Lecture 1. Introduction

The Chinese University of Hong Kong

CSCI3220 Algorithms for Bioinformatics

Lecture outline

1. Course information
2. Introduction to bioinformatics

(Intermission: Background survey)

3. Introduction to genetics and molecular biology
4. Overview of class topics

Part 1

COURSE INFORMATION

Course objectives

- To have fun
- To learn what bioinformatics is about
  - Hopefully, to arouse your interests in this area
- To study some useful algorithms
  - These algorithms are by themselves interesting and fundamentally important
  - To see how theoretical algorithms are used in real-life applications

Teaching staff

- Lecturer
  - Dr. Kevin Yip
    kevinyip@cse.cuhk.edu.hk
    Consultation hours (please make appointment by email):
    - Tuesdays 14:00-16:00
- Teaching assistants
  - Ms. Yizhen Chen (Channie)
    yzchen@cse.cuhk.edu.hk
  - Mr. Zhenghao Zhang (Adam)
    zhzhang@cse.cuhk.edu.hk

Lectures

1. Tuesdays 16:30 – 17:15
2. Thursdays 12:30 – 14:15

- Focusing on concepts and short examples/exercises
- We will have various interactive activities to keep you awake

Tutorials

- Weekly, mainly to go over some exercises together
- Time: Tue 17:30-18:15

Section	Time	Mon	Tue	Wed	Thu	Fri
1	08:30-09:15					
2	09:30-10:15					
3	10:30-11:15					
4	11:30-12:15					
5	12:30-13:15				Lecture	
6	13:30-14:15					
7	14:30-15:15		Kevin's consultation hours			
8	15:30-16:15					
9	16:30-17:15		Lecture			
10	17:30-18:15		Tutorial			
11	18:30-19:15					

Course Web sites

- Course Web site: http://www.cse.cuhk.edu.hk/~kevinyip/csci3220/
  - Lecture notes
  - Information of all the following
- Blackboard (https://blackboard.cuhk.edu.hk/, look for course 2020R1-CSCI3220)
  - Tutorial notes
  - Assignment specifications
  - Assignment collection boxes
  - Lecture recordings – Panopto
  - Announcements – check your link.cuhk.edu.hk email account
- Piazza (https://piazza.com/cuhk.edu.hk/fall2020/csci3220/home, registration required)
  - Questions and discussions
- uReply (http://ureply.mobi/)
  - Interactive tasks
- Online judge system (http://proj.cse.cuhk.edu.hk/csci3220/?locale=en, accessible from CUHK network including VPN)
  - Program submissions
- YouTube channel (https://www.youtube.com/channel/UCk2ozjkbfpteeJUolHWNsMw)
  - Micro-modules on some topics

Reference materials

- No textbooks
- Lecture notes can be downloaded from course Web site
- *Jot your own notes in class*
- References:
  - *Algorithms in Bioinformatics: A Practical Introduction* by Wing-Kin Sung, Chapman & Hall 2009 (with free online materials)
  - *An Introduction to Bioinformatics Algorithms* by Neil C. Jones and Pavel A. Pevzner, MIT Press 2004

Assessment

- Assignments 45%
  - Tentatively 5 of them in total
- Class participation 5%
  - uReply
- Midterm examination 20%
  - During the class on Oct 29
  - Open book, open notes
- Final examination 30%

Promises

- Putting up lecture notes in time
- Suitable teaching pace and level of difficulty
  - Feedback is crucial
- Quick responses to questions
- Prompt and fair grading of assignments

Expectations

- Faculty of Engineering Staff-Student 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?

- Answer #1: Definitions
  - 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?

- Answer #2: My own experience
  - Someone: What is your research area?
  - Kevin: Bioinformatics
  - Someone: Bio...in...? What's that?
  - Kevin: Using computing methods to assist biomedical research

Why do we need bioinformatics?

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

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 (each is called a "haploid genome")
- If we represent DNA as a string with four letters, `A`, `C`, `G` and `T`…

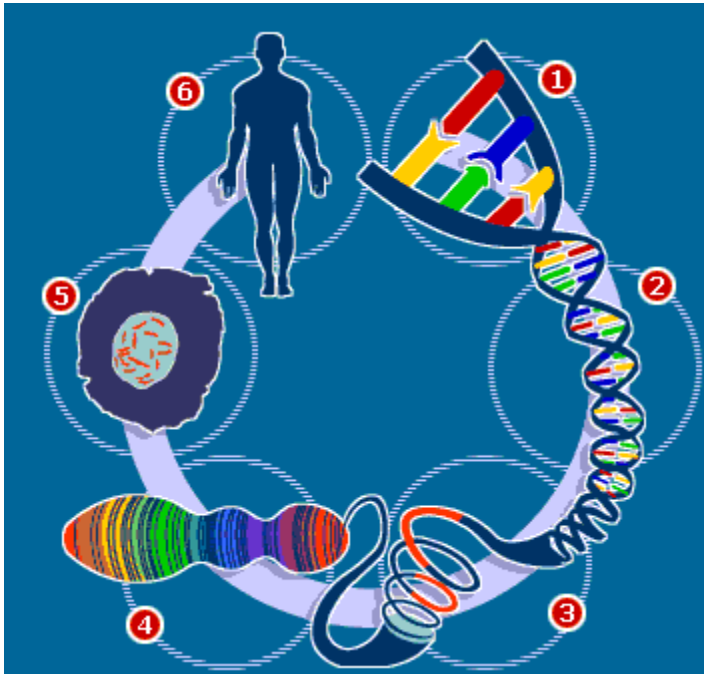


Image credit: news.bbc.co.uk

CTCACCGTCGTAAATCTATGATCTGGCTTGGCCTGCAGT
AGCTCTTTCATTTCGGGCTTATCTAATGCTGACTGGTCG
GTCCTGGCTACGCTCCAAAACGTACGTATTCGGGCCATC
GAGGCTAGCGGCACTTCGAGCGATCTATCGGGAGCTTTG
GCTATCGATCGGGCGATCGATGCTGACGTACGTAGCGCG
CGATCGAGCGCGGCTAGCTAGCGGCATCGTAGCTACGTA
GCTACGGCGCTATTTCGATCGAGTCGTGTCTAGTCGGAT
ATAGCTATGCATCTAGCTGAGGCATCTGAGCGGATCGAT
GCTAGGGCGATCGGAGCTAGCTGAGCTAGCTAGCTGAGC
GCTAGCGAGCGTACGAGCGATCGAGCGAGTCTAGCGAGC
GATTCTAGCGATATACATTAGCCCGATCGTATGCTAGCT
AGGGCTAGCATGCGGATCTATCGAGCGGCTATCTGAGCG
ATTCGATCGAGCGATCTAGCGAGCTATCGATCGAGCCGG

Large data size

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

- To study a certain trait (e.g., a disease), we may need to study the DNA of thousands to millions of people

Deep sequencing of 10,000 human genomes

Amalio Telenti, Levi C. T. Pierce, William H. Biggs, Julia di Iulio, Emily H. M. Wong, Martin M. Fabani, Ewen F. Kirkness, Ahmed Moustafa, Naisha Shah, Chao Xie, Suzanne C. Brewerton, Nadeem Bulsara, Chad Garner, Gary Metzker, Efren Sandoval, Brad A. Perkins, Franz J. Och, Yaron Turpaz, and J. Craig Venter

PNAS October 18, 2016 113 (42) 11901-11906; first published October 4, 2016
https://doi.org/10.1073/pnas.1613365113

- Humans have 20,000-25,000 genes that produce proteins
  - We want to study their pair-wise and higher-order relationships
  - About 3.1×10^8 pairs, 2.6×10^{12} triples, ...

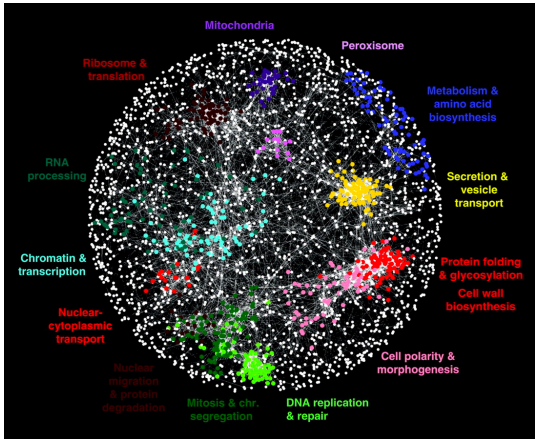


Image credit: University of Leichester; Costanzo et al., *Science* 327(5964):425-431, (2010)

Difficult computational problems

- Given a human genome, where can I find a particular substring?
  - For example, a gene from another species

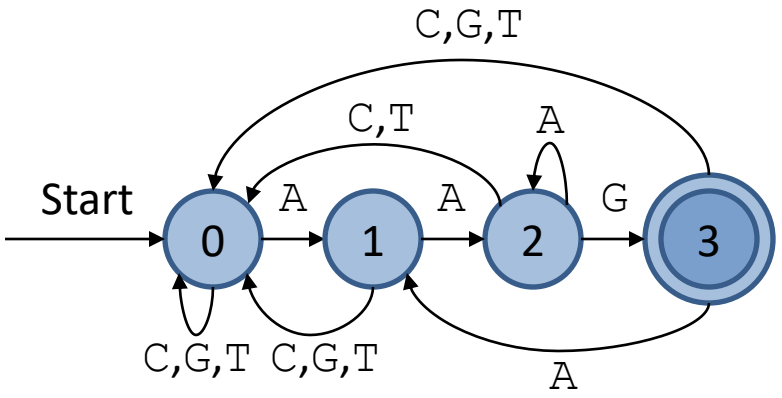
Where is TATACATTAG?

CTCACCGTCGTAAATCTATGATCTGGCTTGGCCTGCAGT
AGCTCTTTCATTTCGGGCTTATCTAATGCTGACTGGTCG
GTCCTGGCTACGCTCCAAAACGTACGTATTCGGGCCATC
GAGGCTAGCGGCACTTCGAGCGATCTATCGGGAGCTTTG
GCTATCGATCGGGCGATCGATGCTGACGTACGTAGCGCG
CGATCGAGCGCGGCTAGCTAGCGGCATCGTAGCTACGTA
GCTACGGCGCTATTTCGATCGAGTCGTGTCTAGTCGGAT
ATAGCTATGCATCTAGCTGAGGCATCTGAGCGGATCGAT
GCTAGGGCGATCGGAGCTAGCTGAGCTAGCTAGCTGAGC
GCTAGCGAGCGTACGAGCGATCGAGCGAGTCTA
GATTCTAGCGATATACATTAGCCCGATCGTATG
AGGGCTAGCATGCGGATCTATCGAGCGGCTATC
ATTCGATCGAGCGATCTAGCGAGCTATCGATCG

Image source: http://img1.etsystatic.com/000/0/6103070/il_fullxfull.203233493.jpg

How to find a string from a long string?

- Given:
  - A long string *s* of length *n* (e.g., `TTCAAGCCGTAAAG`)
  - A short string *r* of length *m* (e.g., `AAG`)
- Goal:
  - Find all occurrences of *r* in *s*
- Methods:
  - Linear search
  - Using a finite automaton of *r*
  - Using a suffix tree of *s*
  - …



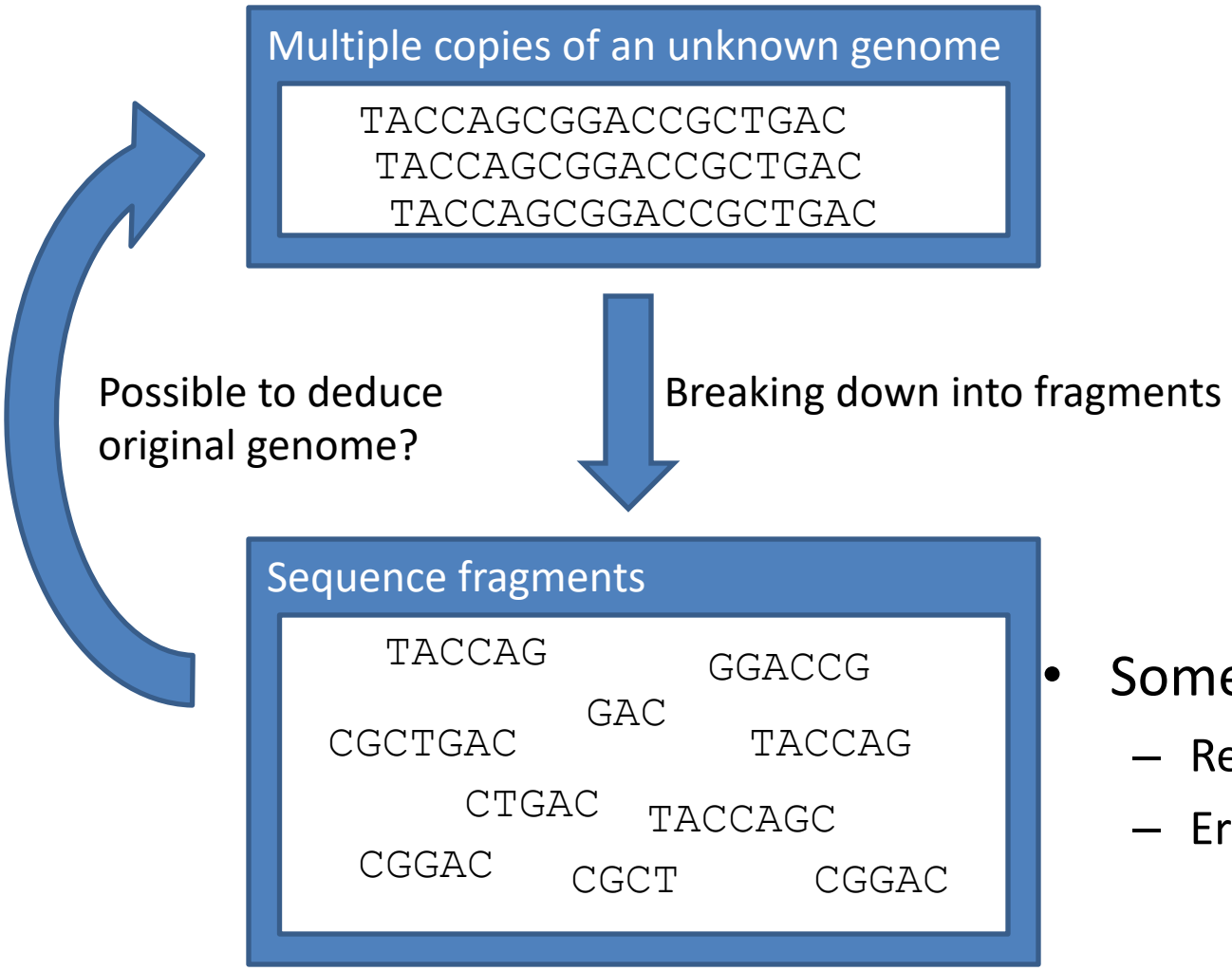
Real life example (1)

- Biomedical scenario: I found an interesting gene in mouse that is related to obesity. Do humans also have this gene?
- Computational definition:
  - I have a short string *r* (say, length *m*=10,000) – the DNA sequence of the mouse gene
  - I have a long string *s* (say, length *n*=3,000,000,000) – the whole set of DNA sequences (the "genome") of human
  - Can I find an occurrence of *r* in *s*?
- Some variations:
  - Inexact match
  - Many *r*'s
  - Many *s*'s

Real life example (2)

- Biomedical scenario: I have many short fragments of DNA from a genome. How do I get back the sequence of the original genome?
- Computational definition:
  - I have billions of short strings r_1, r_2, r_3, ..., each of length 100
  - They are substrings of a long string *s* of length 3,000,000,000
  - Can I reconstruct *s* from the *r*'s (theoretically and practically)?

Real life example (2)



- Some considerations:
  - Repeats in *s*
  - Errors in *r*'s

Real life example (3)

- Biomedical scenario: I have the gene expression profiles of 100 liver cancer patients and 100 healthy controls. How do I find out the genes that may cause the disease?
- Computational definition:
  - I have 100 vectors v_1, v_2, ..., v_{100} (patients) and 100 vectors u_1, u_2, ..., u_{100} (controls)
  - Each vector has 20,000 real-numbered values (one for each gene)
  - That is, we have 200 points in a 20,000-dimentional space
  - Can I find a hyperplane such that all *u*'s are on one side of the plane and all *v*'s are on the other side?
- Some considerations:
  - What if such a hyperplane does not exist?
  - What if there are multiple solutions?
  - How to explain the meaning of the hyperplane to biologists/medical doctors?
  - How to know if the hyperplane is biomedically meaningful?

What is bioinformatics?

- Answer #3: 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

What is bioinformatics?

- Answer #4: Contributions and prospects
  - Very meaningful field, with direct contributions to
    - Medicine
    - Biology
    - Computer science
    - …
  - Cutting-edge, challenging problems
  - A bottleneck in biomedical research
  - Short of qualified people
- A new and growing field with a lot of potentials

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 (fast) in Hong Kong

What is your answer?

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

- An elective subject of your curriculum?
- 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
- Anonymous
- 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?
Algorithms					
Biology					
Bioinformatics					

  2. Did you study biology before? At which level?
  3. Did you take any algorithm courses before? Which one(s)?
  4. Which programming languages can you program in? At which level?

The survey (cont'd)

- Questions:

5. How much do you know about these topics?

	I can teach this topic	I know it	Sort of heard about it	Huh?
Hash table				
Eulerian path				
Dynamic programming				
Big O notation				
Bayes' Theorem				
Maximum likelihood				
Transcription and translation				
Phylogenetic trees				
Massively parallel sequencing				

6. Do you have any special requests for this course?

Part 3

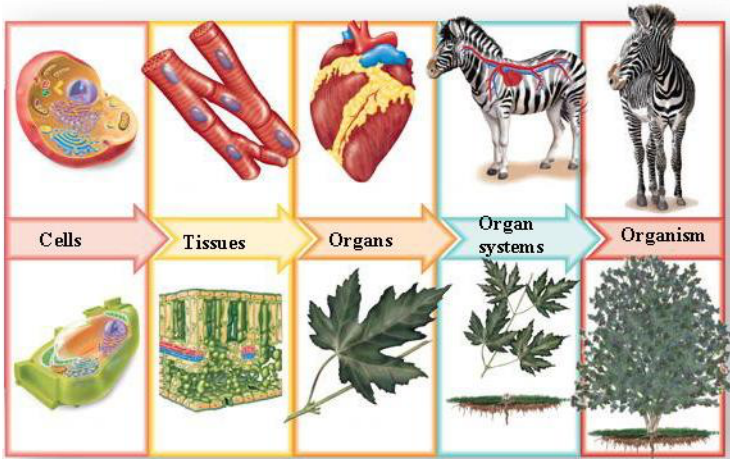
INTRODUCTION TO GENETICS AND MOLECULAR BIOLOGY

Basic biological knowledge

- Useful for
  - Your general knowledge
  - Defining terminology
  - Helping you appreciate the importance of what you are going to learn
- Don’t panic. 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. Revise the materials later when we talk about relevant topics.

Introduction to molecular biology

- Cell: Basic functional unit of life



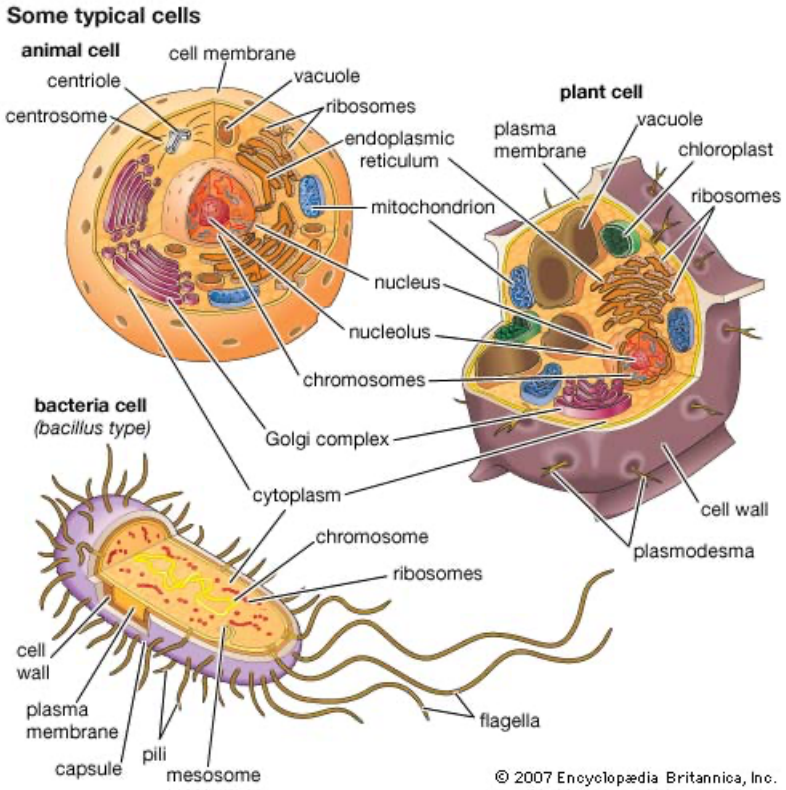
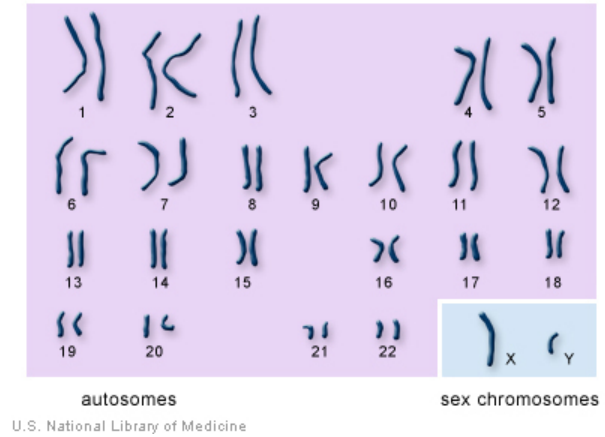


Image credit: http://legacy.hopkinsville.kctcs.edu/sitecore/instructors/Jason-Arnold/VLI/Module%201/m1science/f1-01_levels_of_biologi_c.jpg, http://dbscience5.wikispaces.com/file/view/78585-004-A63E1F47.jpg/51586701/78585-004-A63E1F47.jpg

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)



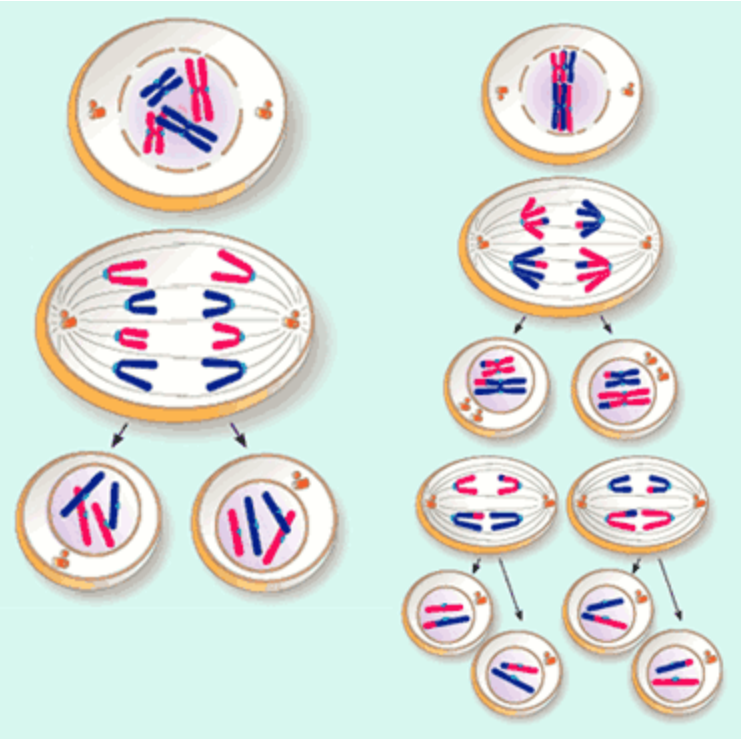
- For higher organisms, chromosomes are in the cell nucleus
- When a cell divides by mitosis, each chromosome is duplicated and both daughter cells have the complete set of chromosomes

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

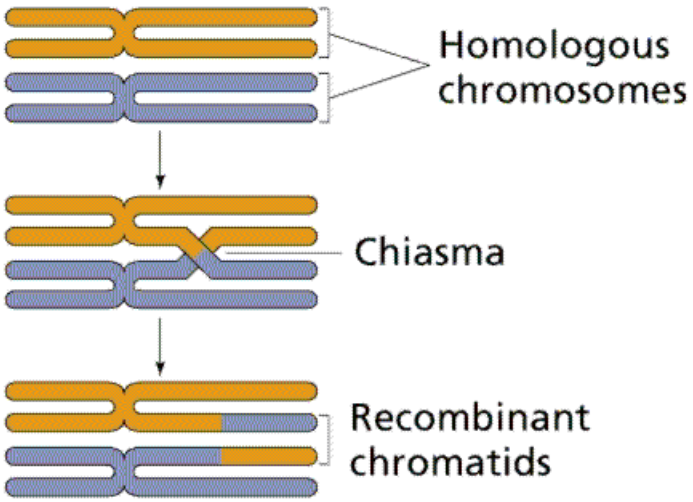
Image credit: http://3.bp.blogspot.com/_207DNIaL-gc/TQk9QRaI5mI/AAAAAAAAAXg/z0Xh8CTgHto/s400/mitosismeiosissummary.gif

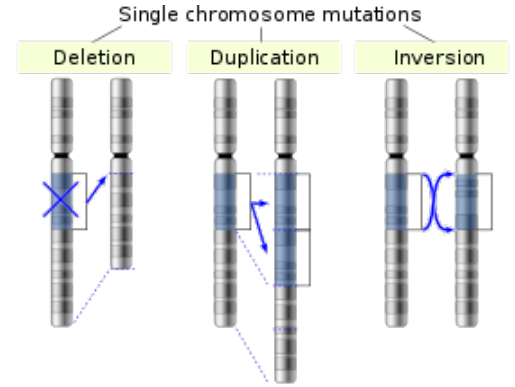
Diploid genome

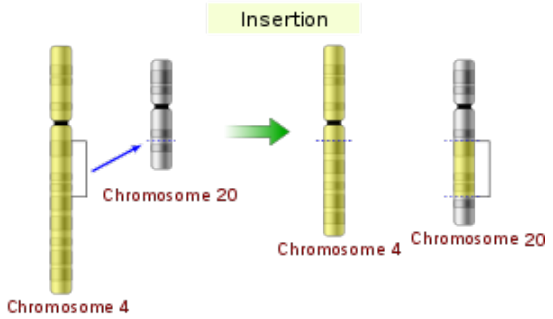
- Why do we need two copies of each chromosome?
  - More combinations: For each of the 23 pairs of chromosomes, only one is passed to each offspring, which creates 2^{23} possible combinations.
  - Error tolerance: If one copy has problems, there is still another copy.
  - Evolution: Having one normal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

How to change?

- Recombination
- Insertion
- Deletion
- …







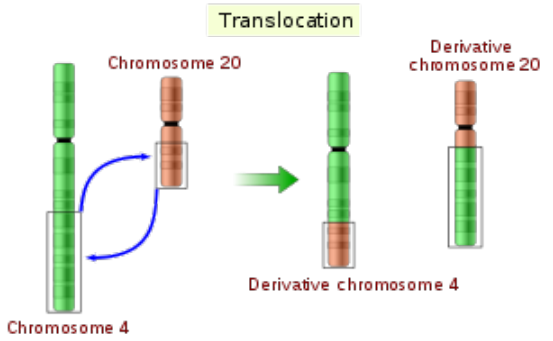


Image credit: http://www2.estrellamountain.edu/faculty/farabee/biobk/Crossover.gif, Wikipedia

Why do changes matter?

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

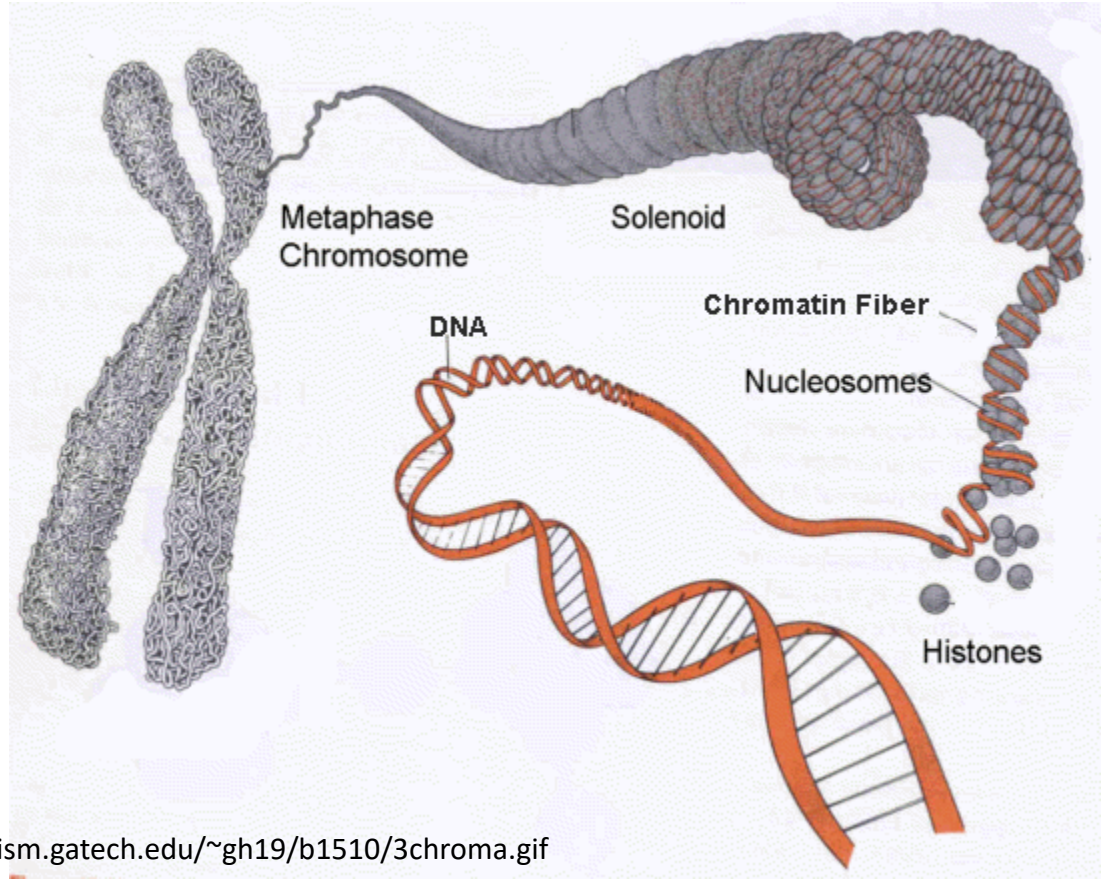
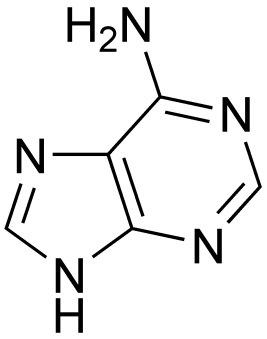


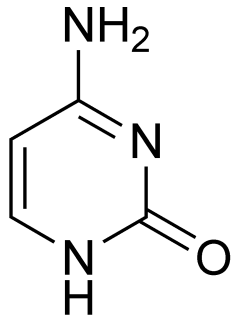
Image credit: http://www.prism.gatech.edu/~gh19/b1510/3chroma.gif

DNA

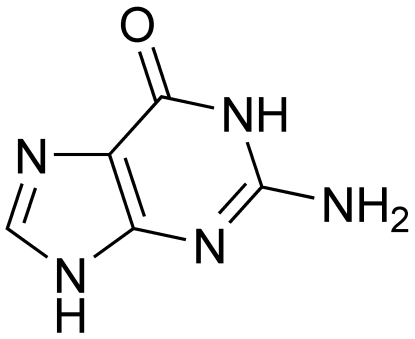
- DNA: DeoxyriboNucleic Acid
  - Two long chains of basic units called nucleotides (bases)
  - Four types of nucleotides:



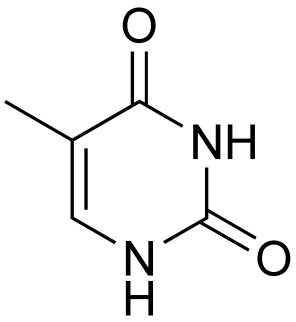
Adenine (A)



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
- Three components of each unit:
  - 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}

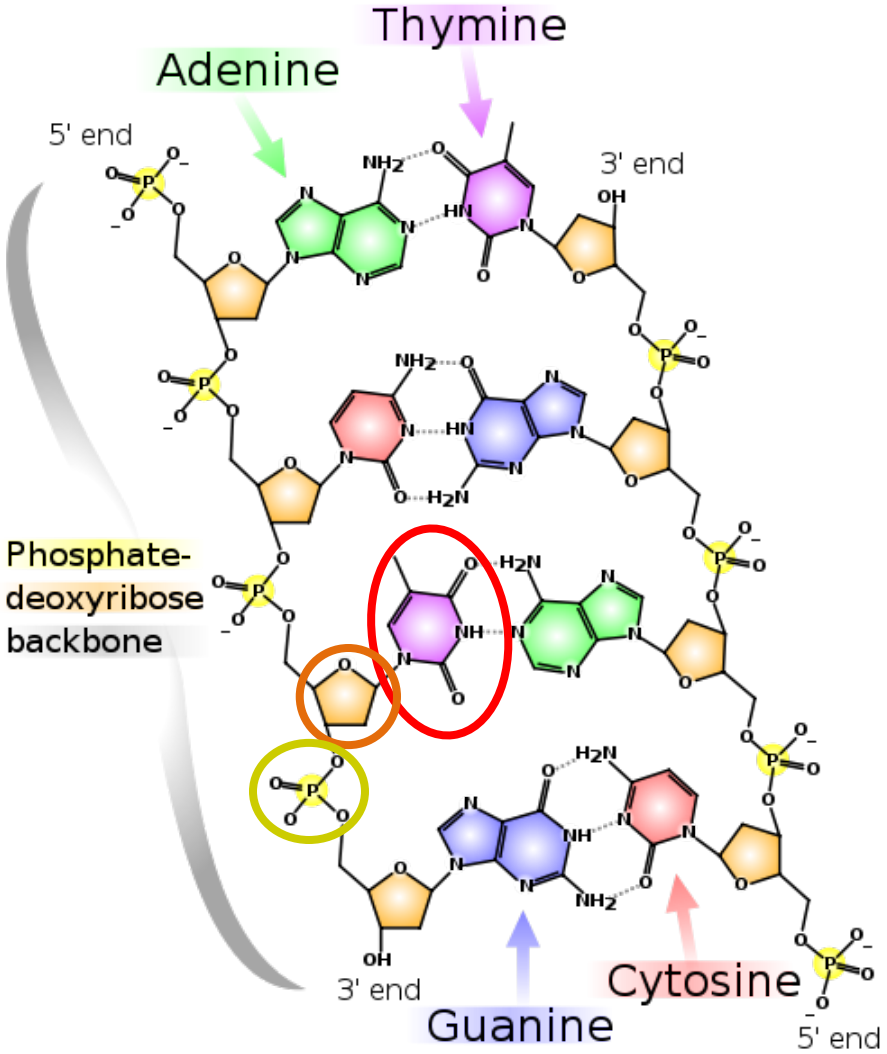
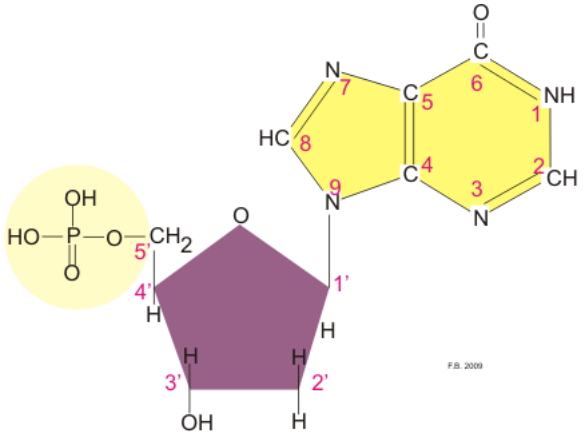


Image credit: Wikipedia

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`

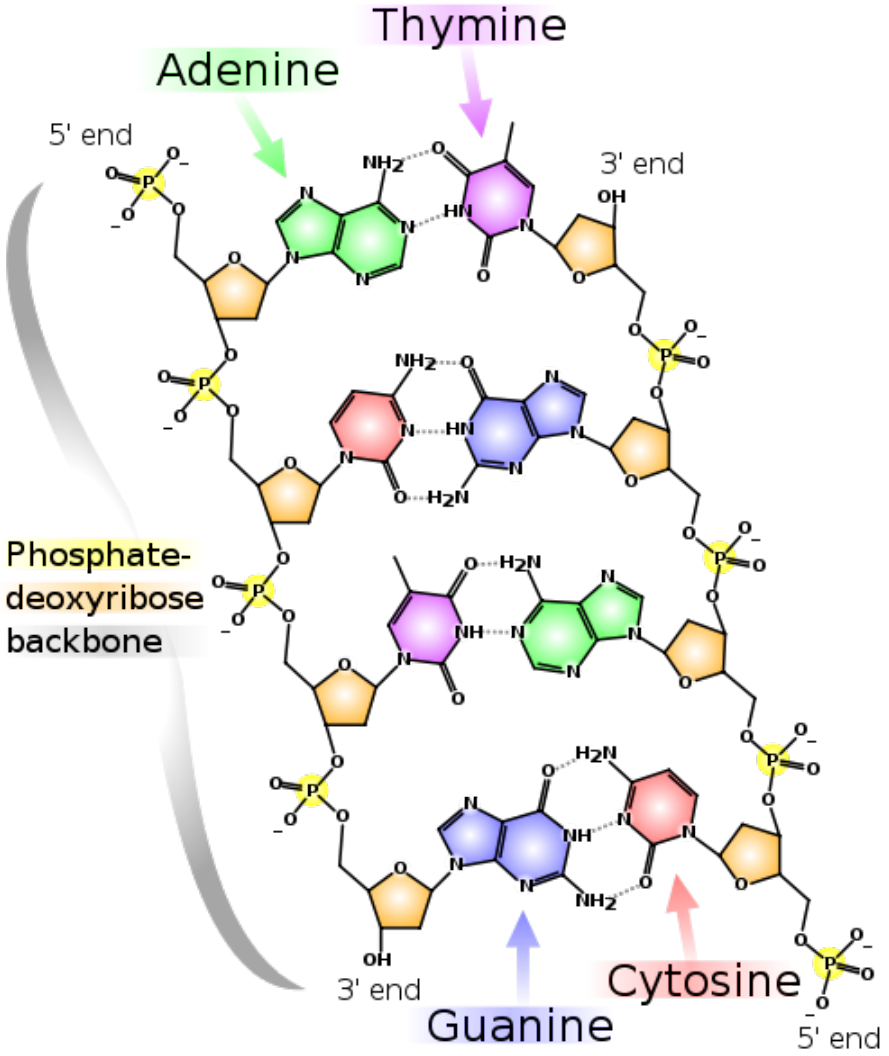


Image credit: Wikipedia, Wikibooks

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"

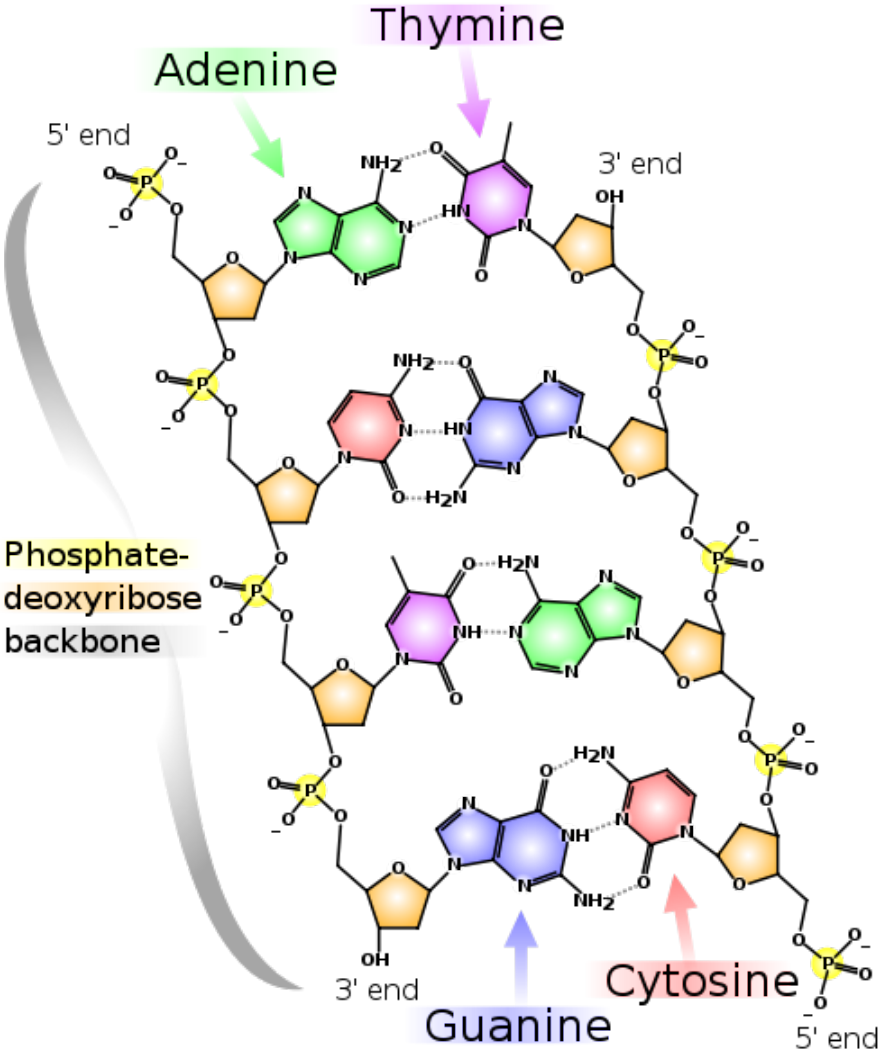


Image credit: Wikipedia

DNA

- The two strands form a double helix structure

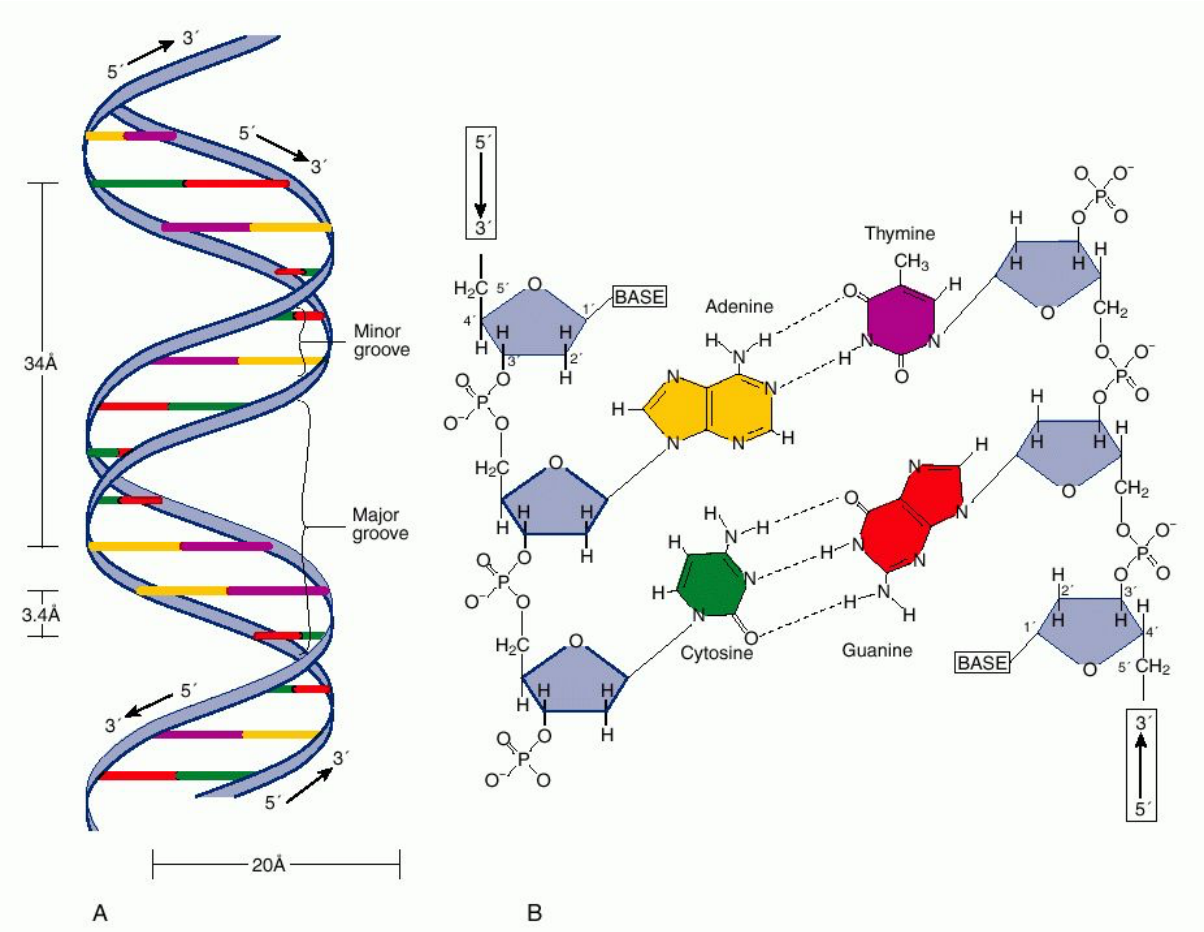


Image credit: http://medical-dictionary.thefreedictionary.com/_/viewer.aspx?path=dorland&name=deoxyribonucleic-acid.jpg

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

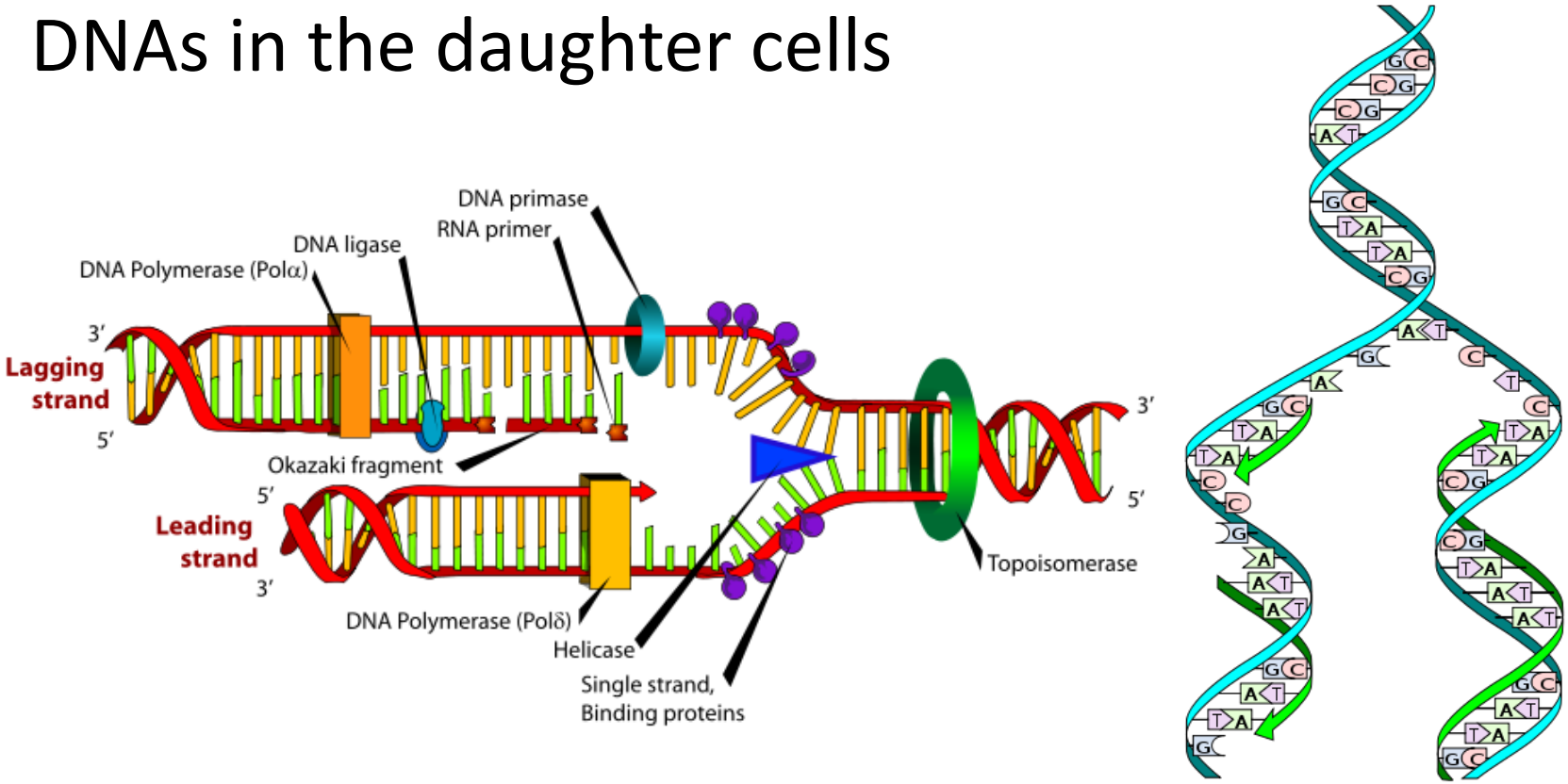


Image credit: Wikipedia

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

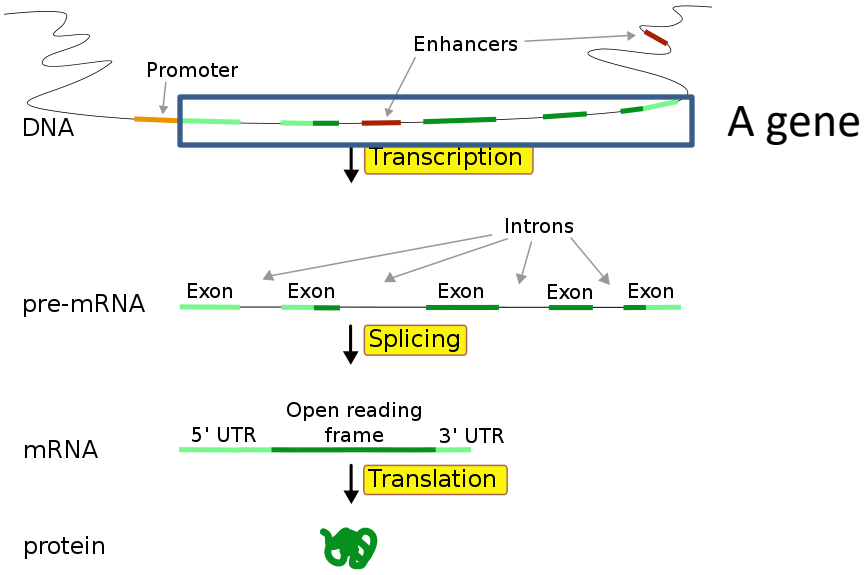


Image credit: Wikipedia

Genes

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

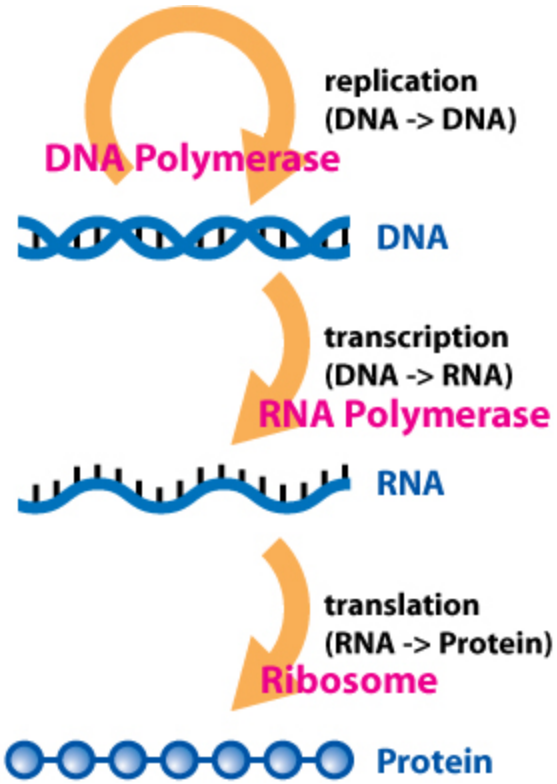


Image credit: Wikipedia

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

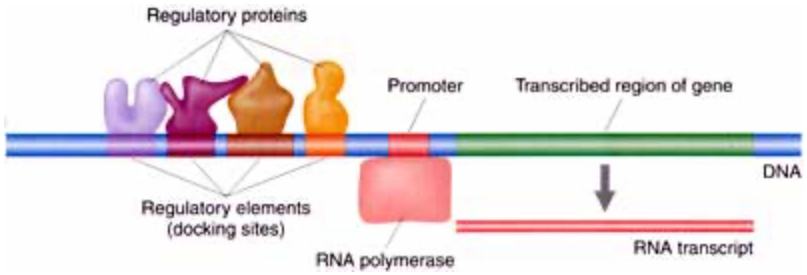
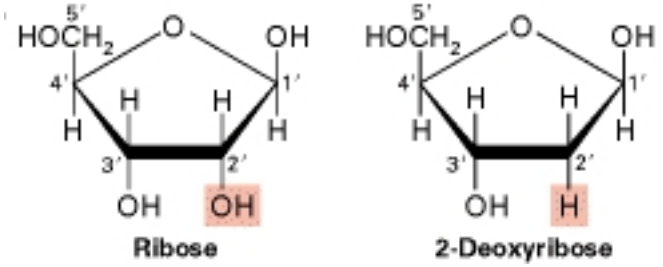


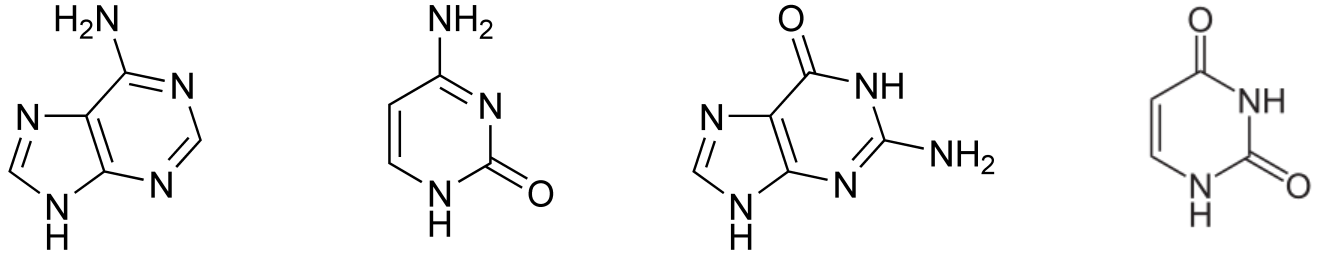
Image credit: http://scienceblogs.com/pharyngula/upload/2007/01/simple_gene_reg.jpg

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)



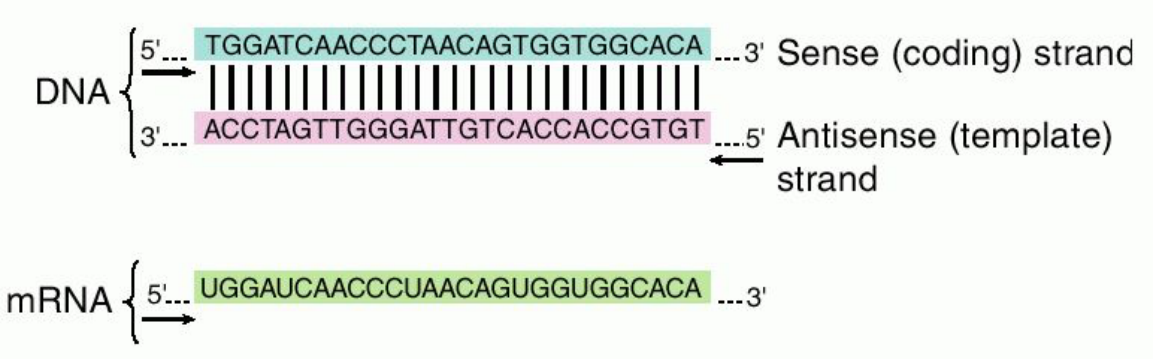
Adenine (A) Cytosine (C) Guanine (G) Uracil (U)

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. Will explain later.
- 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

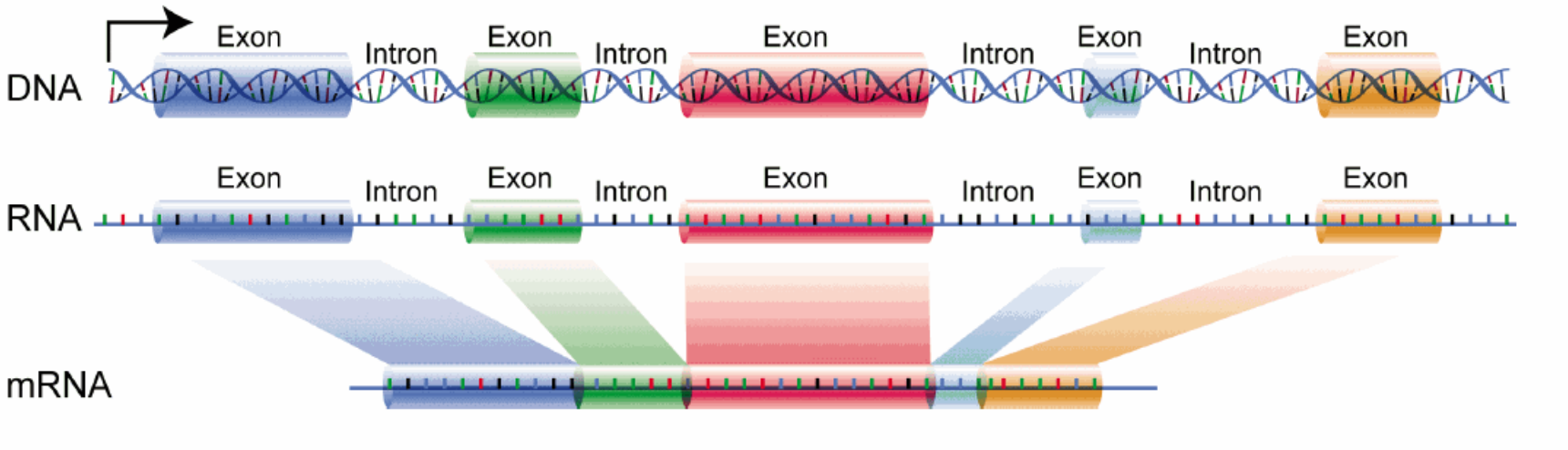


Image credit: Wikipedia

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

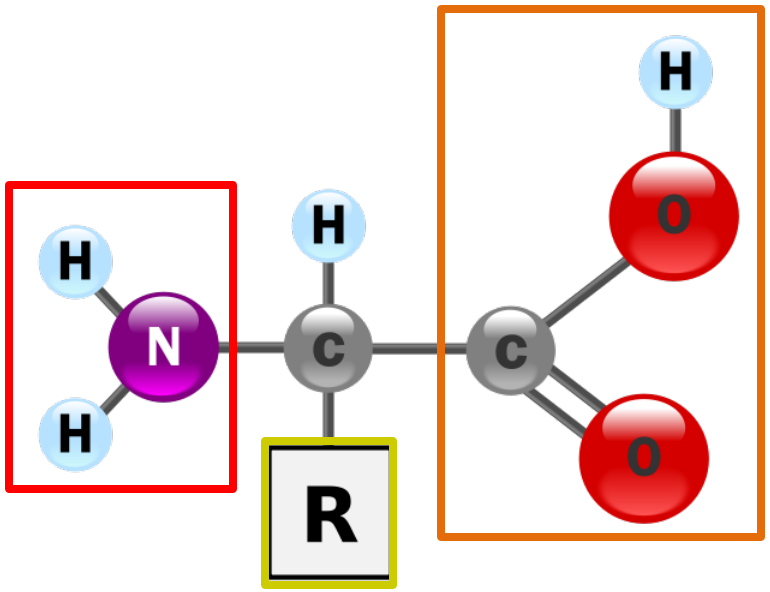
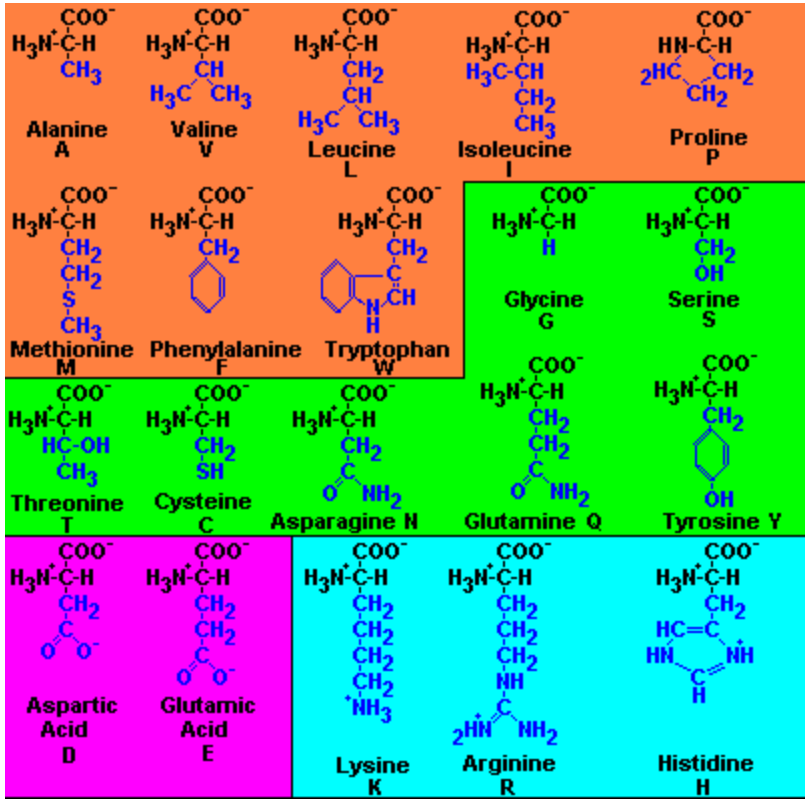


Image credit: Wikipedia

Amino acids

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



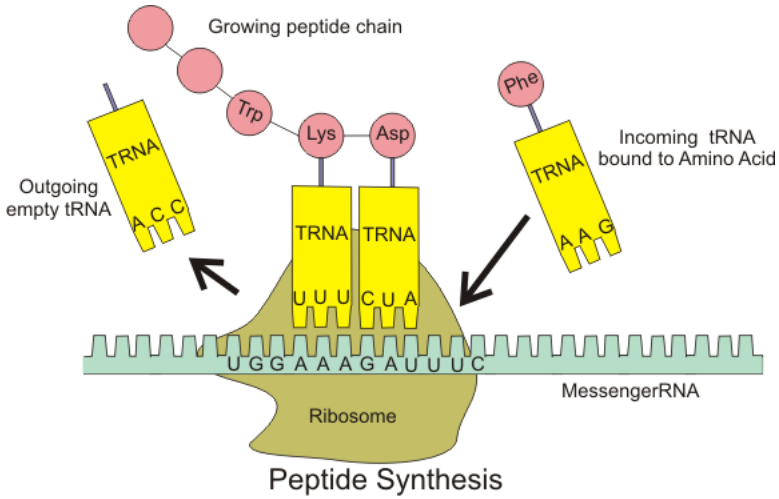
A protein can be represented by a string with the alphabet {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}

- 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)



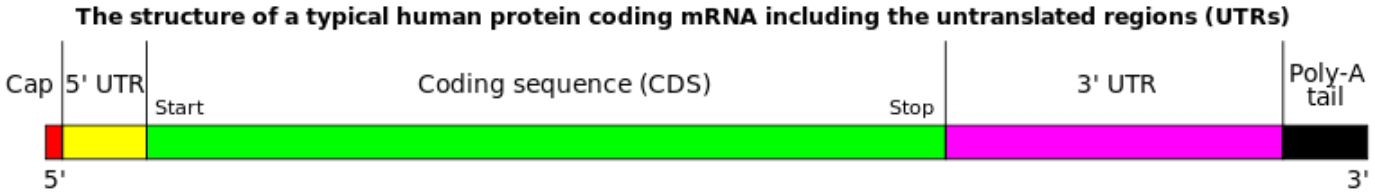


Image credit: http://www.eurekadiscoveries.com/wp-content/uploads/2010/06/Peptide_syn.png, Wikipedia

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

		2nd base							
		U		C		A		G	
1st base	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”

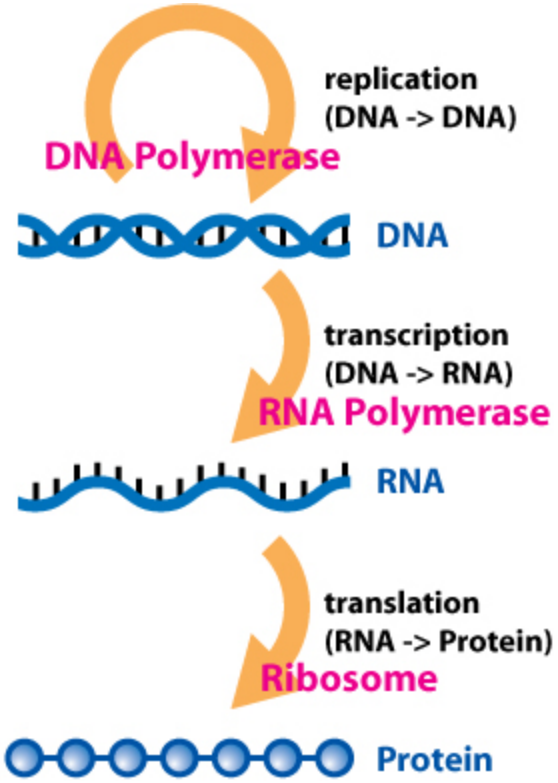


Image credit: Wikipedia

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

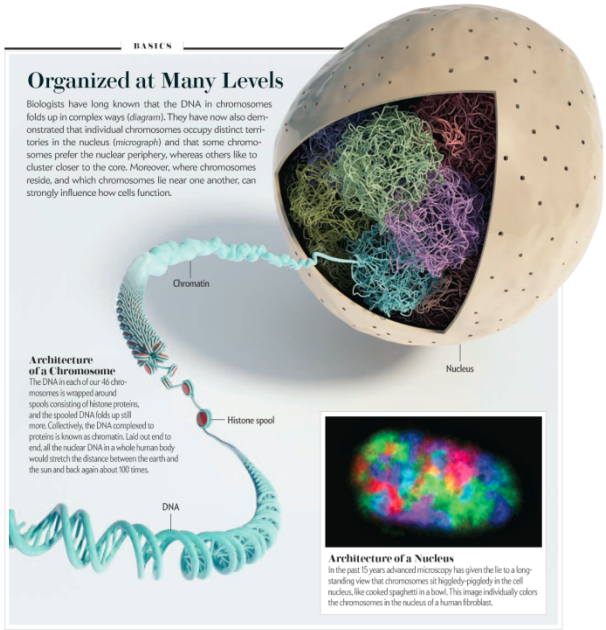
Quick quiz

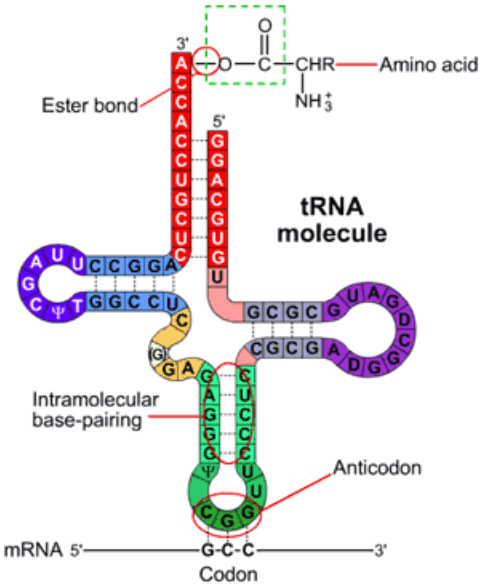
2. What information do we need to fully identify a genomic location?
   - Chromosome, position, strand
3. What information do we need to fully identify a genomic interval (e.g., a gene)?
   - Chromosome, start position, end position, strand

- 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.





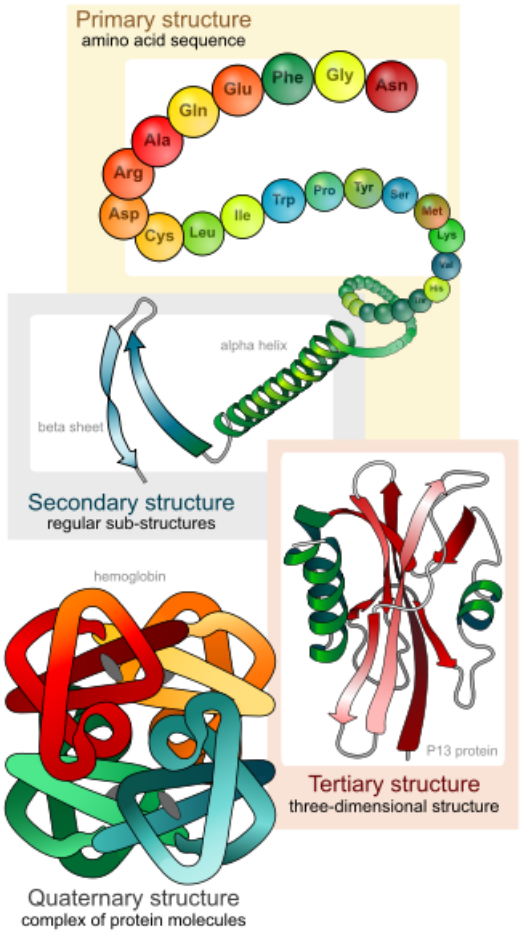


Image credit: Scientific American, http://www.wiley.com/college/boyer/0470003790/structure/tRNA/trna_diagram.gif, Wikipedia

Part 4

OVERVIEW OF CLASS TOPICS

Tentative class schedule

- Topics and tasks

Week	Topic	Tasks
1	Introduction	
2-4	Optimal and heuristic sequence alignment	Assignment #1
5-6	Short read alignment and sequence assembly	Assignment #2
7-9	Sequence motifs and probabilistic models	Assignment #3 Mid-term exam.
10-11	Phylogenetic trees and inheritance	Assignment #4
12	Clustering algorithms	
13	RNA secondary structures prediction	Assignment #5

Tentative class schedule

- Biological objects, data representation and algorithms

Week	Biological objects	Data representation	Some technical topics
2-4	DNA, RNA, Proteins	Strings Tables	Dynamic programming, BLAST, FASTA, ClustalW
5-6	DNA, RNA, Proteins	Strings Graphs	Suffix trie/tree/array, Burrows-Wheeler transform, De Bruijn graph
7-9	DNA, RNA, Proteins	Strings Probability models	Hidden Markov models, Naive Bayes, Logistic regression
10-11	DNA, Proteins	Trees Tables	Maximum parsimony, maximum likelihood, UPGMA, Neighbor-joining
12	All, and their activities	Matrices	k-means, hierarchical clustering, heap, quad tree
13	RNA	Strings	Dynamic programming, stochastic context-free grammar

Epilogue

CASE STUDY, SUMMARY AND FURTHER READINGS

Case study: Mutation and disease

- Classic example: Sickle-cell Anemia
  - Normal red blood cells are round and flexible, able to carry oxygen and travel through blood vessels
  - People with Sickle-cell Anemia have red blood cells that look like a sickle and get stuck in blood vessels

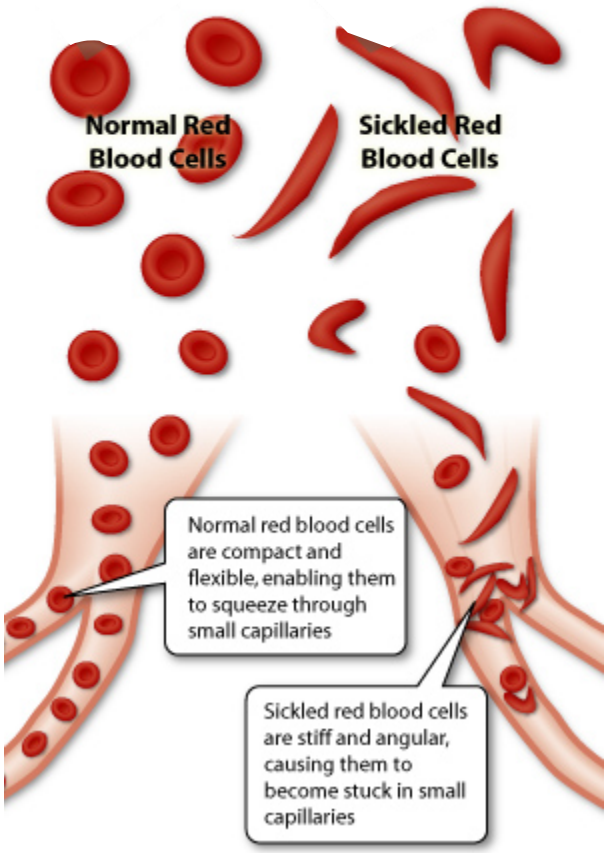


Image credit: http://learn.genetics.utah.edu/content/disorders/whataregd/sicklecell/images/sicklecell.jpg; Wikipedia

Case study: Mutation and disease

- What went wrong?
  - Hemoglobin molecules, which are responsible for carrying oxygen, form a long chain in affected people
  - Why? There is a mutation in one of the hemoglobin genes on chromosome 11

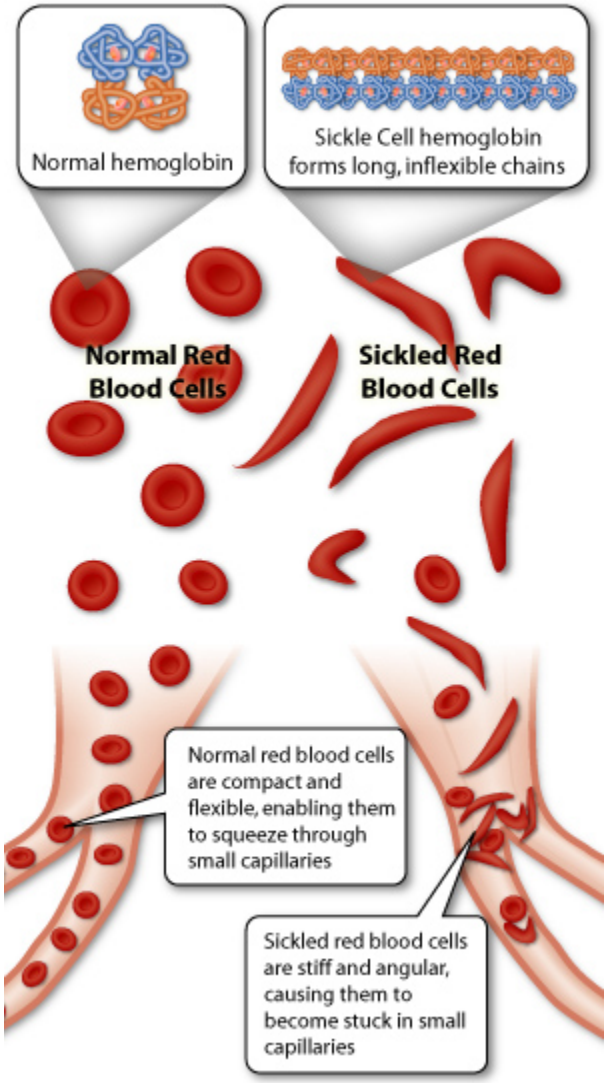


Image credit: http://learn.genetics.utah.edu/content/disorders/whataregd/sicklecell/images/sicklecell.jpg

Case study: Mutation and disease

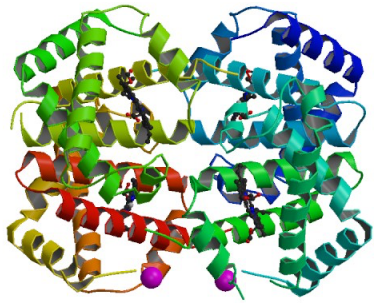
- Mutation in DNA causes change in protein structure

HBB Sequence in Normal Adult Hemoglobin (Hb A):

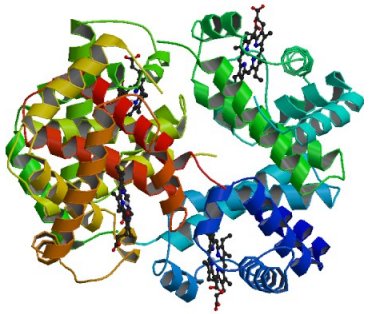
Nucleotide	CTG	ACT	CCT	GAG	GAG	AAG	TCT
Amino Acid	Leu	Thr	Pro	Glu	Glu	Lys	Ser
	3			6			9

HBB Sequence in Mutant Adult Hemoglobin (Hb S):

Nucleotide	CTG	ACT	CCT	GTG	GAG	AAG	TCT
Amino Acid	Leu	Thr	Pro	Val	Glu	Lys	Ser
	3			6			9



Normal human hemoglobin



A common mutated form of human hemoglobin that causes Sickle-cell Anemia

Check out more by yourselves:
https://www.ncbi.nlm.nih.gov/snp/rs334

Image credit: Protein Data Bank, entries 4HHB and 2HBS,
http://www.ornl.gov/sci/techresources/Human_Genome/posters/chromosome/Gifs/HBBmutseq_2.gif

Summary

- Bioinformatics
  - Using computational methods to assist biomedical research
  - Large data size
  - Difficult computational problems
- Basic genetics and molecular biology
  - Mitosis and meiosis
  - DNA $\xrightarrow{\text{(transcription)}}$ RNA $\xrightarrow{\text{(translation)}}$ protein

Further readings

- Chapter 1 of *Algorithms in Bioinformatics: A Practical Introduction*
  - More comprehensive introduction of the basic concepts
  - Free slides available