BMEG3102 Bioinformatics

Lecture 1. Introduction

Qi Dou

Email: qidou@cuhk.edu.hk

Office: Room 1014, 10/F, SHB

BMEG3102 Bioinformatics

The Chinese University of Hong Kong

Lecture outline

- Course information
- Introduction to bioinformatics
- -- (Intermission: Background survey)
- Introduction to genetics and molecular biology
- Data in bioinformatics

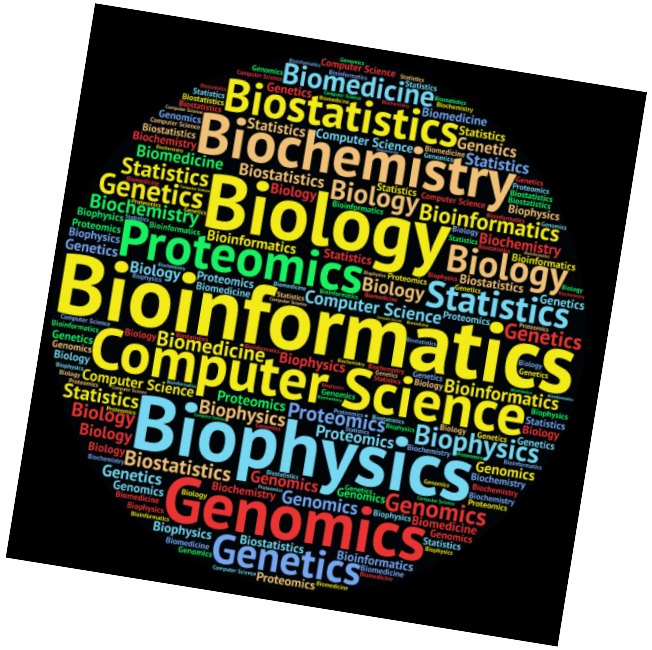


Image credit: https://towardsdatascience.com/a-dummies-intro-to-bioinformatics-e8212ed7c09b

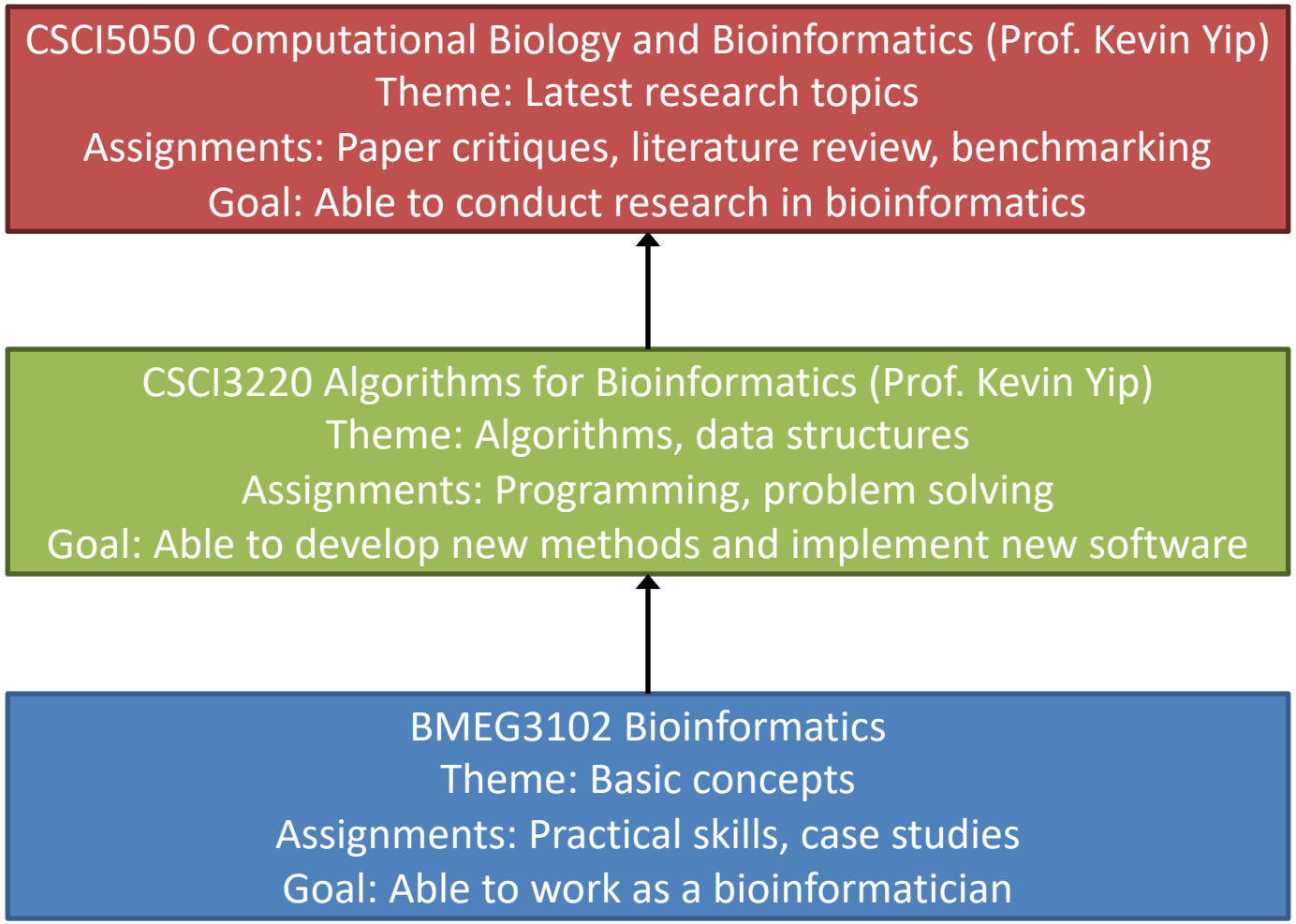
Part 1

COURSE INFORMATION

Course objectives

- **To learn what bioinformatics is about**
  - What it is about
  - Why it is important
  - What the main challenges are
  - Hopefully, to arouse your interests in this area
- **To learn some basic knowledge in bioinformatics**
- **To get hands-on experience in using some tools to solve simple problems**
  - And to know how to discover new resources
  - So that you can perform analyses on your own afterwards

Multi-course design



Acknowledgement to Prof. Kevin Yip for supporting BMEG3102 course materials for this Spring2021 term.

Teaching staff

- **Lecturers**

  – **Dr. Qi DOU**
  Department of Computer Science and Engineering
  qidou@cuhk.edu.hk

  Room 1014, 10/F, Ho Sin-Hang Engineering Building
  Consultation hours: Mon 14:00-16:00

  (please make appointment by email)

  – **Dr. Huating WANG**
  Department of Orthopedics and Traumatology
  huating.wang@cuhk.edu.hk

Teaching staff

- **Teaching assistant**

  - Ms. Yizhen CHEN
  Department of Computer Science and Engineering
  yzchen@cse.cuhk.edu.hk
  Room 1023, 10/F, Ho Sin-Hang Engineering Building
  Consultation hours: Tue 14:00-16:00

Class time & venue

- **Lectures**

  Wednesdays 10:30 – 11:15
  Thursdays 11:30 – 13:15

- **Tutorials**

  Wednesdays 11:30 – 12:15

- **Online teaching (lecture & tutorial)**

  Meeting ID: 956 0245 5157
  Passcode: 5y845q

  https://cuhk.zoom.us/j/95602455157?pwd=RHk5MGdvNkQveFVzMzlMNThmaS8rQT09

Class time summary

Section	Time	Mon	Tue	Wed	Thu	Fri
1	08:30-09:15					
2	09:30-10:15					
3	10:30-11:15			Lecture		
4	11:30-12:15			Tutorial	Lecture	
5	12:30-13:15				Lecture	
6	13:30-14:15	Consultation hours: Qi	Consultation hours: Yizhen			
7	14:30-15:15					
8	15:30-16:15					
9	16:30-17:15					
10	17:30-18:15					

Course websites

- **Course Web site:** http://www.cse.cuhk.edu.hk/~qdou/bmeg3102/
  - Lecture notes, tutorial notes
  - Teaching schedule
- **Blackboard** (https://blackboard.cuhk.edu.hk/, look for course 2020R2-BMEG3102)
  - Lecture notes, tutorial notes
  - Assignment specifications
  - Assignment collection boxes
  - Announcements – check your link.cuhk.edu email account
  - Discussion forum
- **uReply** (http://web.ureply.mobi/getstarted.php)
  - Interactive tasks

Reference materials

- **YouTube channel** (https://www.youtube.com/channel/UCk2ozjkbfpteeJUolHWNsMw)
  - Micro-modules on some topics

- **Book References:**
  - *Fundamental Concepts of Bioinformatics* by Dan E. Krane, Michael L. Raymer and Benjamin Cummings, Pearson Education, 2003
  - *Algorithms in Bioinformatics: A Practical Introduction* by Wing-Kin Sung, Chapman & Hall, 2009 (with free online materials)

Assessment

- **Assignments** 40%
  - Tentatively 4 of them in total
  - Conceptual and practical questions
  - No heavy programming
- **Class participation** 5%
  - uReply
- **Reading presentation** 5%
  - Form groups (2-3 students), provide easy articles
  - Presentation 8min + 2min Q&A
- **Midterm examination** 20%
  - Take home exam
  - Release after lecture, on 11 March
- **Final examination** 30%
  - Closed book, closed notes

Tentative class schedule

Week	Time	Topic	Lecturer	Tasks
1	13-14 Jan	Introduction	Dr. Dou	
2-4	20-21 Jan / 27-28 Jan / 3-4 Feb / 10 Feb	Sequence alignment and searching	Dr. Dou	Assignment #1
5		(Lunar new year)		
6-7	18 Feb / 24-25 Feb	Mutation models and molecular phylogenetics	Dr. Dou	Assignment #2
8-9	3-4 Mar / 10-11 Mar	Motifs and domains	Dr. Dou	Midterm exam.
10	17-18 Mar	Human genetics and genetic disorders	Dr. Wang	
11	24-25 Mar	High-throughput data processing and analysis	Dr. Dou	Assignment #3
12	31 Mar, 1 Apr	(Reading week)	Dr. Dou	
13	8 Apr	Functional annotations	Dr. Dou	Presentation
14-15	14-15 Apr / 21-22 Apr	Molecular structures	Dr. Dou	Assignment #4
?				Final exam.

Promises

- Putting up lecture notes in time
- Suitable teaching pace and level of difficulty
  - feedback from you is crucial
- Quick responses to emails
- Timely announcements in Blackboard
- Prompt and fair gradings of assignments

Expectations

- Attending lectures, punctuality
- *Active class participation*
- Finishing assignments in time
  - Special note on academic honesty: CUHK has rigorous policies against dishonest acts such as plagiarism.

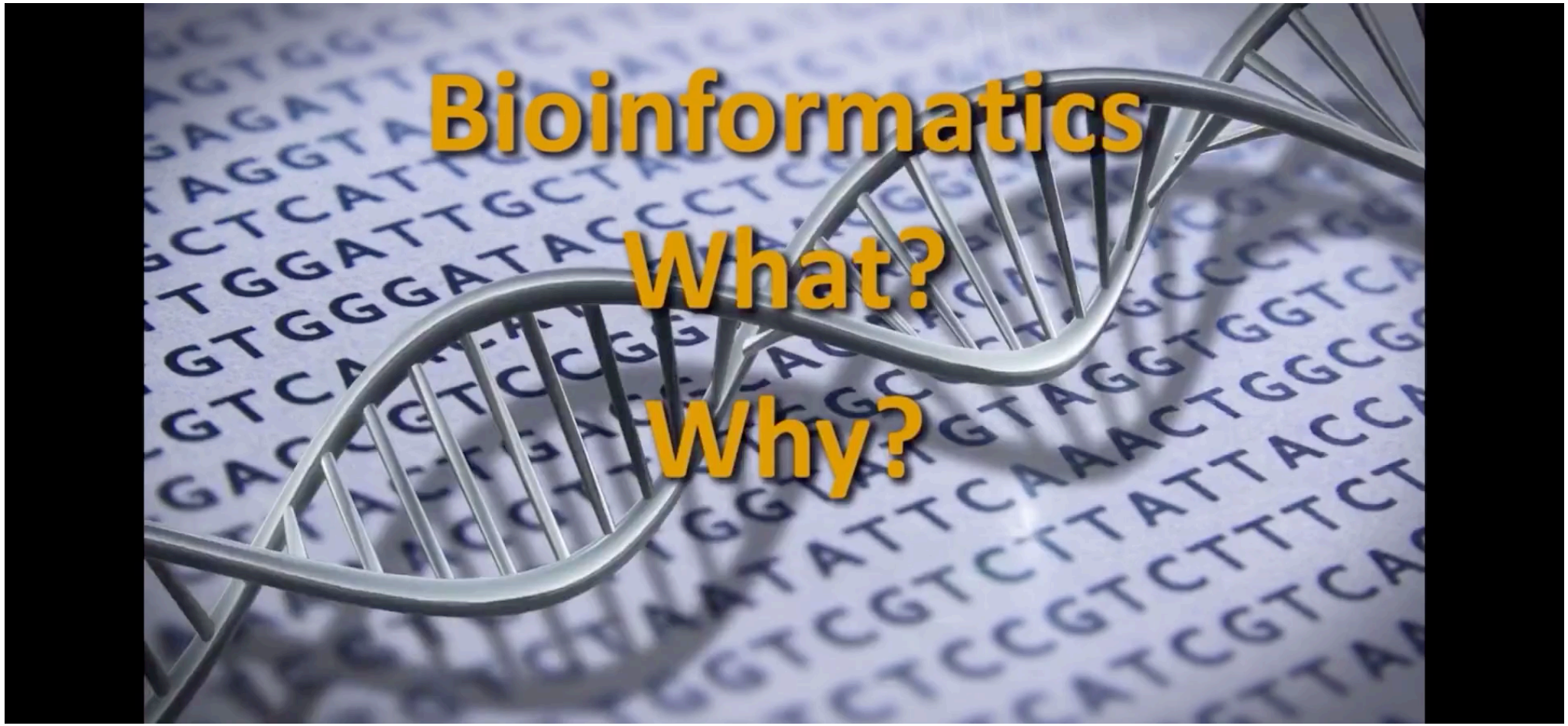
Part 2

INTRODUCTION TO BIOINFORMATICS

What is bioinformatics?

- **What is bioinformatics?**
  - –Bio-informatics
  - –Bio: Biology, the study of life and living organisms [Wikipedia]
  - –Informatics: Information science [Webster]
  - –Bioinformatics: Application of computer science and information technology to the field of biology and medicine [Wikipedia]

What is bioinformatics?



video credit to: https://www.youtube.com/watch?v=v1cTNhiZ2_c

Why we need bioinformatics?

- **Why do we need computing methods to assist biomedical research?**
  - Large data size
  - Difficult computational problems

A toy example?

AAACGTACGTATTCGGGCCATCGAGGCTAGCGGCACTTC
GAGCGATCTATCGGGAGCTTTGGCTATCGATCGGGCGAT
CGATGCTGACGTACGTAGCGCGCGATCGAGCGCGGCTAG
CTAGCGGCATCGTAGCTACGTAGCTACGGCGCTATTTCG
ATCGAGTCGTGTCTAGTCGGATATAGCTATGCATCTAGC
TGAGGCGATCTGAGCGGATCGATGCTAGGGCGATCGGAG
CTAGCTGAGCTAGCTAGCTGAGCGCTAGCGAGCGTACGA
GCGATCGAGCGAGTCTAGCGAGCGATTCTAGCGATCGAG
CGTCTACGATCGTATGCTAGCTAGGGCTAGCATGCGGAT
CTATCGAGCGGCTATCTGAGCGATTCGATCGAGCGATCT
AGCGAGCTATCGATCGAGCCGGCTCACCGTCGTAAATCT
ATGATCTGGCTTGGCCTGCAGTAGCTCTTTCATTTCGGG
CTTATCTAATGCTGACTGGTCGGTCCTGGCTACGCTCCA

Find
TACGA?

A toy example?

Find TACGA?

Large data size

Question 1:

How many cells do we have in our body?

- Each adult human has

  10^{13}-10^{14} cells

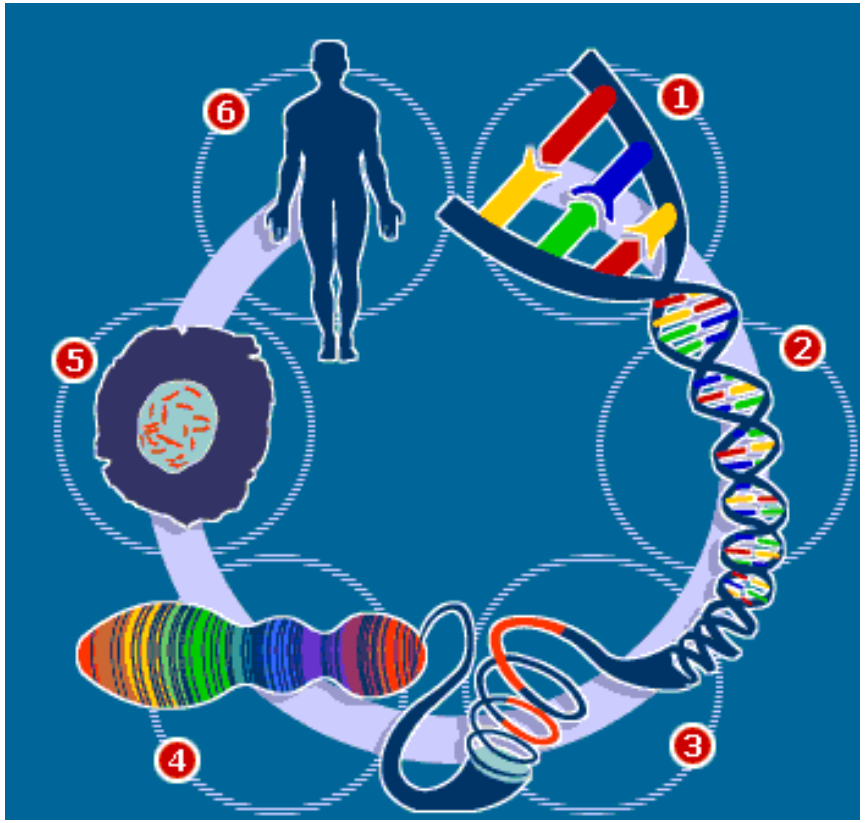


Image credit: news.bbc.co.uk

Large data size

- Each adult human has 10^{13}-10^{14} cells
- Most of them contain two copies of DNA with 3×10^{9} nucleotides (the “haploid genome”)
- If we represent DNA as a string with four letters, `A`, `C`, `G` and `T`…

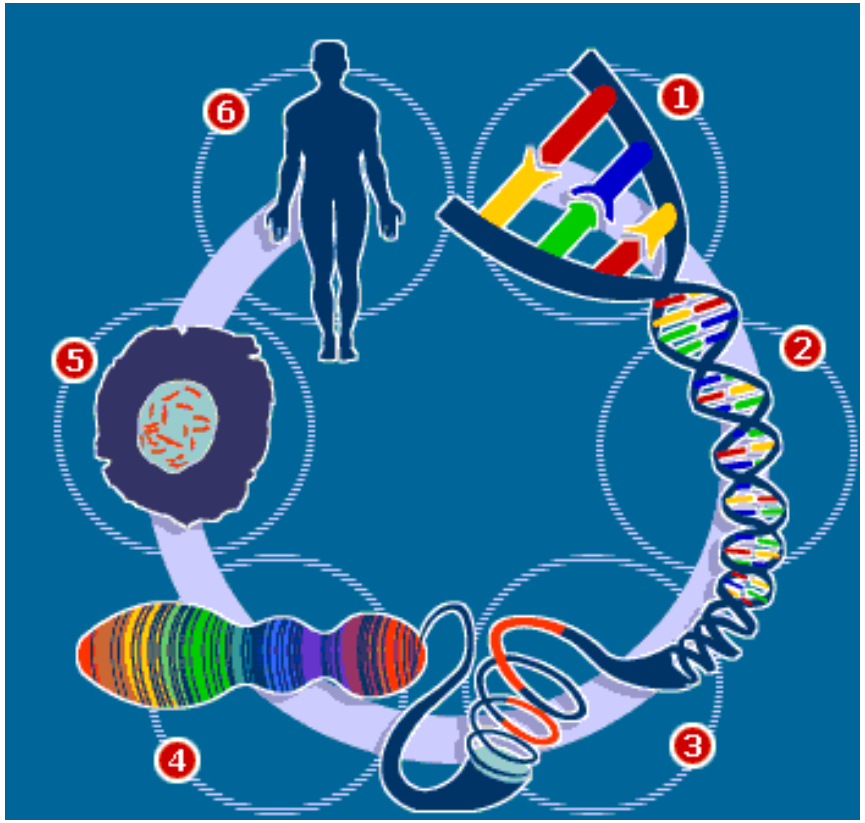


Image credit: news.bbc.co.uk

Large data size

AAACGTACGTATTCGGGCCATCGAGGCTAGCGGCACTTC
GAGCGATCTATCGGGAGCTTTGGCTATCGATCGGGCGAT
CGATGCTGACGTACGTAGCGCGCGATCGAGCGCGGCTAG
CTAGCGGCATCGTAGCTACGTAGCTACGGCGCTATTTCG
ATCGAGTCGTGTCTAGTCGGATATAGCTATGCATCTAGC
TGAGGCGATCTGAGCGGATCGATGCTAGGGCGATCGGAG
CTAGCTGAGCTAGCTAGCTGAGCGCTAGCGAGCGTACGA
GCGATCGAGCGAGTCTAGCGAGCGATTCTAGCGATCGAG
CGTCTACGATCGTATGCTAGCTAGGGCTAGCATGCGGAT
CTATCGAGCGGCTATCTGAGCGATTCGATCGAGCGATCT
AGCGAGCTATCGATCGAGCCGGCTCACCGTCGTAAATCT
ATGATCTGGCTTGGCCTGCAGTAGCTCTTTCATTTCGGG
CTTATCTAATGCTGACTGGTCGGTCCTGGCTACGCTCCA

Large data size

- The last page contains about 500 characters
  - Need 6,000,000 pages to show the human genome
  - Printed in 130 books

Large data size

- The last page contains about 500 characters
  - Need 6,000,000 pages to show the human genome
  - Printed in 130 books

Leicester scientists print human genome in 130 books

Scientists at the University of Leicester have printed the whole of the human genome to show just how much information it takes to make up one human body.

They say it has taken 130 book volumes, which would take 95 years to read.

Large data size

- Some large projects study the DNA of many people
  - E.g., study a certain disease, typical way is to compare two groups of people (diseased v.s. healthy)
  - Just compare one person each is useless
  - To find out the locations in the DNA that are related to the disease,
  - Need to have large number of samples to compare with (e.g., 5000 w/ disease, 5000 w/o disease)

    Find things that are common in one group, that is different in the other group

Large data size

- Humans have 20,000-25,000 genes that produce proteins
  - These proteins do not work alone, they work cooperatively
  - Generate a combinatory number of functions
  - We want to study their pair-wise and higher-order relationships
  - About 3.1×10^8 pairs, 2.6×10^{12} triples, ...
  - Grow up very quickly, generate large data size
  - After collecting the data/information, we can then make a lot of hypothesis
    e.g., for some proteins, if they are not present in the cells anymore, the cells can not survive well
    for some triples, if they are dissociated, then certain morphology of the cell will be affected
  - Observe a lot of things in the same experiment, common in nowadays biology

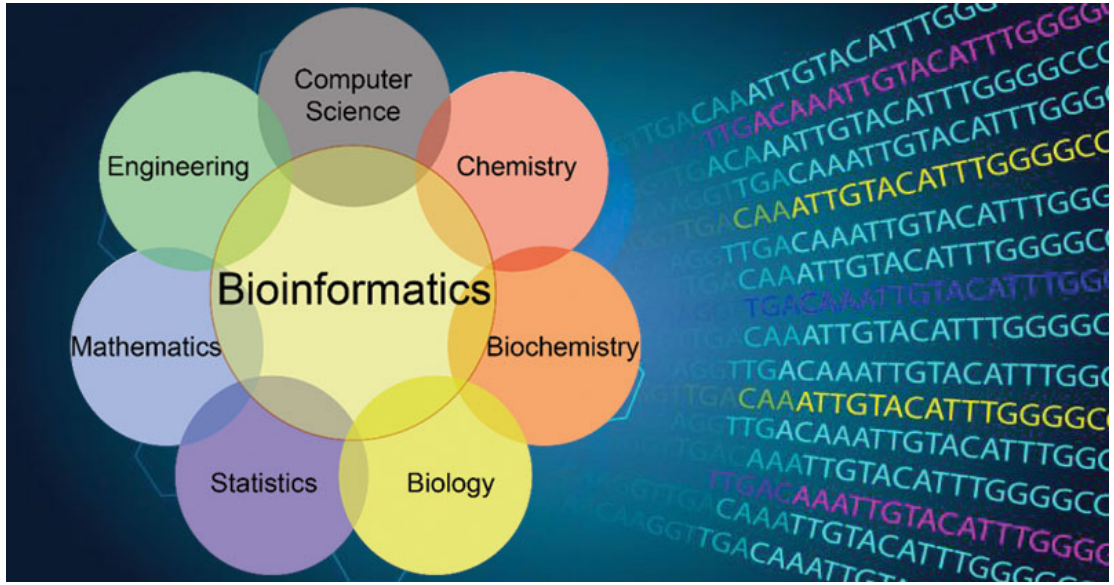
Real life example

- **Biomedical scenario:** I have sequenced the DNA of 100 lung cancer samples and 100 controls, how do I find out the disease-associated genetic variants?
  - Find out all the disease associated genetic variates, i.e., positions in DNA that can help distinguish whether it is a cancerous sample or non-cancer sample
  - In other words, form a computational model

- **Basic problems:**
  - Identifying genetic variants in each sample, identify those differences among the different samples
    - String comparisons -- **Computer science**
  - Determining whether the genetic variants can separate the two groups of samples well
    - Testing how different two groups of points are -- **Statistics**
  - Identifying the most confident set of variants for experimental validation and functional study
    - Using knowledge about lung cancer to select -- **Biology and medicine**

What is bioinformatics?

- **Related fields**
  - –Computer science
    - Algorithms
    - Database management
    - Machine learning
    - Software engineering
    - …
  - –Statistics
  - –Biology
    - Molecular biology
    - Genetics
    - …
  - –Biotechnology
  - –Medicine
  - –…



- A multi-disciplinary area that solves hard biomedical problems by combining the knowledge from many fields

Image credit: https://microbenotes.com/bioinformatics-introduction-and-applications/

What is bioinformatics?

- **Contributions and prospects**
  - –A new and rapidly growing field with a lot of potentials
  - –Very meaningful field, with direct contributions to
    - Medicine
    - Biology
    - Computer science
    - …
  - –Cutting-edge, a lot of challenging problems
  - –A bottleneck in biomedical research
  - –Short of qualified people

Career

- Where can we find jobs for bioinformaticians?
  - Universities
  - Research institutes
  - Hospitals
  - Pharmaceutical companies
  - Biotechnology companies
  - Sequencing centers
  - Personal genomics companies
  - …
- Good prospects worldwide, growing in Hong Kong

What is your answer?

What will be your own answer at the end of this semester?

What is bioinformatics to you?

- An interesting course that you have taken?
- A research area that you want to study in your graduate school?
- An area in which you want to develop your career?

Intermission

BACKGROUND SURVEY

Purpose

- To determine…
  - –Materials to be covered
  - –Ways of presentation
  - –Teaching pace and level of difficulty

The survey

- Go to uReply now if you have Internet access

- Questions:

1. What do you want to learn from this course? (Check one for each row.)

	Yes, a lot!	Yes	Not too much, please	No!!!	What is it?
Bioinformatics concepts					
Bioinformatics algorithms					
Bioinformatics tools					
Bioinformatics databases					
Bioinformatics research					

The survey

- Questions:

    2. Did you study biology before? At which level?
    3. List the programming languages that you know, and your proficiency level.

The survey

- Questions:

4. How much do you know about these topics?

	I can teach this topic	I know it	Sort of heard about it	Huh?
Hash tables				
Big-O notation				
Statistical testing				
First-order differential equations				
Missense mutations				
6-frame translation				
Sanger sequencing				
Massively parallel sequencing				

5. Write down any special requests that you may have for this course.

Part 3

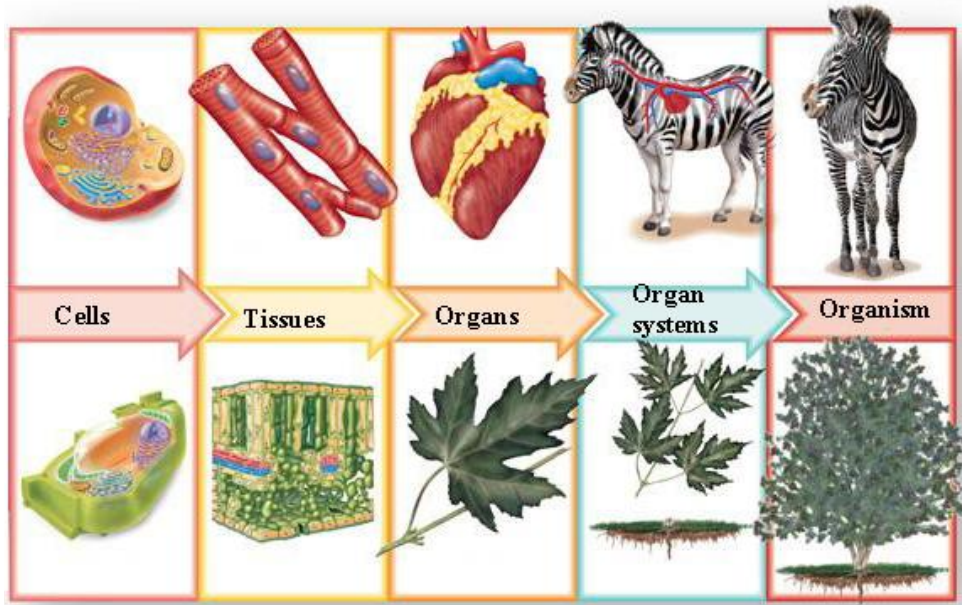
INTRODUCTION TO GENETICS AND MOLECULAR BIOLOGY

Basic biological knowledge

- Useful for
  - To refresh your general knowledge
  - Defining terminology
  - Helping you appreciate the importance of what you are going to learn
- Don’t panic:
  - You should have learned most things before. This is just a revision.
  - This is not a biology/biochemistry class. You don’t need to memorize everything.
  - Treat it as something fun.
- Use this set of slides as a reference.

Introduction to molecular biology

- Cell: Basic functional unit of life



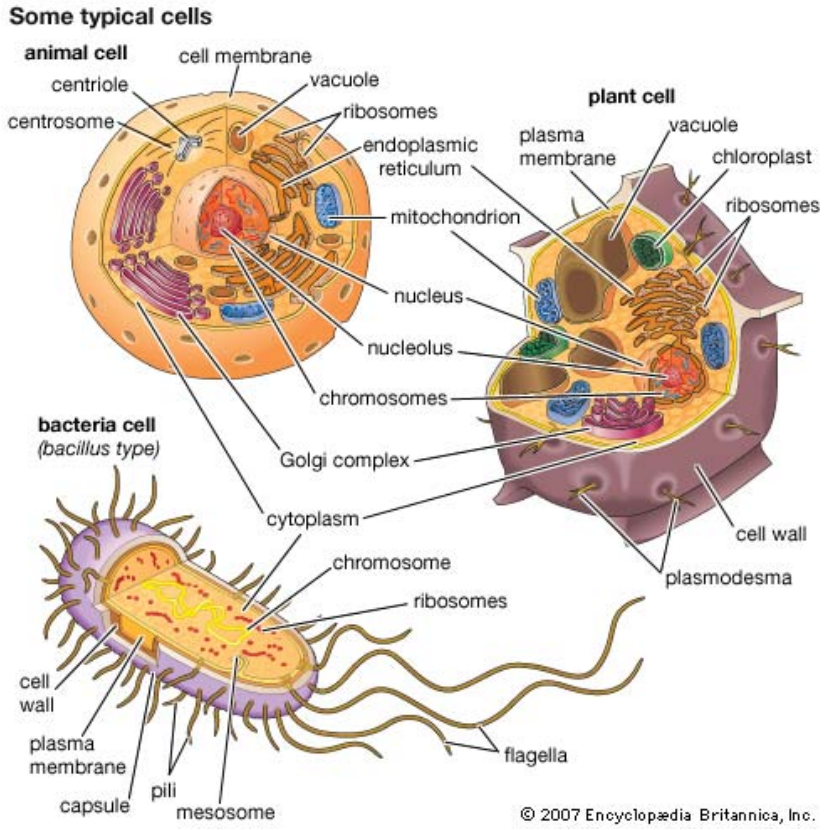


Image credit: http://legacy.hopkinsville.kctcs.edu/sitecore/instructors/Jason-Arnold/VLI/Module%201/m1science/f1-01_levels_of_biologi_c.jpg, http://dbscience5.wikispaces.com/file/view/78585-004-A63E1F47.jpg/51586701/78585-004-A63E1F47.jpg

Chromosome

- In human, each DNA-containing somatic cell has 23 pairs of chromosomes (one from father, one from mother)
  - –Chr1, Chr2, …, Chr22, ChrX, ChrY
  - –Male: XY; Female: XX
  - –(Mitochondrial DNA)
- For higher organisms, chromosomes are in the cell nucleus
- When cell divides by mitosis, each chromosome is duplicated and both daughter cells have the complete set of chromosomes

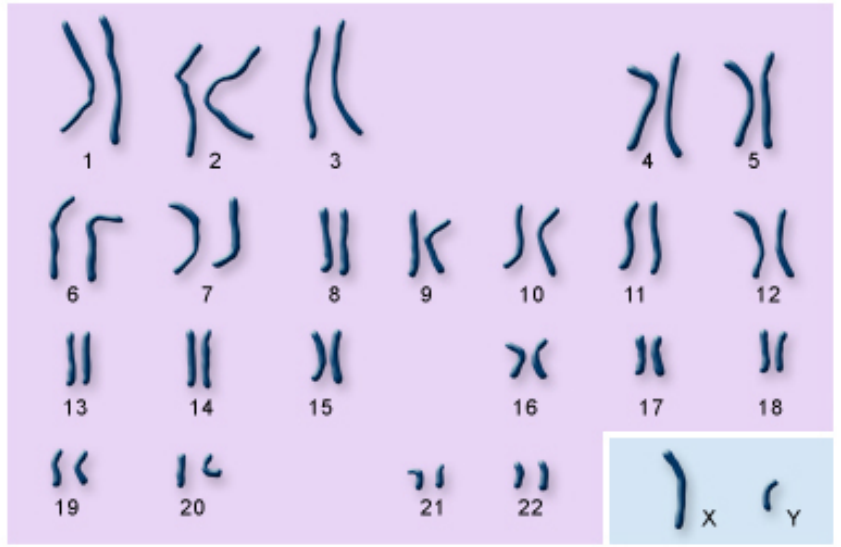
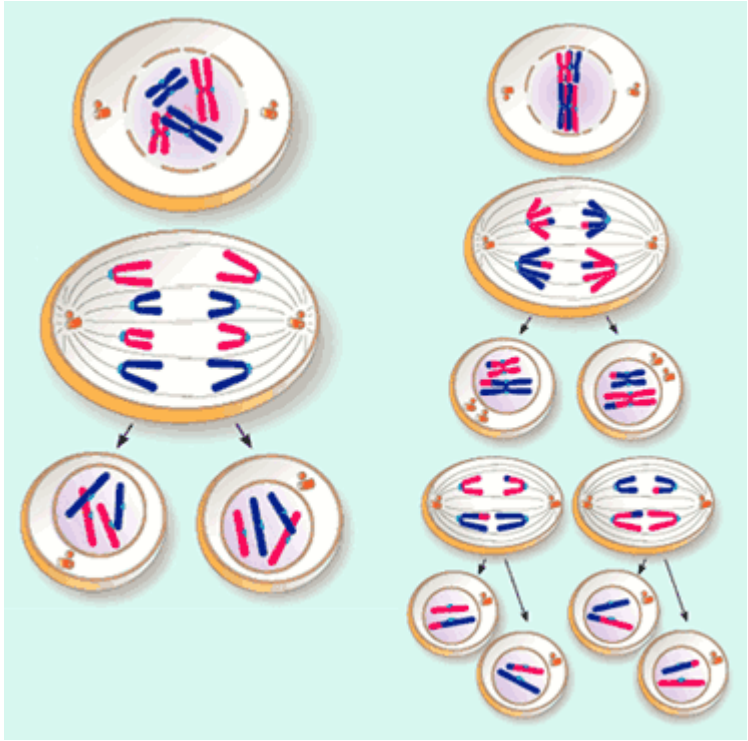


Image credit: http://ghr.nlm.nih.gov/handbook/illustrations/chromosomes.jpg

Chromosome and inheritance

- Each germ cell contains only one of each pair of chromosomes by a process called meiosis

Mitosis:
- Resulting in two cells
- Diploid: Each has 23 pairs



Meiosis:
- Resulting in four cells
- Haploid: Only one copy of each chromosome

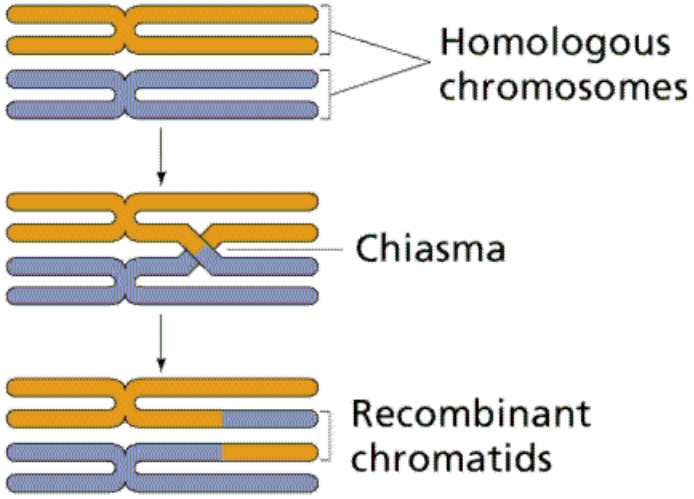
Image credit: http://3.bp.blogspot.com/_207DNIaL-gc/TQk9QRaI5mI/AAAAAAAAAXg/z0Xh8CTgHto/s400/mitosismeiosissummary.gif

Diploid genome

- Why we need two copies of each chromosome?
  - More combinations: For each of the 23 pairs of chromosomes, only one is passed to each offspring, which creates 2^{23} possible combinations.
  - Error tolerance: If one copy has problem, there is still another copy.
  - Evolution: Having one normal copy, the other is more free to change, sometimes resulting in an overall advantage.

How to change?

- Recombination
- Insertion
- Deletion
- …



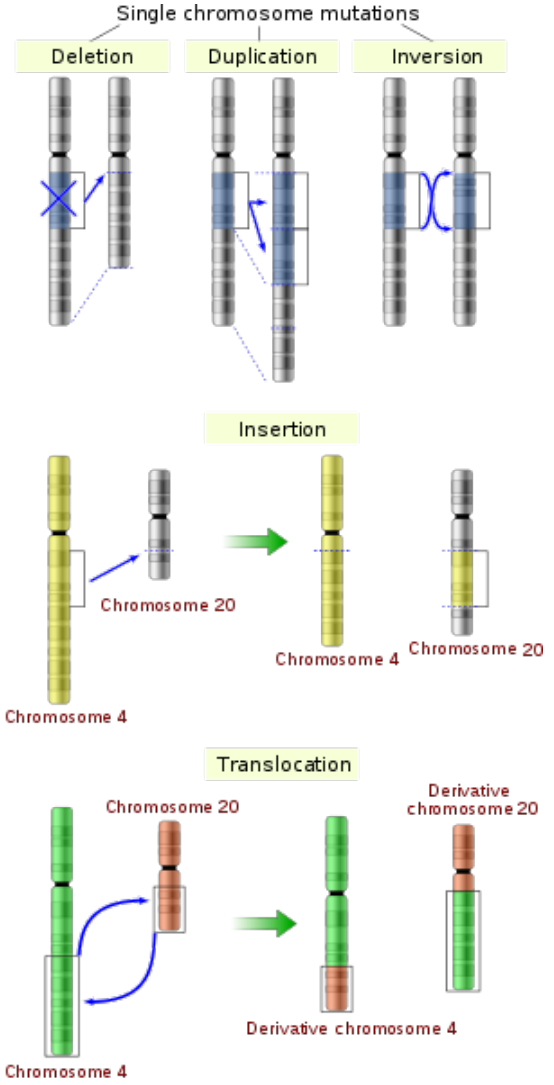


Image credit: http://www2.estrellamountain.edu/faculty/farabee/biobk/Crossover.gif, Wikipedia

Why do changes matter?

- Need to know what's in a chromosome
  - Chromosome → chromatin → DNA

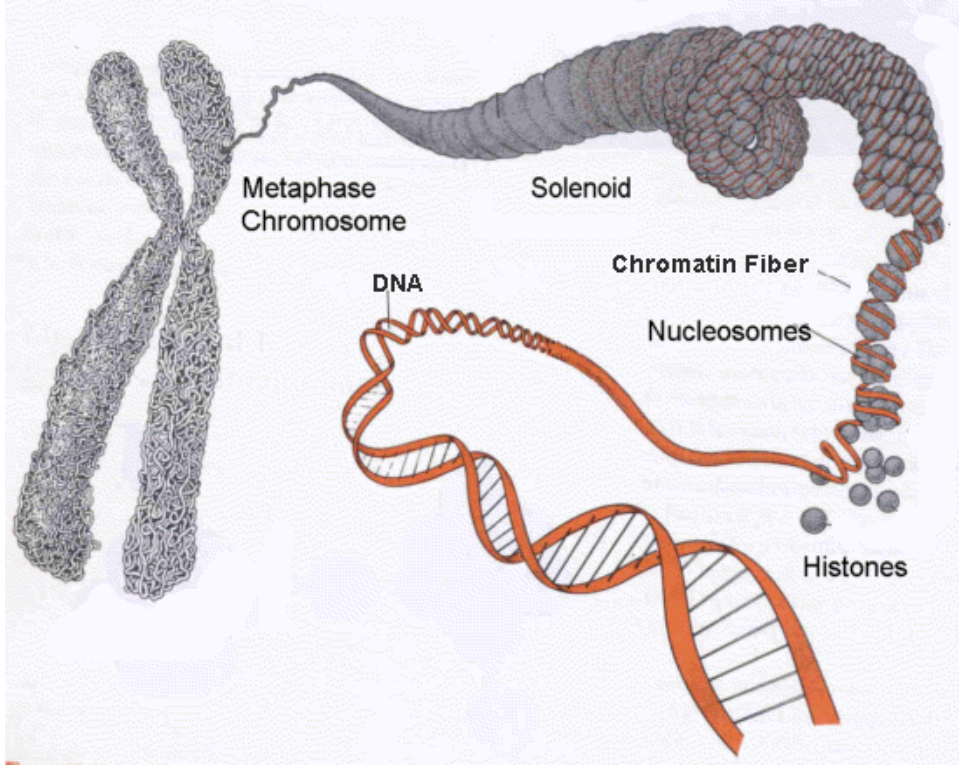
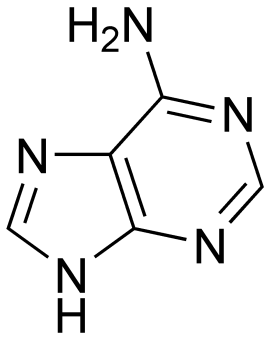


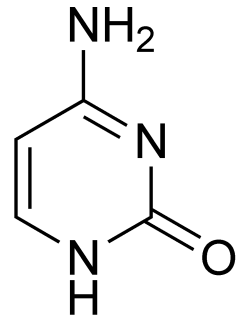
Image credit: http://www.prism.gatech.edu/~gh19/b1510/3chroma.gif

DNA

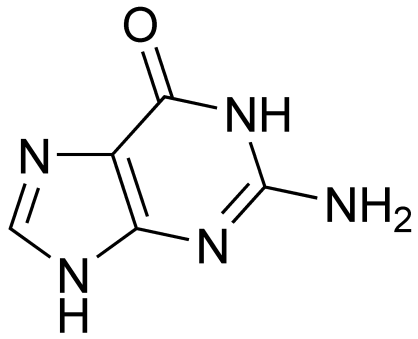
- DNA: DeoxyriboNucleic Acid
  - Two long chains of basic units called nucleotides (bases)
  - Four types of nucleotides:



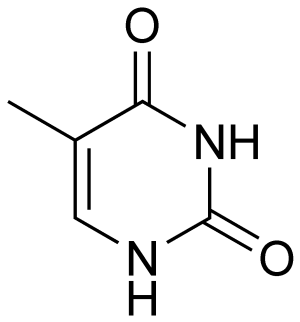
Adenine (A)



Cytosine (C)



Guanine (G)



Thymine (T)

  - C and T have 1 ring, and are called pyrimidines
  - A and G have 2 rings, and are called purines

Image credit: Wikipedia

DNA

- Nucleotides can join together through strong phosphate backbone to form one strand
- Each unit has three components:
  - Nitrogenous base
  - Pentose sugar (ribose)
  - Phosphate
- Different DNA molecules differ only in the base, so we can represent a DNA strand simply by a string with the alphabet {A, C, G, T}

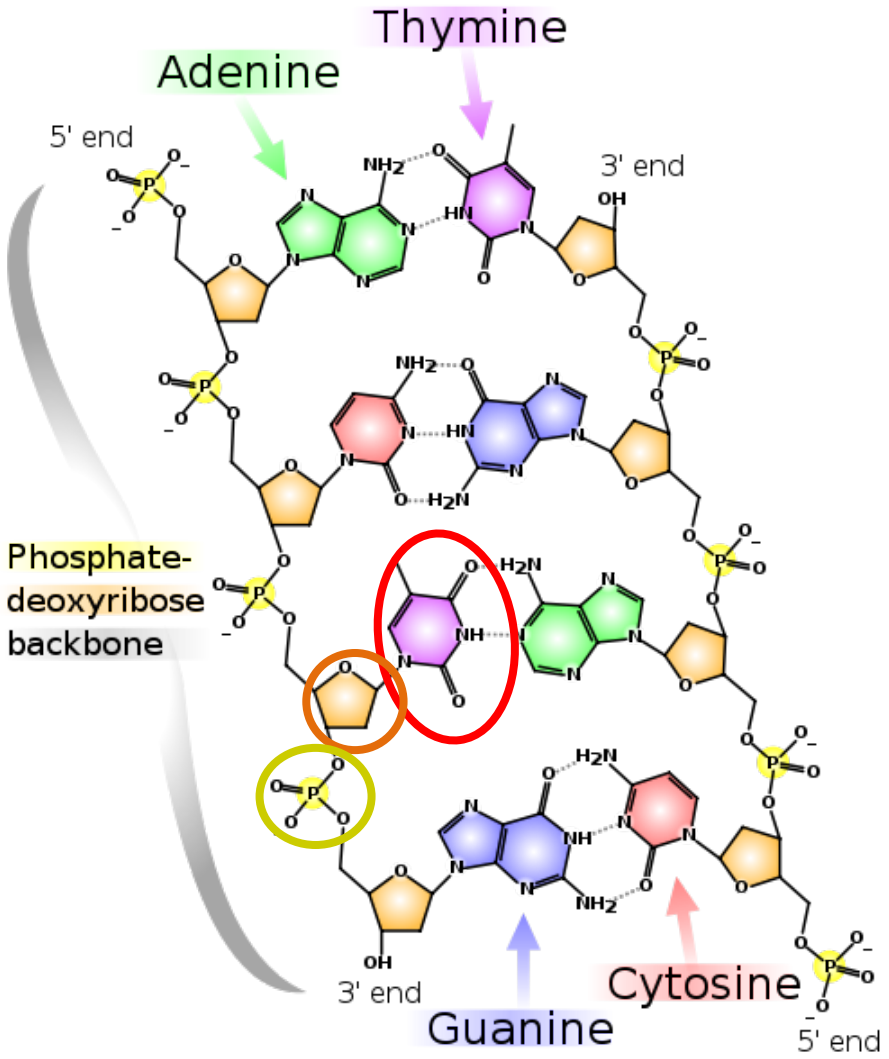
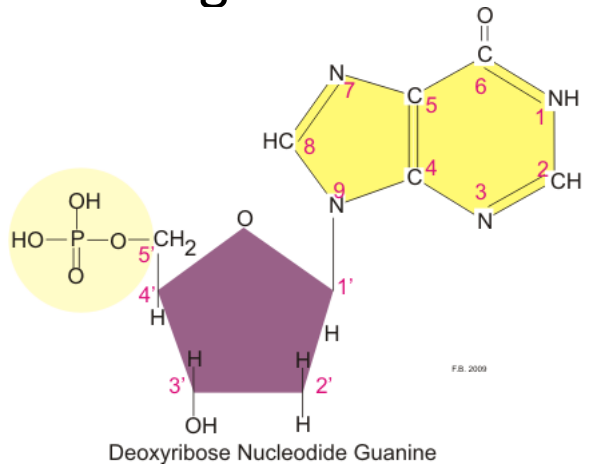


Image credit: Wikipedia

DNA

- The carbon atoms in the pentose sugar are numbered



- When we represent a strand, we go from the 5’ end towards the 3’end
  - Left strand: `ACTG`
  - Right strand: `CAGT`

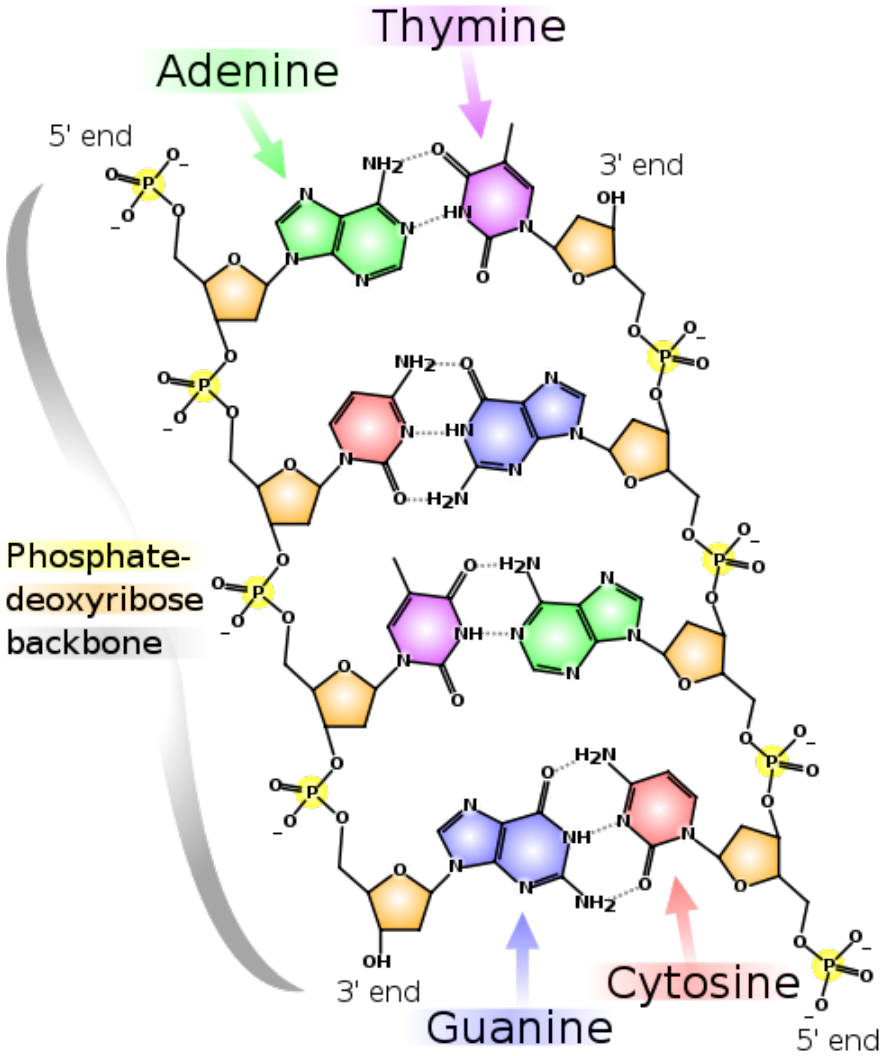


Image credit: Wikipedia

DNA

- Two strands join together through weak hydrogen bonds
  - –A and T can form two hydrogen bonds
  - –C and G can form three hydrogen bonds
  - –(Almost) always true: A paired with T, C paired with G - “reverse complementarity”
  - –When both strands are considered at the same time, the basic unit is a “base pair”

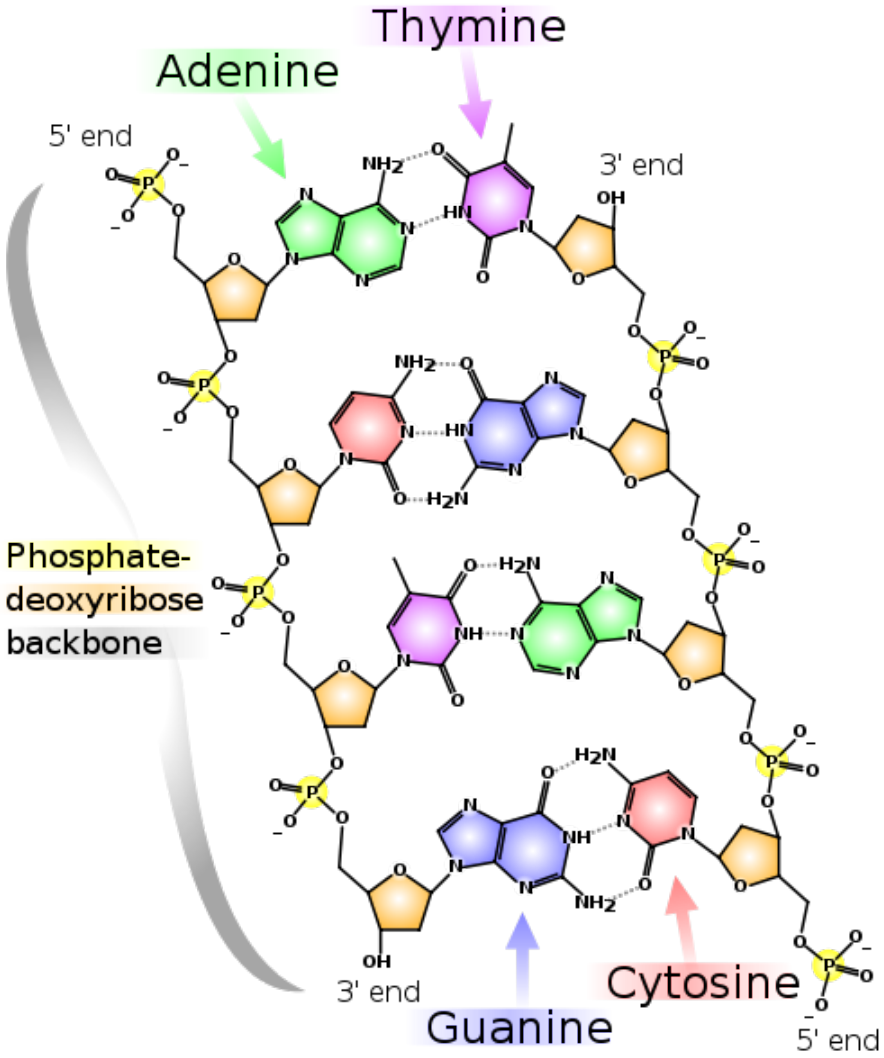


Image credit: Wikipedia

DNA

- The two strands form a double helix structure

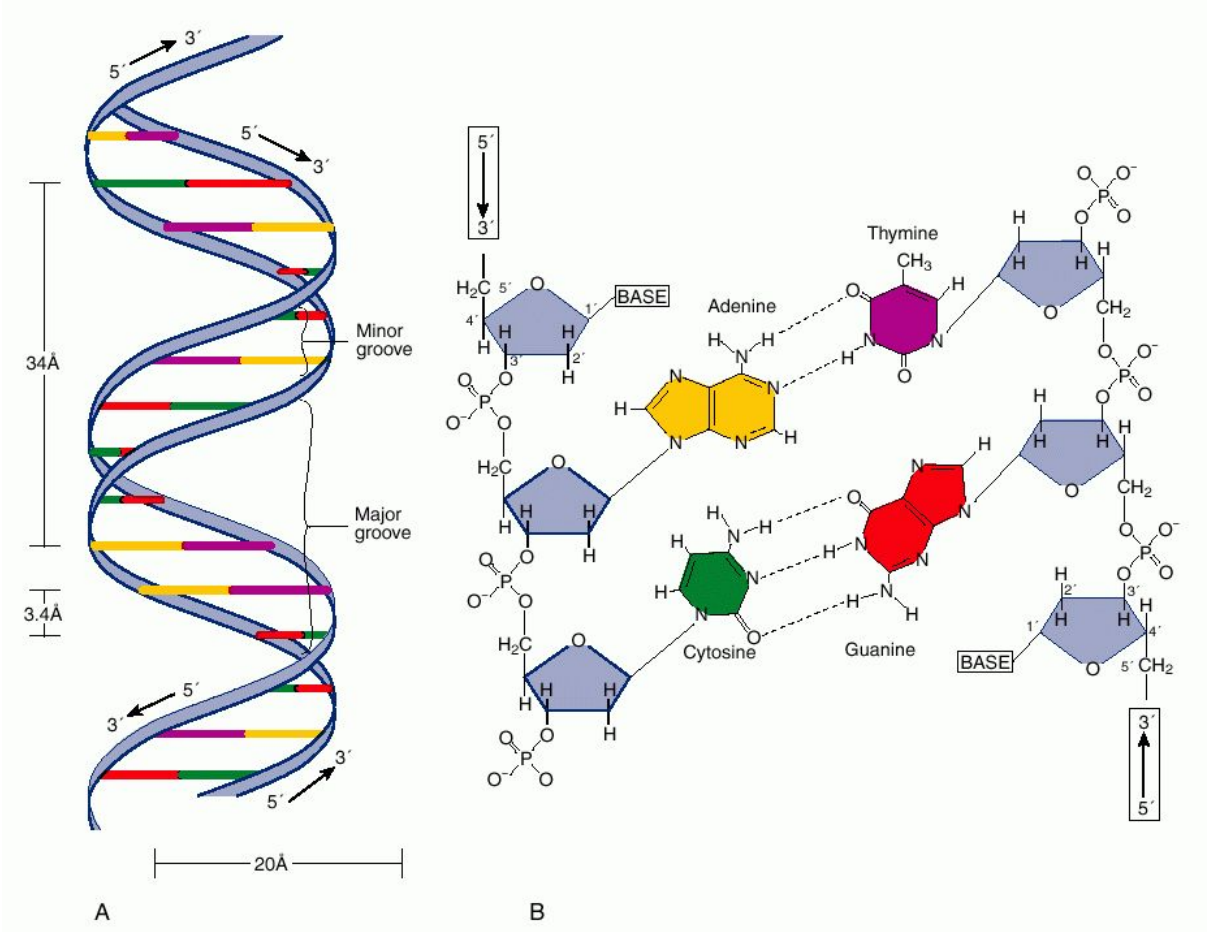


Image credit: http://medical-dictionary.thefreedictionary.com/_/viewer.aspx?path=dorland&name=deoxyribonucleic-acid.jpg

Quick quiz

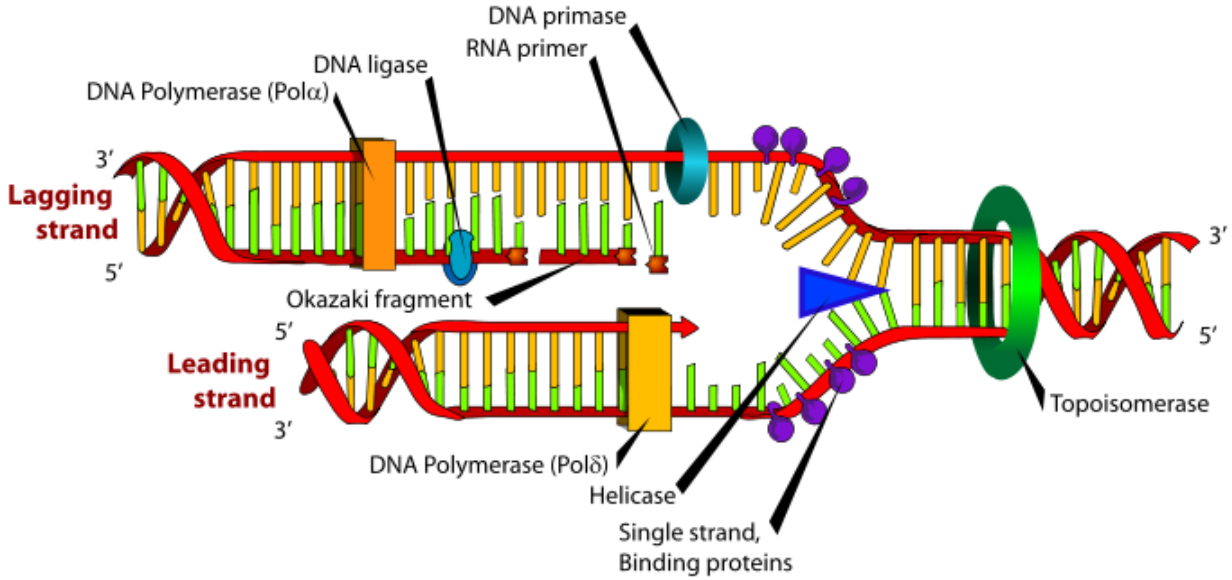
1. If I have ACCGGTC on the forward strand, what do I have on the reverse strand?
   - TGGCCAG
   - If we also consider the orientation, we have the following:

```
        1234567
+ 5′ ACCGGTC 3′
- 3′ TGGCCAG 5′
```

   - It is quite common for biologists to use the 5'-to-3' direction and say the answer is GACCGGT
   - Best to specify both the sequence and the orientation.

DNA replication

- Before a cell divides by mitosis, the two strands serve as templates to build up new DNAs in the daughter cells



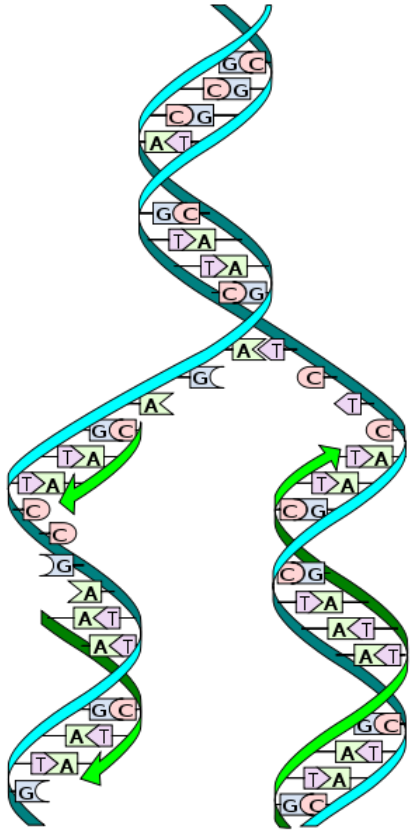


Image credit: Wikipedia

But what does DNA do?

- Frank answer: Nobody completely knows what roles each of the 3 billion base pairs plays
- But: There are some well-studied regions called genes

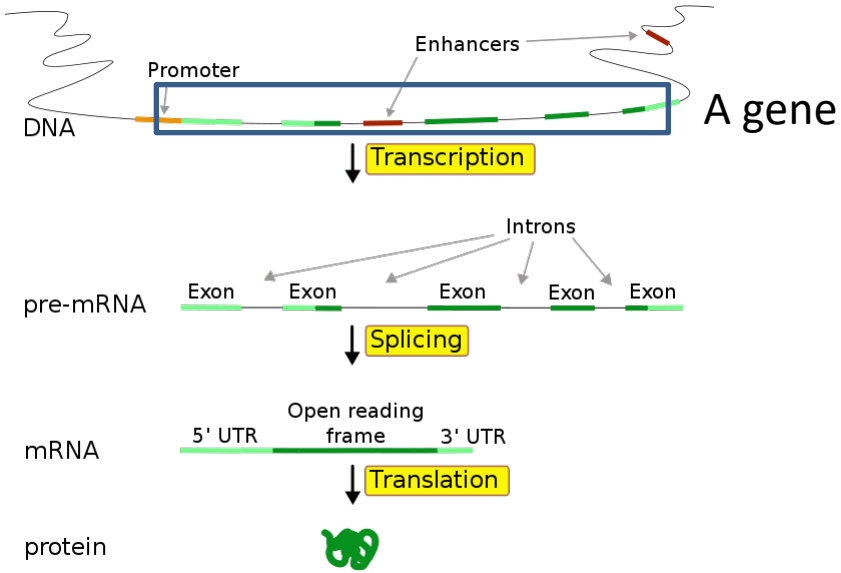


Image credit: Wikipedia

Genes

- Classic view
- "central dogma" of molecular biology:
  - –DNA transcribes to RNA
    - •Transcription
  - –RNA translates to protein
    - •Translation

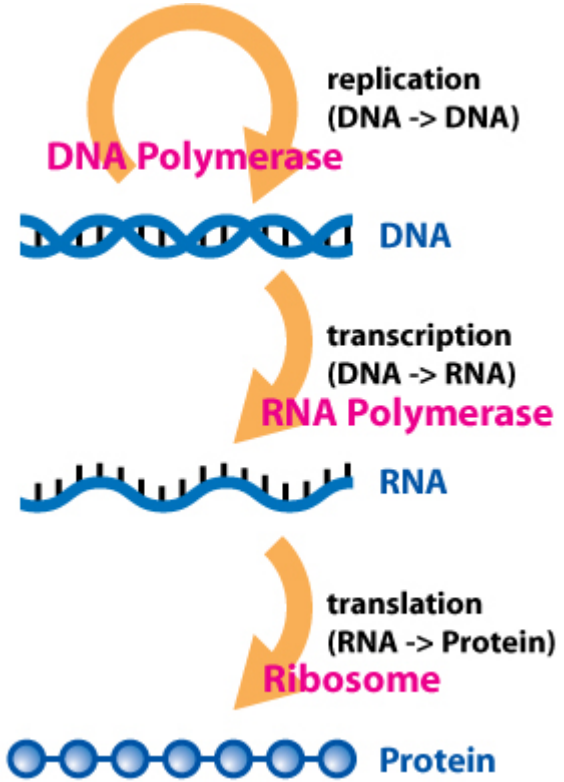


Image credit: Wikipedia

First level: DNA

- Special nucleotide sequences on DNA define different gene regions:
  - Where the transcription machinery (RNA polymerase) should be loaded
  - Where transcription should start
  - Where transcription should end
  - Where the on/off switches (regulatory elements) are

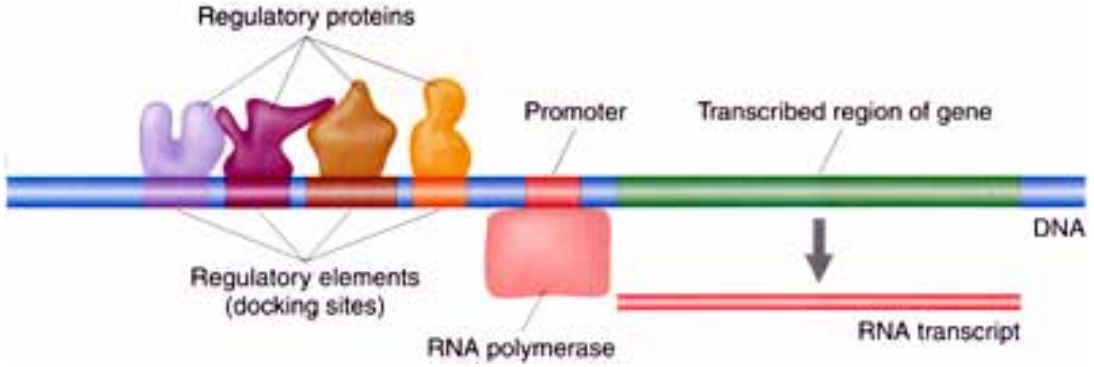
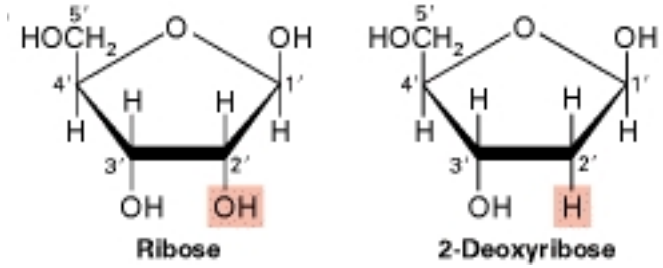


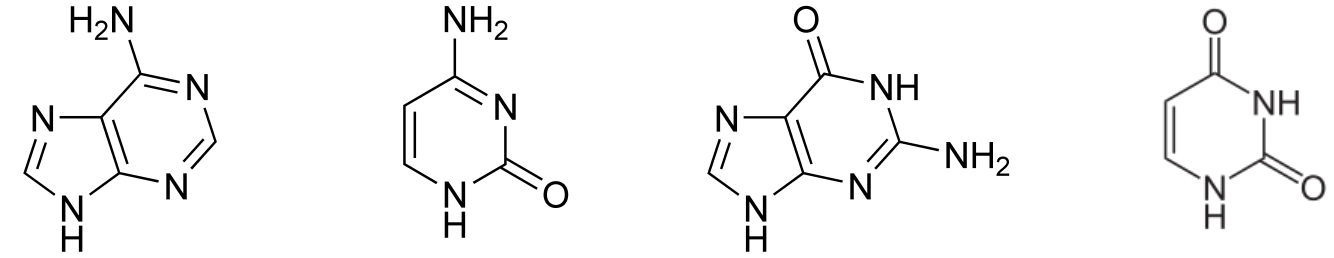
Image credit: http://scienceblogs.com/pharyngula/upload/2007/01/simple_gene_reg.jpg

Second level: RNA

- RNA: Ribonucleic acid
  - –Additional hydroxyl group at 2’ carbon as compared to DNA (that’s why DNA is “deoxy…”)



  - –Also, four types commonly found (note: U instead of T)



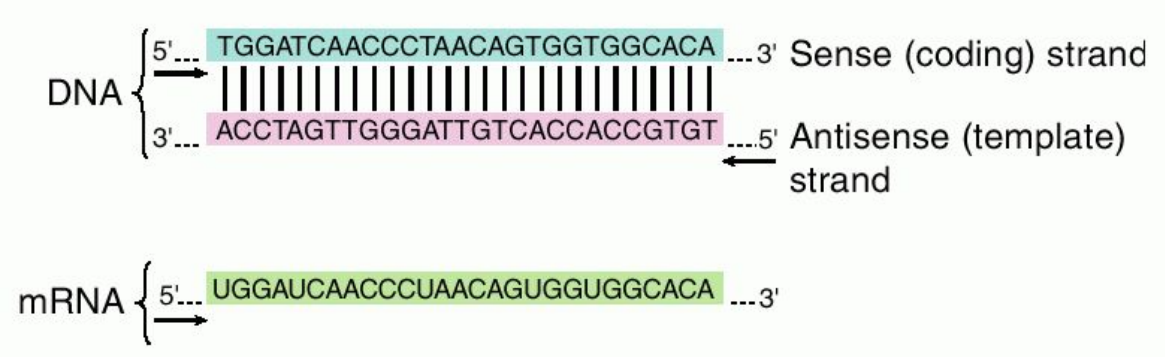
Adenine (A) Cytosine (C) Guanine (G) Uracil (U)

Image credit: http://www.ncbi.nlm.nih.gov/books/NBK21514/bin/ch4f1b.jpg, Wikipedia

DNA to RNA: Transcription

- DNA serves as template. Rule:

Template DNA	Resulting RNA
A	U (not T)
C	G
G	C
T	A



  - Determined according to the template strand
  - "Coding" in "coding strand" means protein coding.

- RNA has only one strand.

Image credit: http://img.tfd.com/dorland/antisense.jpg

Splicing

- For higher organisms, some parts of the RNA called “introns” are spliced, leaving the “exons” in the mature RNA

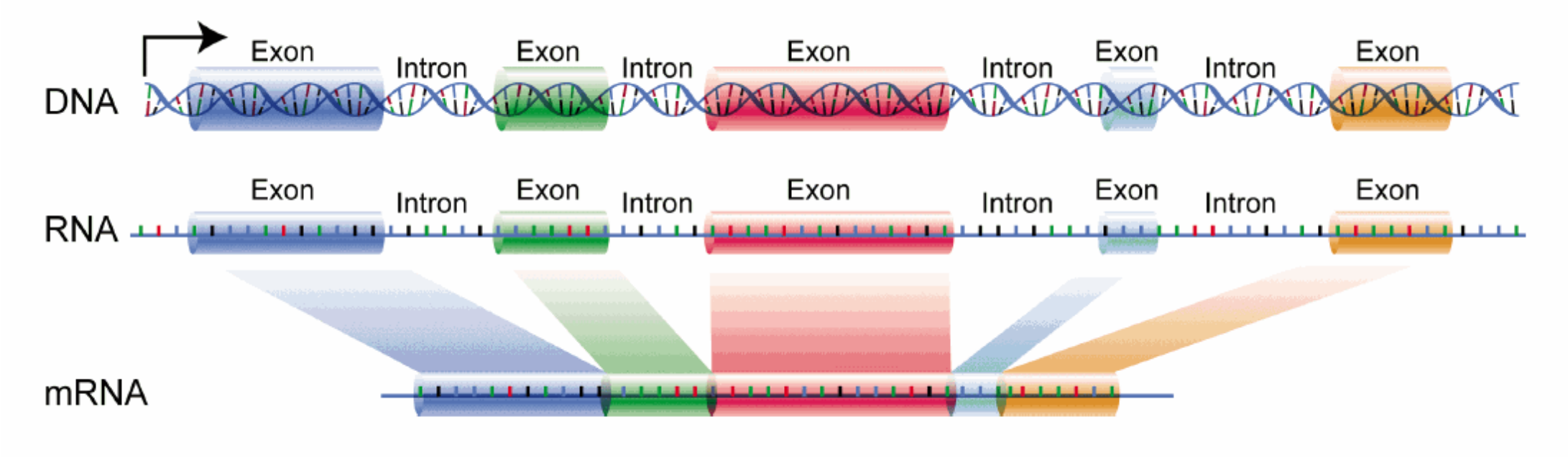


Image credit: Wikipedia

Third level: Protein

- Protein: A chain of amino acids, folded into a particular structure
- Amino acid: 20 common types, all with three components:
  - –Amine group
  - –Carboxylic acid group
  - –Side chain
- The 20 types only differ in the side chain

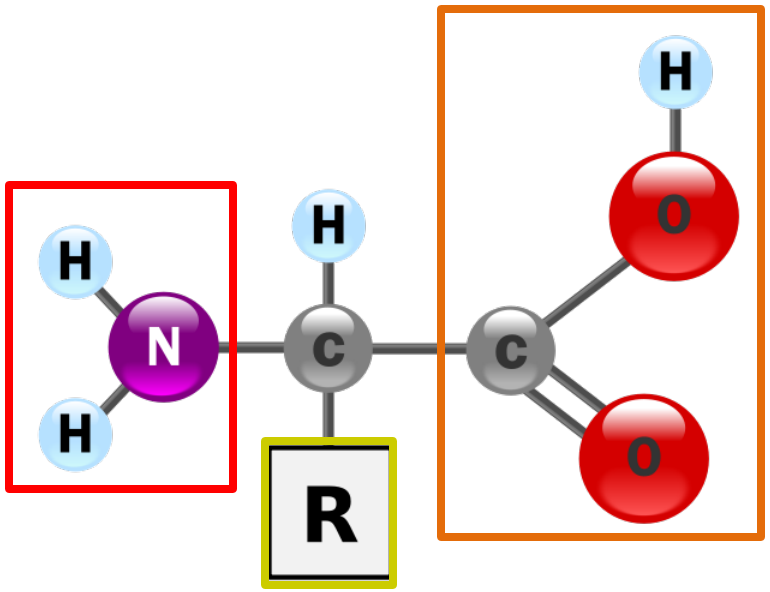
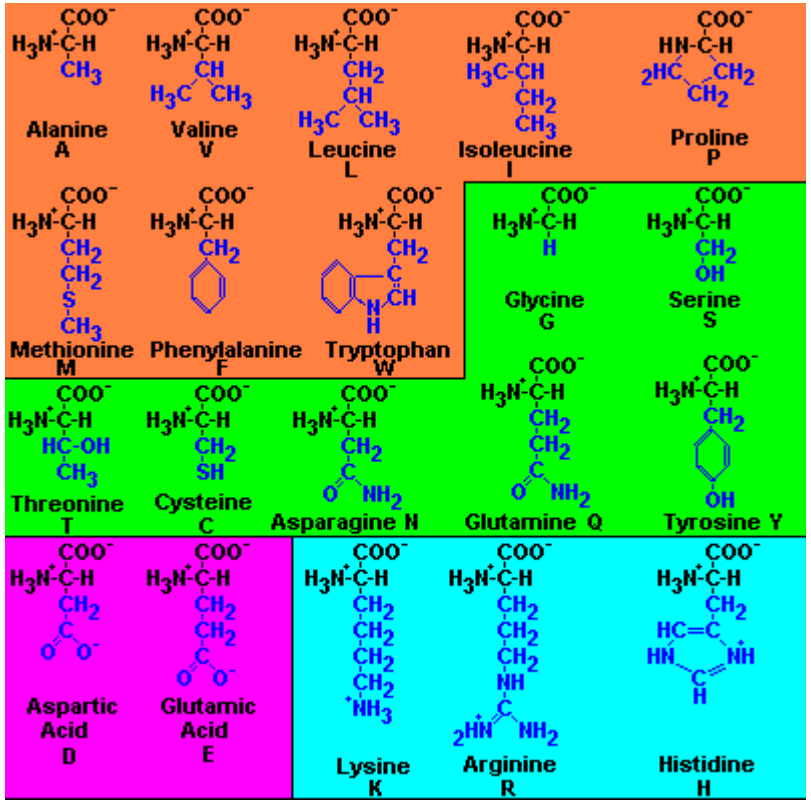


Image credit: Wikipedia

Amino acids

- The 20 common types (side chains in blue):



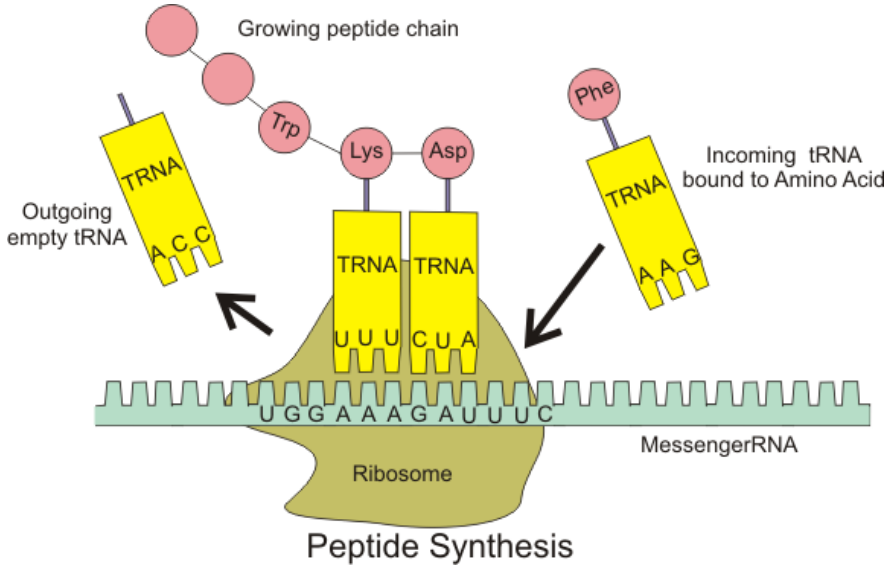
A protein can be represented by a string with the alphabet {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}

- Which 6 are missing?
- B, J, O, U, X, Z

Image credit: http://www.molecularstation.com/molecular-biology-images/data/510/AminoAcids.gif

RNA to protein: Translation

- RNA enters a big machinery (the ribosome), free amino acids assemble into a chain according to the RNA sequence
  - These RNAs deliver messages from DNA to protein, that's why they are called "messenger RNAs" (mRNAs)
  - Again, some signals determine where translation should start and where to stop. The remaining parts are called the "untranslated regions" (UTRs)



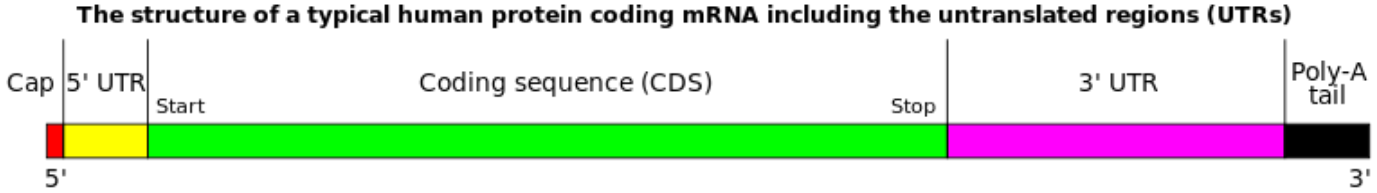


Image credit: http://www.eurekadiscoveries.com/wp-content/uploads/2010/06/Peptide_syn.png, Wikipedia

Coding table

- How to determine which amino acid to add?
  - –Every three nucleotides form a unit called “codon”
  - –The amino acid to add is based on the codon
  - –Note: start/stop, redundancy

1st base	2nd base U		2nd base C		2nd base A		2nd base G	
U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine
	UUC	(Phe/F) Phenylalanine	UCC	(Ser/S) Serine	UAC	(Tyr/Y) Tyrosine	UGC	(Cys/C) Cysteine
	UUA	(Leu/L) Leucine	UCA	(Ser/S) Serine	UAA	Stop (*Ochre*)	UGA	Stop (*Opal*)
	UUG	(Leu/L) Leucine	UCG	(Ser/S) Serine	UAG	Stop (*Amber*)	UGG	(Trp/W) Tryptophan
C	CUU	(Leu/L) Leucine	CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine
	CUC	(Leu/L) Leucine	CCC	(Pro/P) Proline	CAC	(His/H) Histidine	CGC	(Arg/R) Arginine
	CUA	(Leu/L) Leucine	CCA	(Pro/P) Proline	CAA	(Gln/Q) Glutamine	CGA	(Arg/R) Arginine
	CUG	(Leu/L) Leucine	CCG	(Pro/P) Proline	CAG	(Gln/Q) Glutamine	CGG	(Arg/R) Arginine
A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine
	AUC	(Ile/I) Isoleucine	ACC	(Thr/T) Threonine	AAC	(Asn/N) Asparagine	AGC	(Ser/S) Serine
	AUA	(Ile/I) Isoleucine	ACA	(Thr/T) Threonine	AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine
	AUG[A]	(Met/M) Methionine	ACG	(Thr/T) Threonine	AAG	(Lys/K) Lysine	AGG	(Arg/R) Arginine
G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine
	GUC	(Val/V) Valine	GCC	(Ala/A) Alanine	GAC	(Asp/D) Aspartic acid	GGC	(Gly/G) Glycine
	GUA	(Val/V) Valine	GCA	(Ala/A) Alanine	GAA	(Glu/E) Glutamic acid	GGA	(Gly/G) Glycine
	GUG	(Val/V) Valine	GCG	(Ala/A) Alanine	GAG	(Glu/E) Glutamic acid	GGG	(Gly/G) Glycine

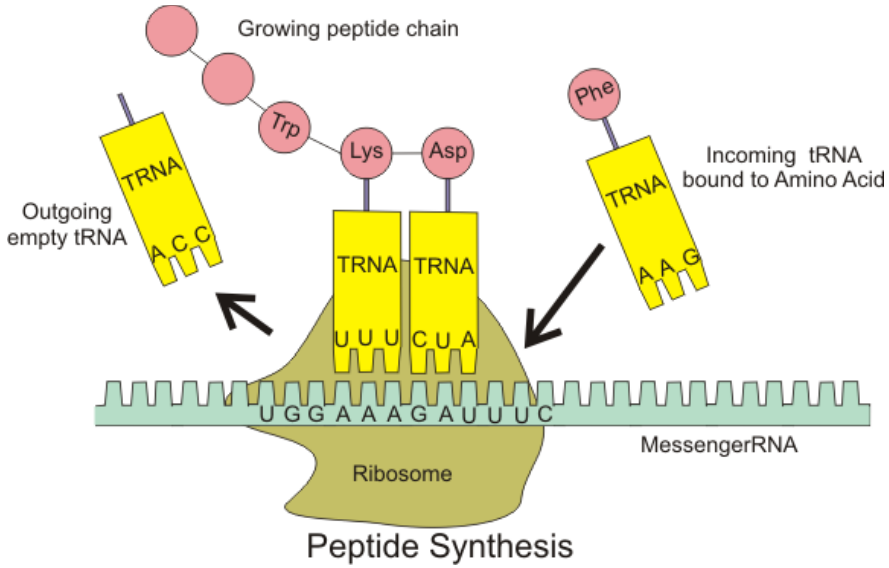


Image credit: Wikipedia

The whole process

- Now the meaning of “coding strand” is clear: the final amino acid sequence can be read out from the coding strand

- Note:
  - Not all RNAs are translated. Those do not are called non-coding RNAs (ncRNAs)
  - When two amino acids join together to form a peptide bond, a water molecule is expelled. Therefore, the remaining is called a “residue”

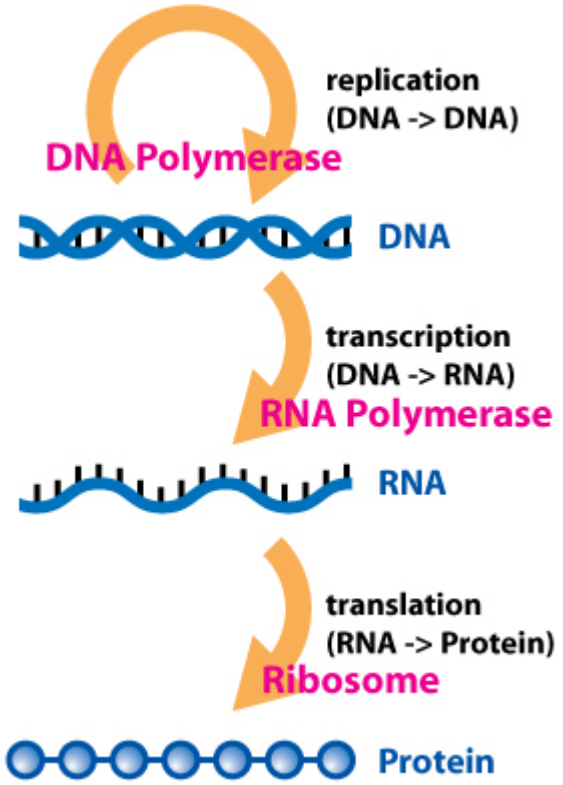


Image credit: Wikipedia

Coding and template strands revisited

- If we specify the sequence of a gene, we always specify its sequence on the coding strand

1st base	2nd base: U		2nd base: C		2nd base: A		2nd base: G	
U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine
	UUC	(Phe/F) Phenylalanine	UCC	(Ser/S) Serine	UAC	(Tyr/Y) Tyrosine	UGC	(Cys/C) Cysteine
	UUA	(Leu/L) Leucine	UCA	(Ser/S) Serine	UAA	Stop (*Ochre*)	UGA	Stop (*Opal*)
	UUG	(Leu/L) Leucine	UCG	(Ser/S) Serine	UAG	Stop (*Amber*)	UGG	(Trp/W) Tryptophan
C	CUU	(Leu/L) Leucine	CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine
	CUC	(Leu/L) Leucine	CCC	(Pro/P) Proline	CAC	(His/H) Histidine	CGC	(Arg/R) Arginine
	CUA	(Leu/L) Leucine	CCA	(Pro/P) Proline	CAA	(Gln/Q) Glutamine	CGA	(Arg/R) Arginine
	CUG	(Leu/L) Leucine	CCG	(Pro/P) Proline	CAG	(Gln/Q) Glutamine	CGG	(Arg/R) Arginine
A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine
	AUC	(Ile/I) Isoleucine	ACC	(Thr/T) Threonine	AAC	(Asn/N) Asparagine	AGC	(Ser/S) Serine
	AUA	(Ile/I) Isoleucine	ACA	(Thr/T) Threonine	AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine
	AUG[A]	(Met/M) Methionine	ACG	(Thr/T) Threonine	AAG	(Lys/K) Lysine	AGG	(Arg/R) Arginine
G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine
	GUC	(Val/V) Valine	GCC	(Ala/A) Alanine	GAC	(Asp/D) Aspartic acid	GGC	(Gly/G) Glycine
	GUA	(Val/V) Valine	GCA	(Ala/A) Alanine	GAA	(Glu/E) Glutamic acid	GGA	(Gly/G) Glycine
	GUG	(Val/V) Valine	GCG	(Ala/A) Alanine	GAG	(Glu/E) Glutamic acid	GGG	(Gly/G) Glycine

DNA

Coding strand: 5′–CGACATGGAGGGTCCAGTGAAATGCTATTAACGTG–3′

Template strand: 3′–GCTGTACCTCCCAGGTCACTTTACGATAATTGCAC–5′

RNA

Pre-mRNA: 5′–CGACAUGGAGGGUCCAGUGAAAUGCUAUUAACGUG–3′

Mature mRNA: 5′–CGACAUGGAGG UGAAAUGCUAUUAACGUG–3′

Amino acids: NH3–M E V K C Y *–COOH

Key:
Intron
Exons
Untranslated regions (UTRs)
Coding sequence (CDS)

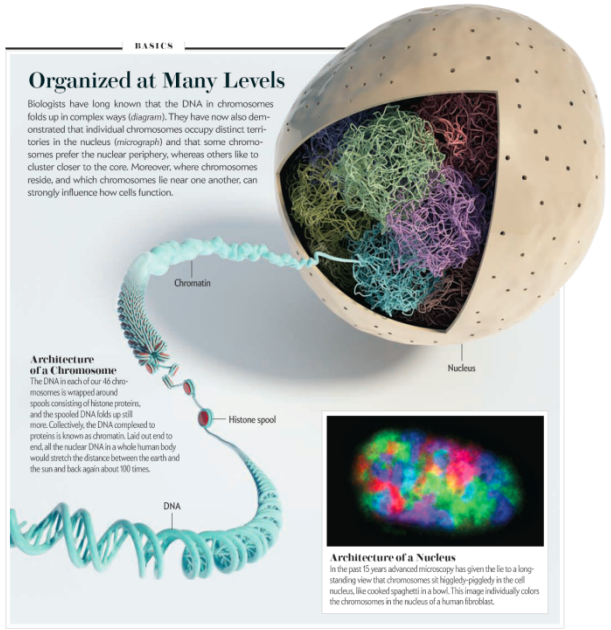
Image credit: Wikipedia

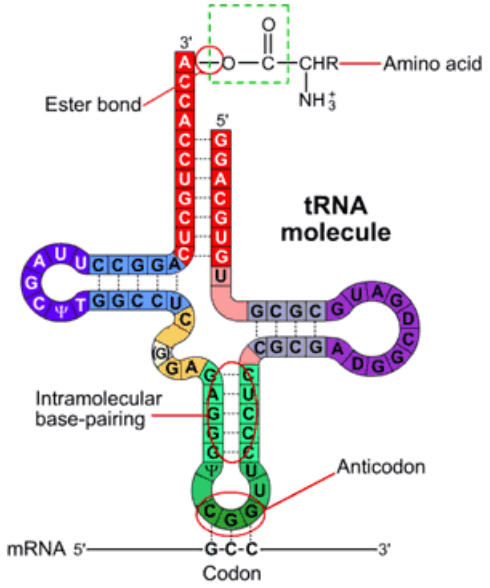
Coding and template strands revisited

- What information do we need to fully identify a genomic location?
  - Chromosome, position, strand
- What information do we need to fully identify a genomic interval (e.g., a gene)?
  - Chromosome, start position, end position, strand
- Note: Most biologists implicitly assume a “1-based, both sides inclusive” indexing scheme
  - Which means the first position is counted as 1, and chr1:10-20 means the tenth to twentieth positions (11 positions/nucleotides/base pairs in total)
  - We will also assume this indexing scheme except when we deal with some particular file formats

Structures

- RNA and proteins are not simply long chains of molecules. Like DNA, they are highly structured.
- Function is related to structure.





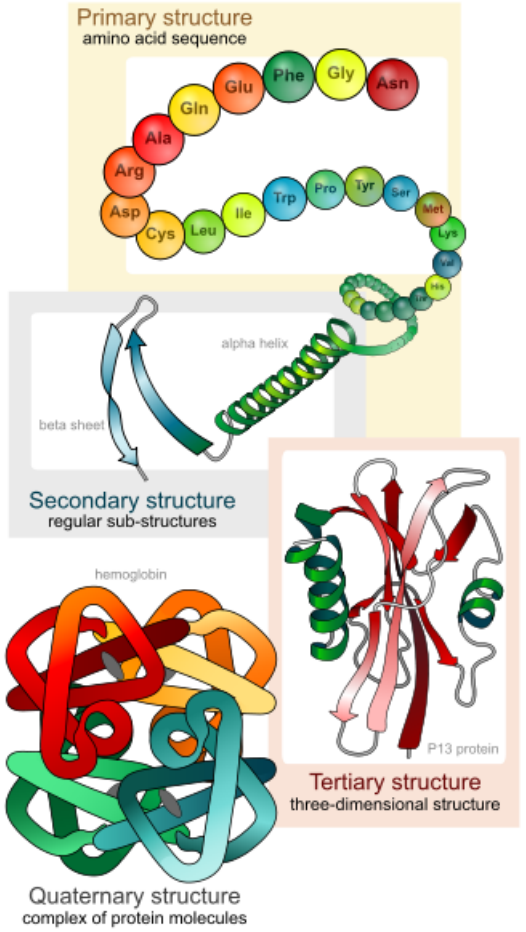


Image credit: Scientific American, http://www.wiley.com/college/boyer/0470003790/structure/tRNA/trna_diagram.gif, Wikipedia

Part 4

DATA IN BIOINFORMATICS

Two approaches to biological research

- Traditionally, biologists study in detail a small number of objects at a time
  - Hypothesis-driven
  - Bottom-up approach

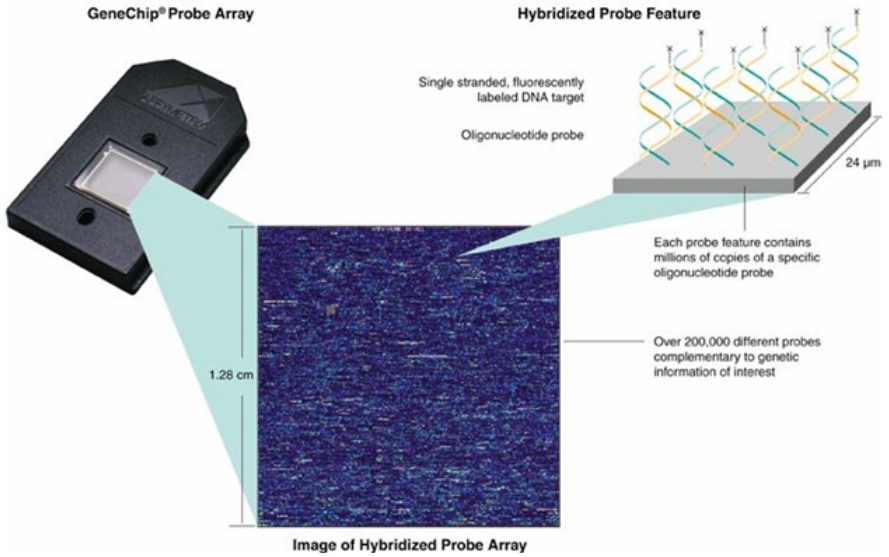
Two approaches to biological research

- Traditionally, biologists study in detail a small number of objects at a time
  - –Hypothesis-driven
  - –Bottom-up approach

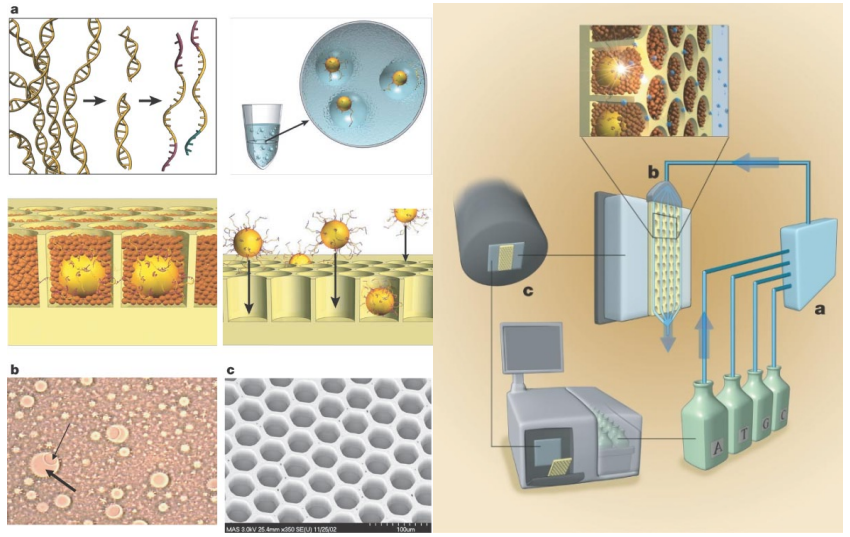
- An alternative approach is to generate a lot of experimental data, identify interesting patterns, and pick some to study further
  - –Data-driven
  - –Top-down approach

Data, data and data

- The second approach was driven by technologies that allow for the production of an enormous amount of data in a short time
  - We will study some of them in more detail in this course



Microarrays

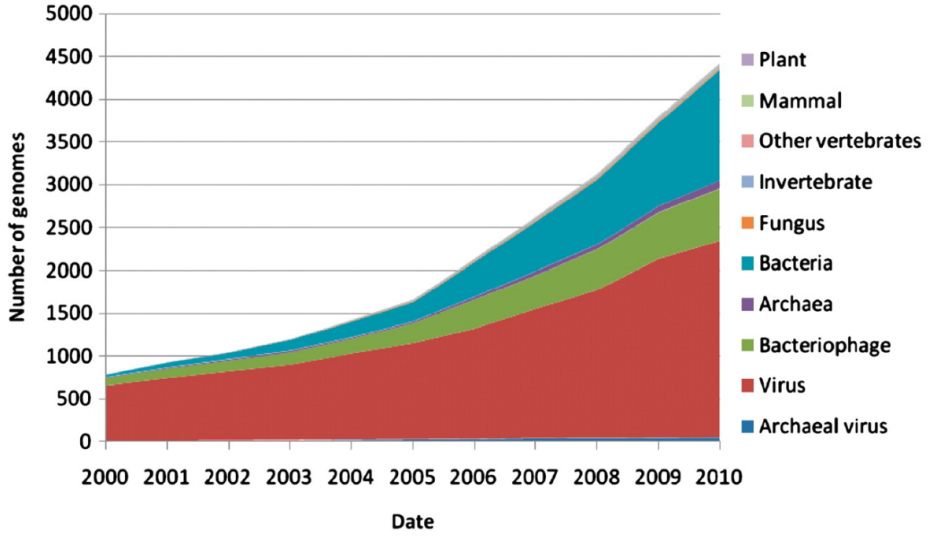


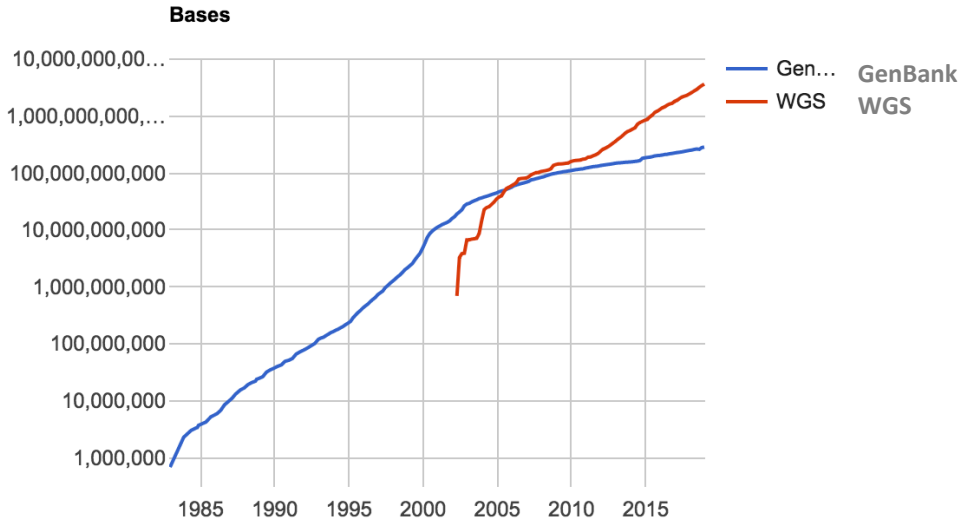
Massively parallel sequencing

Image credit: Affymetrix, Margulies et al., *Nature* 437:376-380, (2005)

Statistics

- Number of nucleotides and sequences in GenBank, and number of complete genomes (WGS: whole-genome shotgun)





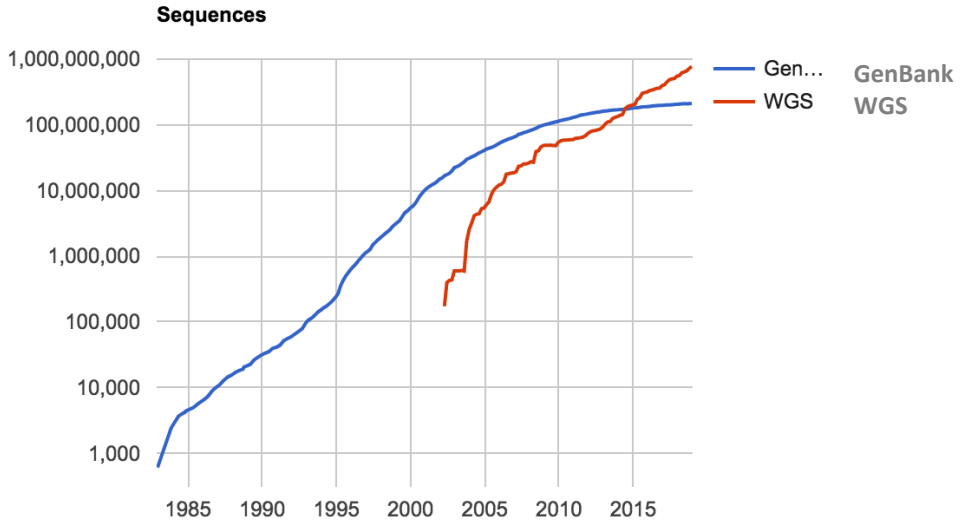


Image credit: Cochrane et al., *Nucleic Acids Research* 39(S1):D15-D18, (2010), http://www.ncbi.nlm.nih.gov/genbank/statistics

Milestones in genome sequencing

Genome	Type	Size	Completed year	Time needed	Cost (USD)
Bacteriophage MS2	Virus (RNA)	3,569nt	1976	?	?
Bacteriophage ΦX174	Virus (DNA)	5,368bp	1977	?	?
Haemophilus influenzae	Bacteria	1.8Mb	1995	?	?
Saccharomyces cerevisiae	Fungus (yeast)	12.1Mb	1996	?	?
Caenorhabditis elegans	Nematode (worm)	100Mb	1998	?	?
Arabidopsis thaliana	Plant	157Mb	2000	?	?
Homo sapiens	Mammal (human)	3.2Gb	2003	15 years	3B
Craig Venter	Mammal (human)	2.8Gb	2007	5 years	100M
James Watson	Mammal (human)	6Gb (diploid)	2008	4 months	1.5M
YanHuang 1 (Chinese)	Mammal (human)	~3Gb	2008	2 months	0.5M
Neanderthal	Mammal	3.2Gb	2010	4 years	6.4M
Anyone	Mammal (human)	~3Gb	2011	1 week	10K
Anyone (30x coverage)	Mammal (human)	~3Gb	2020	<1 week	<1K

International endeavors

Consortium	Purpose
The Human Genome Project (HGP)	Sequence the human genome
The International HapMap Project	Develop haplotype map of the human genome
Encyclopedia of DNA Elements (ENCODE)	Catalog and characterize human DNA elements
Model Organism Encyclopedia of DNA Elements (modENCODE)	Catalog and characterize model organism DNA elements
1000 Genomes Project	Identify most genetic variants with at least 1% frequencies
The Cancer Genome Atlas (TCGA)	Build an atlas of genomic changes in cancer genomes
...	...

Image credit: HGP, HapMap, ENCODE, modENCODE, TCGA

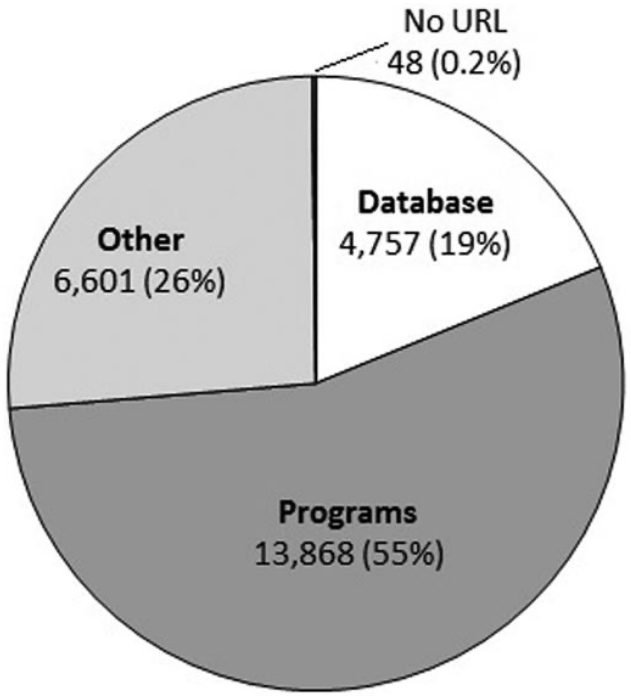
Classification

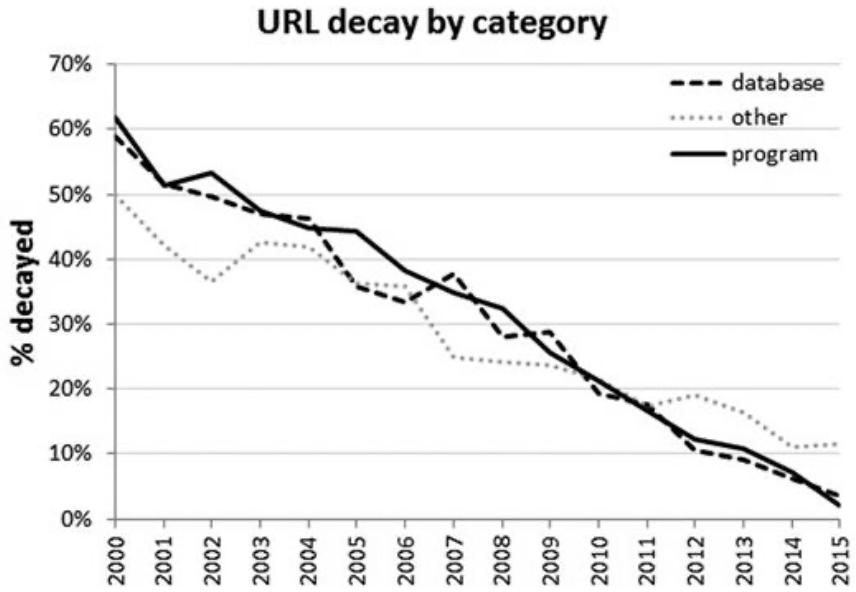
a. By data type
   - Sequences
   - Annotations
   - Motifs and domains
   - Variations
   - Phylogenies trees
   - Structures
   - Expression
   - Networks
   - Publications

b. By lecture materials
   - Data representation and file formats
   - Data origin and acquisition
   - Databases and tools

We have data, then...?

- Many databases and analysis tools developed





–We need to ensure good quality and availability

Image credit: Wren et al., *Nucleic Acids Research* 45(7):3627-3633, (2017)

Working with data

- Some examples
  - Small-scale (tens to thousands of points)
    - Data: Sequence of a gene, a protein structure, a microarray dataset, ...
    - Tools: Excel, R, Matlab, ...
  - Medium-scale (thousands to millions of points)
    - Data: SNP list of a genome, a protein-protein interaction network, simulation trace of the molecular motions of a small protein, ...
    - Tools: Perl, Python, Java, ...
  - Large-scale (millions to billions of points)
    - Data: Raw sequencing reads, whole-genome alignment of 10 species, global ocean survey, ...
    - Tools: C, Oracle, parallelized and tailor-made software, ...

Epilogue

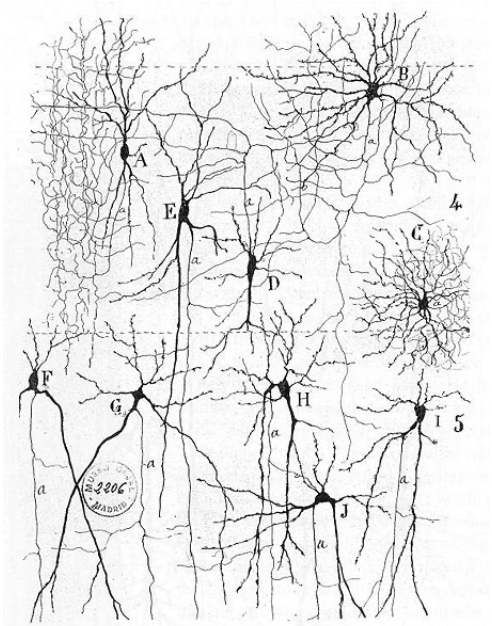
CASE STUDY, SUMMARY AND FURTHER READINGS

Case study: Computer and Biology

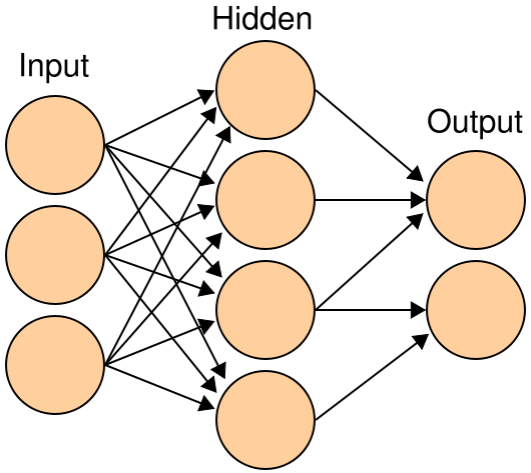
- Computer and biology are related in multiple ways:
  - Computer can help study biology
    - Bioinformatics, computational biology
  - Computer algorithms are inspired by biology
  - Computational problems can be solved by biology
  - New biological systems can be designed according to principles used in computer systems

Case study: Computer and Biology

- Computational algorithms inspired by biology
- Artificial neural network: A network of mathematical functions for modeling some complex concepts (e.g., text recognition)



A biological neural network



An artificial neural network

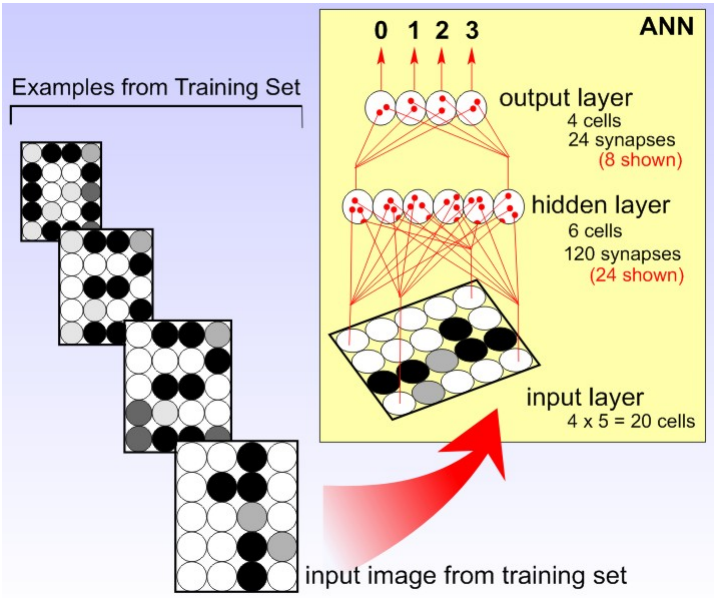


Image sources: http://upload.wikimedia.org/wikipedia/en/thumb/1/1a/Cajal_actx_inter.jpg/456px-Cajal_actx_inter.jpg, http://upload.wikimedia.org/wikipedia/commons/thumb/e/e4/Artificial_neural_network.svg/560px-Artificial_neural_network.svg.png, http://www.highlights-in-neurobiology.com/wp-content/uploads/2013/03/neural-network-new.jpg

Summary

- **Bioinformatics**
  - Using computational methods to assist biomedical research
  - Large data size
  - Difficult computational problems
- **Revisit genetics and molecular biology**
- **There are many data types in bioinformatics, and a huge amount of data produced**
- **Further readings:**

  Chapter 1 of *Algorithms in Bioinformatics: A Practical Introduction*
  - More comprehensive introduction of the basic concepts
  - Free slides: https://www.comp.nus.edu.sg/~ksung/algo_in_bioinfo/slides/Ch1_intro.pdf

Thanks for your attention!