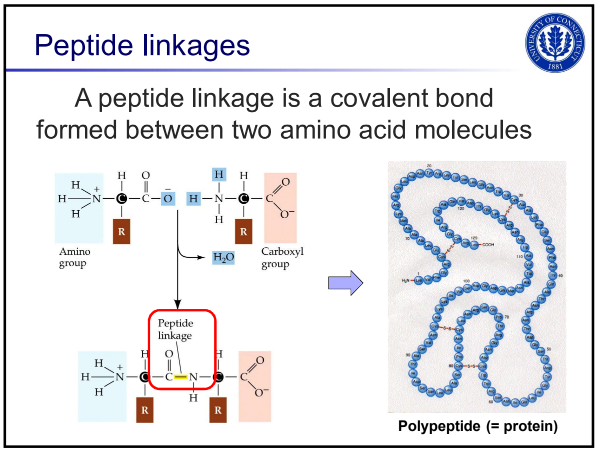


Proteins

- ❑ Amino acids are organized in long chains to form a protein
- ❑ There is one **main chain** (backbone) and many **side chains** with a 3D arrangement
- ❑ This 3D arrangement gives each protein its particular characteristics



Possible 3D arrangements



Peptide linkages

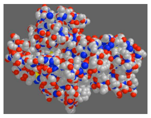
A peptide linkage is a covalent bond formed between two amino acid molecules

Polypeptide (= protein)

An example

Hexokinase is an enzyme protein involved in glycolysis (457 amino acids)

```
              5         10         15         20         25         30
  1 A A S X D X S L V E V H X X V F I V P P X I L Q A V V S I A
 31 T T R X D D X D S A A A S I P M V P G W V L K Q V X G S Q A
 61 G S F L A I V M G G G D L E V I L I X L A G Y Q E S S I X A
 91 S R S L A A S M X T T A I P S D L W G N X A X S N A A F S S
121 X E F S S X A G S V P L G F T F X E A G A K E X V I K G Q I
151 T X Q A X A F S L A X L X K L I S A M X N A X F P A G D X X
181 X X V A D I X D S H G I L X X V N Y T D A X I K M G I I F G
211 S G V N A A Y W C D S T X I A D A A D A G X X G G A G X M X
241 V C C X Q D S F R K A F P S L P Q I X Y X X T L N X X S P X
271 A X K T F E K N S X A K N X G Q S L R D V L M X Y K X X G Q
301 X H X X X A X D F X A A N V E N S S Y P A K I Q K L P H F D
331 L R X X X D L F X G D Q G I A X K T X M K X V V R R X L F L
361 I A A Y A F R L V V C X I X A I C Q K K G Y S S G H I A A X
391 G S X R D Y S G F S X N S A T X N X N I Y G W P Q S A X X S
421 K P I X I T P A I D G E G A A X X V I X S I A S S Q X X X A
451 X X S A X X A
```



Example #3

A polypeptide is composed of 10 amino acids. How many different amino acid sequences are possible?

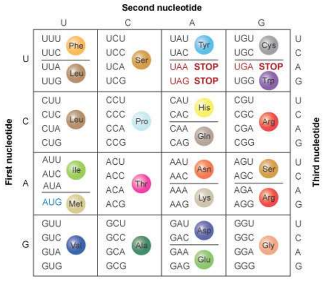
Example #3

A polypeptide is composed of 10 amino acids. How many different amino acid sequences are possible?

<u>Answer:</u> 10^{20} different sequences

Translation (mRNA → Protein)

The process of translating a mRNA sequence into a protein sequence

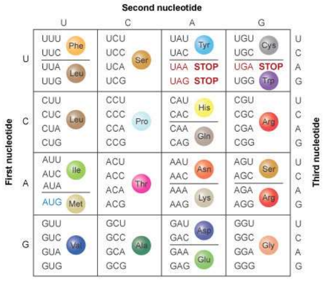


Second nucleotide

First nucleotide	U	C	A	G	Third nucleotide
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 (mRNA → Protein)

The process of translating a mRNA sequence into a protein sequence



Second nucleotide

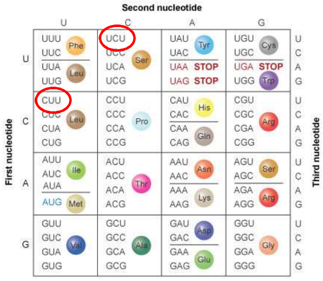
First nucleotide	U	C	A	G	Third nucleotide
U	UUU UUC Phe UUA UUG Leu	UCU UCC UCA UCG Ser	UAU UAC Tyr UAA STOP UAG STOP	UGU UGC Cys UGA STOP UGG Trp	U C A G
C	CUU CUC CUA CUG Leu	CCU CCC CCA CCG Pro	CAU CAC His CAA CAG Gln	CGU CGC CGA CGG Arg	U C A G
A	AUU AUC AUA Ile AUG Met	ACU ACC ACA ACG Thr	AAU AAC 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

Each set of 3 bases corresponds to one specific amino acid

Why do we need at least 3 bases to code for one amino acid?

Translation (mRNA → Protein)

The process of translating a mRNA sequence into a protein sequence



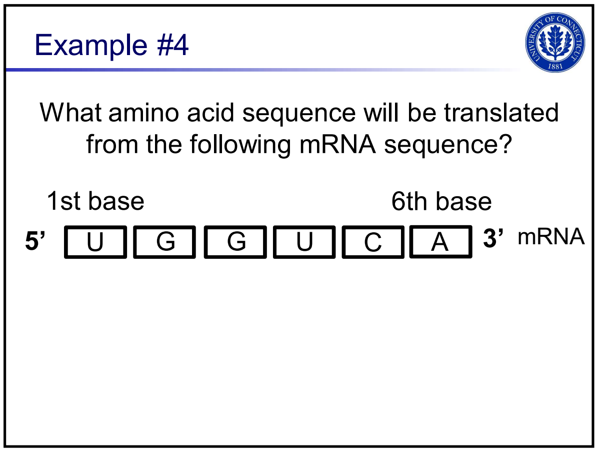
Second nucleotide

First nucleotide	U	C	A	G	Third nucleotide
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

Example:

CUU (mRNA)
→ leu (amino acid)

UCU (mRNA)
→ ser (amino acid)

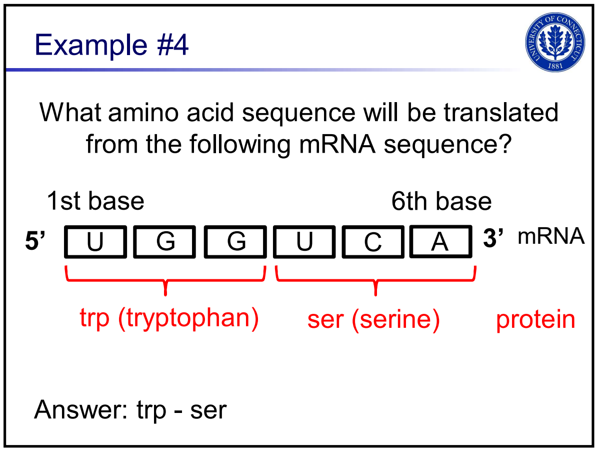


Example #4

What amino acid sequence will be translated from the following mRNA sequence?

1st base 6th base

5’ U G G U C A 3’ mRNA



Example #4

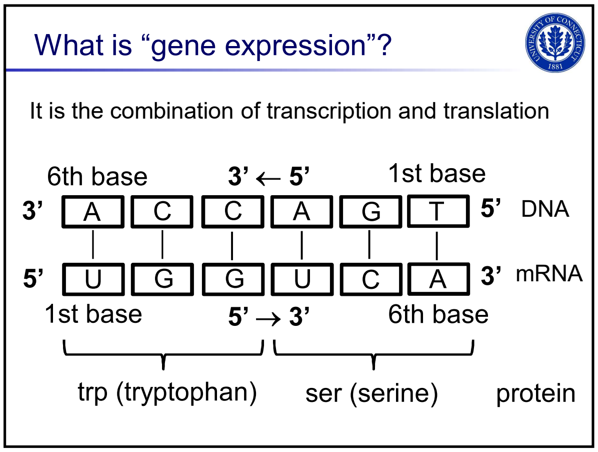
What amino acid sequence will be translated from the following mRNA sequence?

1st base 6th base

5’ U G G U C A 3’ mRNA

trp (tryptophan) ser (serine) protein

Answer: trp - ser



What is “gene expression”?

It is the combination of transcription and translation

6th base 3’ ← 5’ 1st base

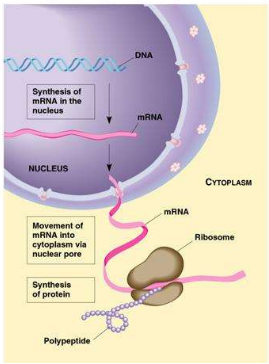
3’ A C C A G T 5’ DNA

5’ U G G U C A 3’ mRNA

1st base 5’ → 3’ 6th base

trp (tryptophan) ser (serine) protein

What is “gene expression”?

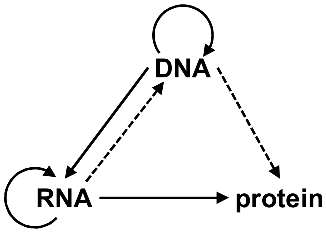

DNA
Synthesis of mRNA in the nucleus
mRNA
NUCLEUS
CYTOPLASM
mRNA
Movement of mRNA into cytoplasm via nuclear pore
Ribosome
Synthesis of protein
Polypeptide

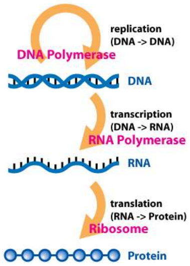
DNA → mRNA → Protein
transcription translation

The central dogma

source: *http://www.youtube.com/watch?v=41_Ne5mS2ls*

The central dogma





The dominant path of information flow in biology is from DNA to RNA to proteins