CS681: Advanced Topics in Computational Biology

Can Alkan
EA224
calkan@cs.bilkent.edu.tr

http://www.cs.bilkent.edu.tr/~calkan/teaching/cs681/

CS681

- Class hours:
  - Wed 9:40 - 10:30; Fri 10:40 - 12:30
- Class room: EE317
- Office hour: Tue + Thu 13:00-14:00
- Grading:
  - 1 project: 50%
  - Class participation: 10%
  - Paper presentation & summary report: 40%

CS681

- Textbook: None
- Recommended Material
  - An Introduction to Bioinformatics Algorithms (Computational Molecular Biology), Neil Jones and Pavel Pevzner, MIT Press, 2004
  - Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Richard Durbin, Sean R. Eddy, Anders Krogh, Graeme Mitchison, Cambridge University Press
  - Bioinformatics: The Machine Learning Approach, Second Edition, Pierre Baldi, Soren Brunak, MIT Press
  - Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology, Dan Gusfield, Cambridge University Press
  - Scientific journals

CS681

- This course is about **algorithms** in the field of bioinformatics / computational biology; mostly genomics:
  - What are the problems?
  - What algorithms are developed for what problem?
  - What is missing / needs advances in the field.
  - Possible research directions for graduate students.

CS681: Assumptions

- You are assumed to know/understand
  - Advanced algorithms
    - Dynamic programming, greedy algorithms, graph theory
    - CS473 is required
    - CS573 is better
    - CS570 is recommended
  - Programming: C, C++, Java
- You don’t *have to* be a “biology expert” but MBG 101 or 110 would be beneficial

INTRODUCTION, CONCEPTS AND TERMS

Bioinformatics & Computational Biology

- Bioinformatics: Development of methods based on computer science for problems in biology &medicine
  - Sequence analysis (combinatorial and statistical/probabilistic methods)
  - Graph theory
  - Data mining
  - Database
  - Statistics
  - Image processing
  - Visualization
  - …..

CS 481 and CS 681

- Computational biology: Application of computational methods to address questions in biology & medicine

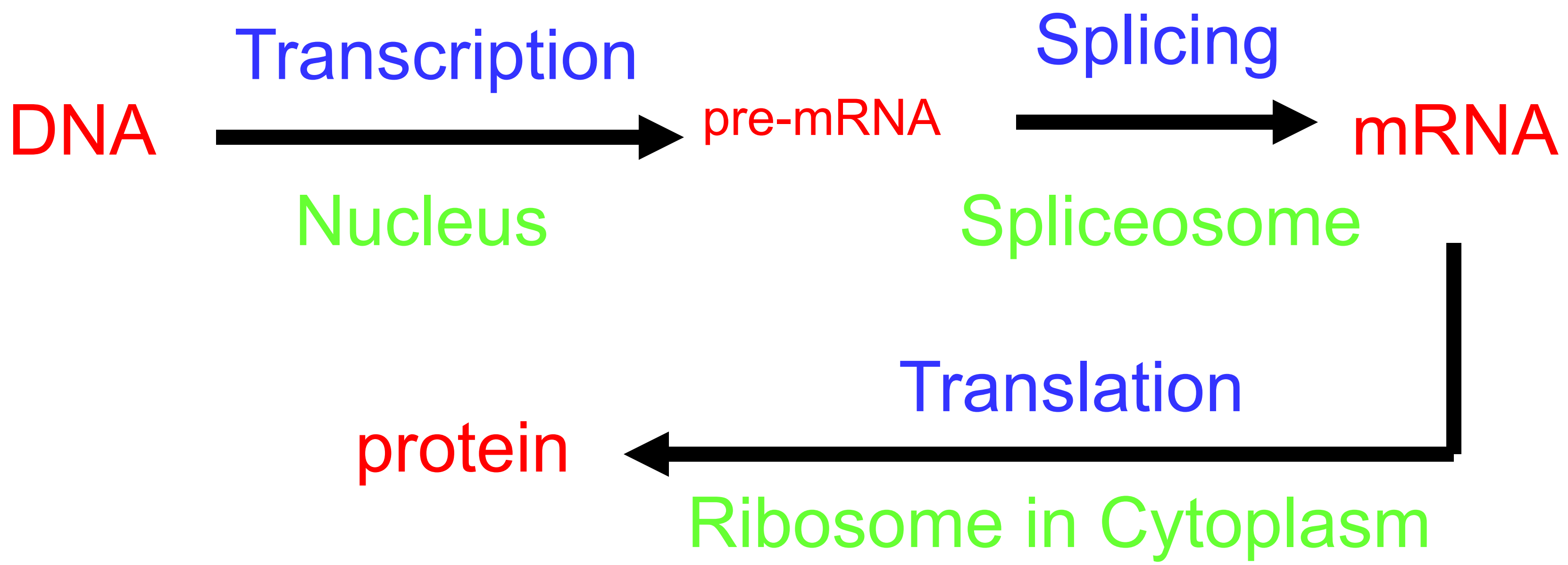
Concepts

- Gene: discrete units of hereditary information located on the chromosomes and consisting of DNA.
- Genetics: study of inherited phenotypes
- Genotype: The genetic makeup of an organism
- Phenotype: the physical expressed traits of an organism
- Genome: entire hereditary information of an organism
- Genomics: analysis of the whole genome (that is, the DNA content for most organisims; RNA content for retroviruses)
- Transcriptome: set of all RNA molecules
- Proteome: set of all protein molecules

All life depends on 3 critical molecules

- DNAs
  - Hold information on how cell works
- RNAs
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein
- Proteins
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body’s major components (e.g. hair, skin, etc.)
- For a computer scientist, these are all strings derived from three alphabets.

Central dogma of biology



- Base Pairing Rule: A and T or U is held together by 2 hydrogen bonds and G and C is held together by 3 hydrogen bonds.
- Note: Some RNA stays as RNA (ie tRNA,rRNA, miRNA, snoRNA, etc.).

Alphabets

DNA:

$\Sigma = \{A, C, G, T\}$

A pairs with T; G pairs with C

RNA:

$\Sigma = \{A, C, G, U\}$

A pairs with U; G pairs with C

Protein:

$\Sigma = \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}$ *and*

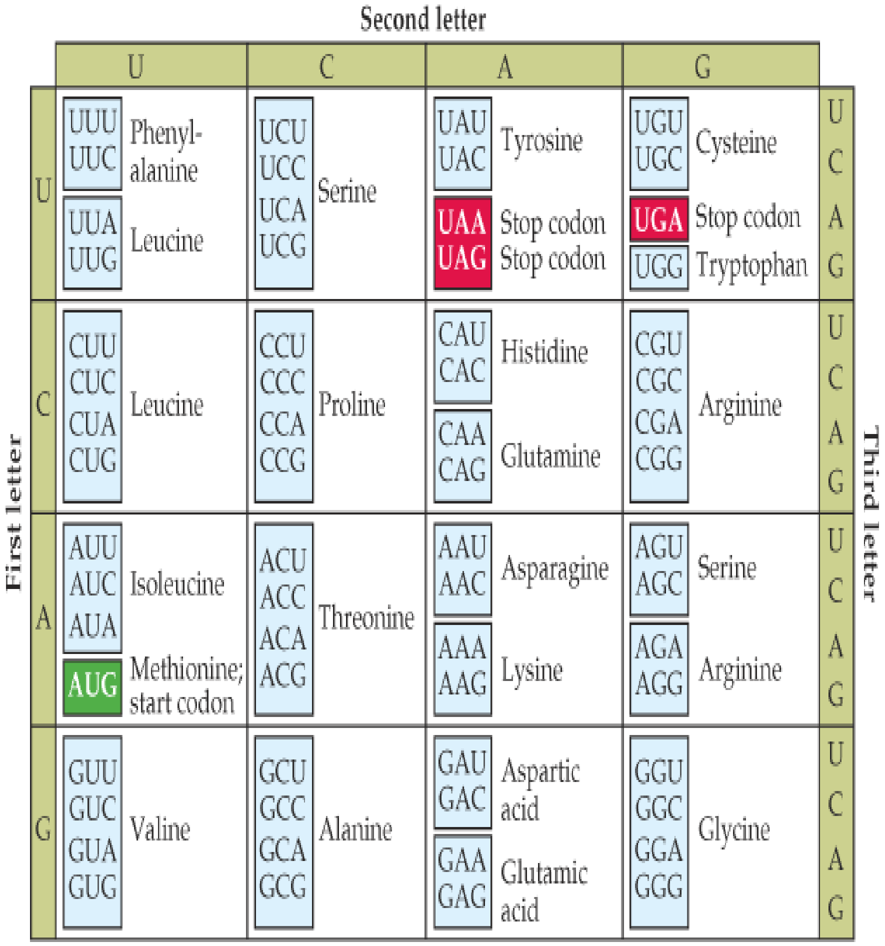
B = N | D

Z = Q | E

X = *any*

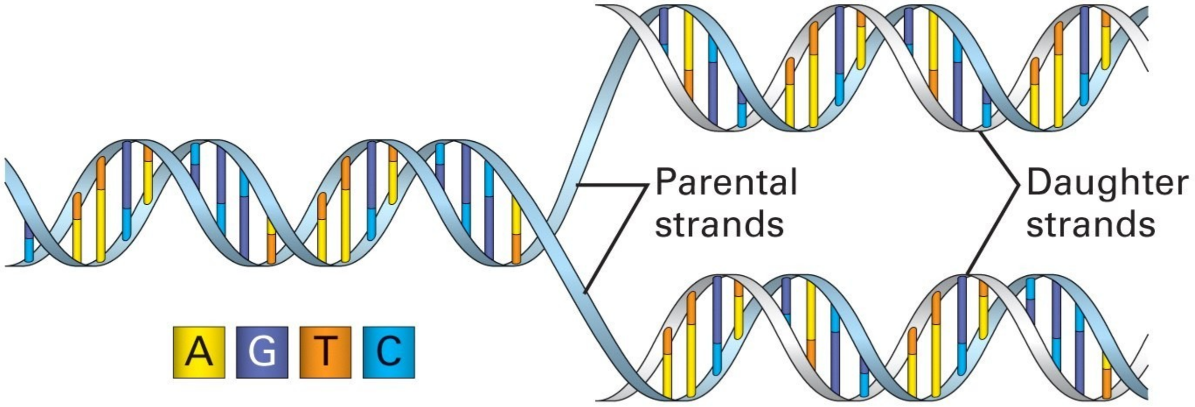
Cell Information: Instruction book of Life

- DNA, RNA, and Proteins are examples of strings written in either the four-letter nucleotide of DNA and RNA (A C G T/U)
- or the twenty-letter amino acid of proteins. Each amino acid is coded by 3 nucleotides called **codon**. (Leu, Arg, Met, etc.)



	Second letter				
First letter	U	C	A	G	Third letter
U	UUU Phenylalanine	UCU Serine	UAU Tyrosine	UGU Cysteine	U
	UUC	UCC	UAC	UGC	C
	UUA Leucine	UCA	UAA Stop codon	UGA Stop codon	A
	UUG	UCG	UAG Stop codon	UGG Tryptophan	G
C	CUU Leucine	CCU Proline	CAU Histidine	CGU Arginine	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA Glutamine	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU Isoleucine	ACU Threonine	AAU Asparagine	AGU Serine	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA	AAA Lysine	AGA Arginine	A
	AUG Methionine; start codon	ACG	AAG	AGG	G
G	GUU Valine	GCU Alanine	GAU Aspartic acid	GGU Glycine	U
	GUC	GCC	GAC	GGC	C
	GUA	GCA	GAA Glutamic acid	GGA	A
	GUG	GCG	GAG	GGG	G

DNA: The Code of Life

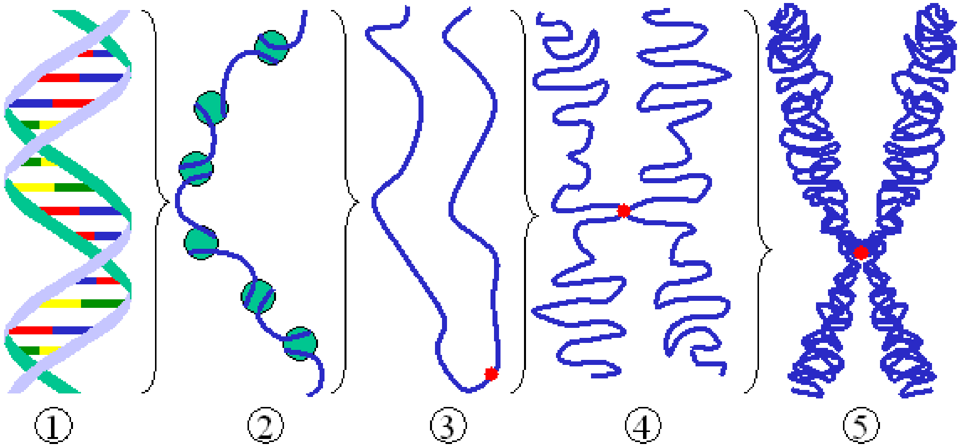


- The structure and the four genomic letters code for all living organisms
- Adenine, Guanine, Thymine, and Cytosine which pair A-T and C-G on complimentary strands.

DNA is organized into Chromosomes

- Chromosomes:
  - Found in the nucleus of the cell which is made from a long strand of DNA, “packaged” by proteins called *histones*. Different organisms have a different number of chromosomes in their cells.
  - Human genome has 23 pairs of chromosomes
    - 22 pairs of *autosomal* chromosomes (chr1 to chr22)
    - 1 pair of *sex* chromosomes (chrX+chrX or chrX+chrY)
- Ploidy: number of sets of chromosomes
  - Haploid (n): one of each chromosome
    - Sperm & egg cells; hydatidiform mole
  - Diploid (2n): two of each chromosome
    - All other cells in mammals (human, chimp, cat, dog, etc.)
  - Triploid (3n), Tetraploid (4n), etc.
    - Tetraploidy is common in plants

Genetic Information: Chromosomes



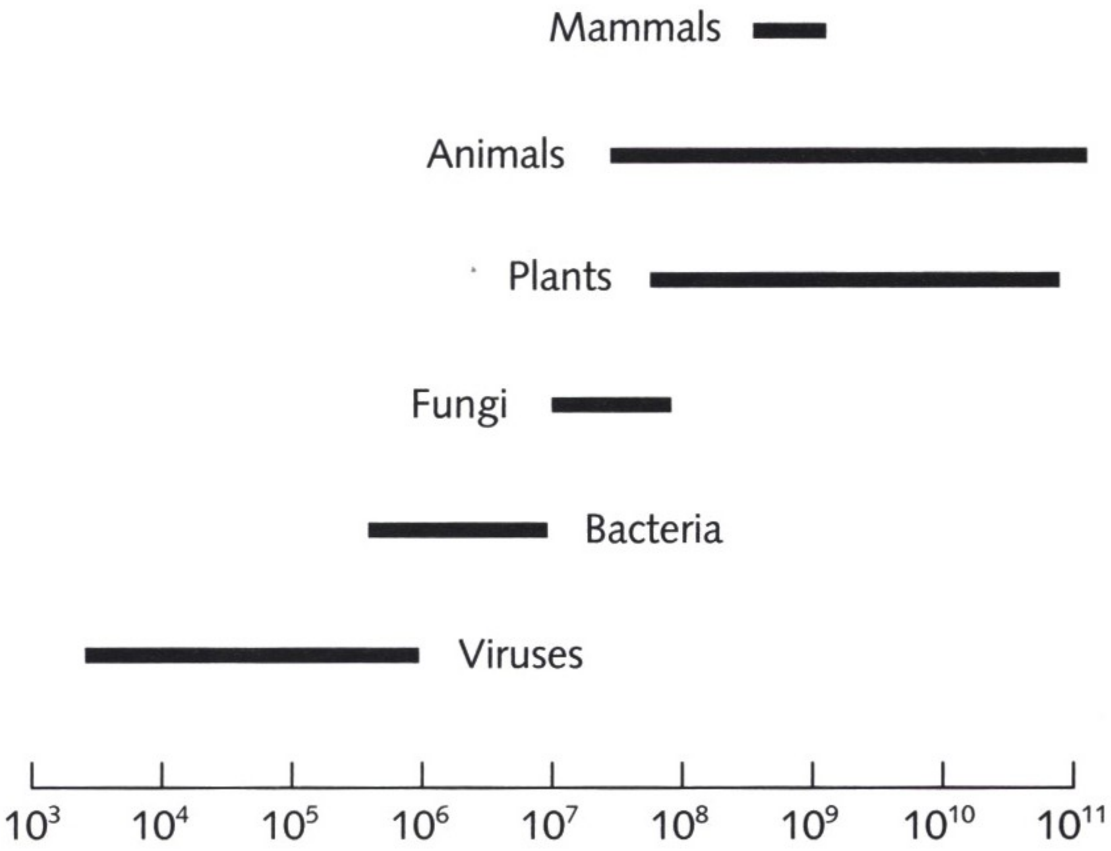
- (1) Double helix DNA strand.
- (2) Chromatin strand (**DNA** with **histones**)
- (3) Condensed chromatin during interphase with **centromere**.
- (4) Condensed chromatin during prophase
- (5) Chromosome during metaphase

Euchromatin: Lightly packed DNA; gene rich, often active
Heterochromatin: Tightly packed DNA; usually repetitive; structural functions

Genomes

- Definition (again): the entire collection of hereditary material
  - Most organisms: DNA content
  - Retroviruses (like HIV, influenza): RNA content
- Eukaryotes can have 2-3 genomes:
  - Nuclear (default)
  - Mitochondrial
  - Plastid
- Libraries & instruction sets for the cells
- Identical in most cells, except the immune system cells
- Germline DNA: material that may be transmitted to the child (germ cell)
  - Germ cell: cells that give rise to gametes (sperm/egg)
- Somatic DNA: material in cells other than germ cells & gametes
  - Changes in somatic cells do not transmit to offspring

How big are genomes?



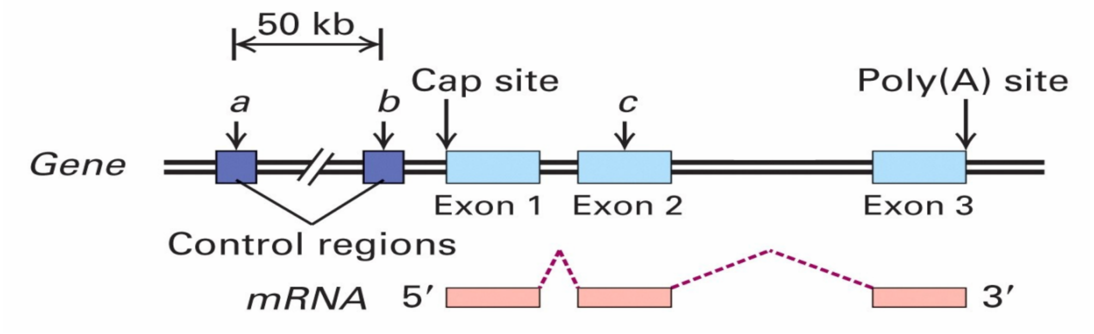
How big are genomes?

Organism	Genome Size (Bases)	Estimated Genes
Human (*Homo sapiens*)	3 billion	20,000
Laboratory mouse (*M. musculus*)	2.6 billion	20,000
Mustard weed (*A. thaliana*)	100 million	18,000
Roundworm (*C. elegans*)	97 million	16,000
Fruit fly (*D. melanogaster*)	137 million	12,000
Yeast (*S. cerevisiae*)	12.1 million	5,000
Bacterium (*E. coli*)	4.6 million	3,200
Human immunodeficiency virus (HIV)	9700	9

Genome "table of contents"

- Genes (~35%; but only 1% are coding exons)
  - Protein coding
  - Non-coding (ncRNA only)
- Pseudogenes: genes that lost their expression ability:
  - Evolutionary loss
  - Processed pseudogenes
- Repeats (~50%)
  - Transposable elements: sequence that can copy/paste themselves. Typically of virus origin.
  - Satellites (short tandem repeats [STR]; variable number of tandem repeats [VNTR])
  - Segmental duplications (5%)
    - Include genes and other repeat elements within

Genes



- Subsequences of DNA that are transcribed into RNA
  - Some encode for proteins, some do not
- Regulatory regions: up to 50 kb upstream of +1 site
- Exons: protein coding and untranslated regions (UTR)
  1 to 178 exons per gene (mean 8.8)
  8 bp to 17 kb per exon (mean 145 bp)
- Introns: sequence between exons; spliced out before translation
  average 1 kb – 50 kb per intron
- Gene size: Largest – 2.4 Mb (Dystrophin). Mean – 27 kb.

Genes can be switched on/off

- In an adult multicellular organism, there is a wide variety of cell types seen in the adult. eg, muscle, nerve and blood cells.
- The different cell types contain the *same* DNA.
- This differentiation arises because different cell types express different genes.
- Type of gene regulation mechanisms:
  - Promoters, enhancers, methylation, RNAi, etc.

Pseudogenes

- “Dead” genes that lost their coding ability
- Evolutionary process:
  - Mutations cause:
    - Early stop codons
    - Loss of promoter / enhancer sequence
- Processed pseudogenes:
  - A real gene is transcribed to mRNA, introns are spliced out, then reverse transcribed into cDNA
  - This cDNA is then reintegrated into the nuclear genome

Repeats

- Transposons (mobile elements): generally of viral origin, integrated into genomes millions of years ago
- Can copy/paste; most are fixed, some are still active
  - Retrotransposon: intermediate step that involves transcription (RNA)
  - DNA transposon: no intermediate step

Retrotransposons

- LTR: long terminal repeat
- Non-LTR:
  - LINEs: Long Interspersed Nucleotide Elements
    - L1 (~6 kbp full length, ~900 bp trimmed version): Approximately 17% of human genome
      - They encode genes to copy themselves
  - SINEs: Short Interspersed Nucleotide Elements
    - *Alu* repeats (~300 bp full length): Approximately 1 million copies = ~10% of the genome
      - They use cell's machinery to replicate
      - Many subfamilies; AluY being the most active, AluJ most ancient

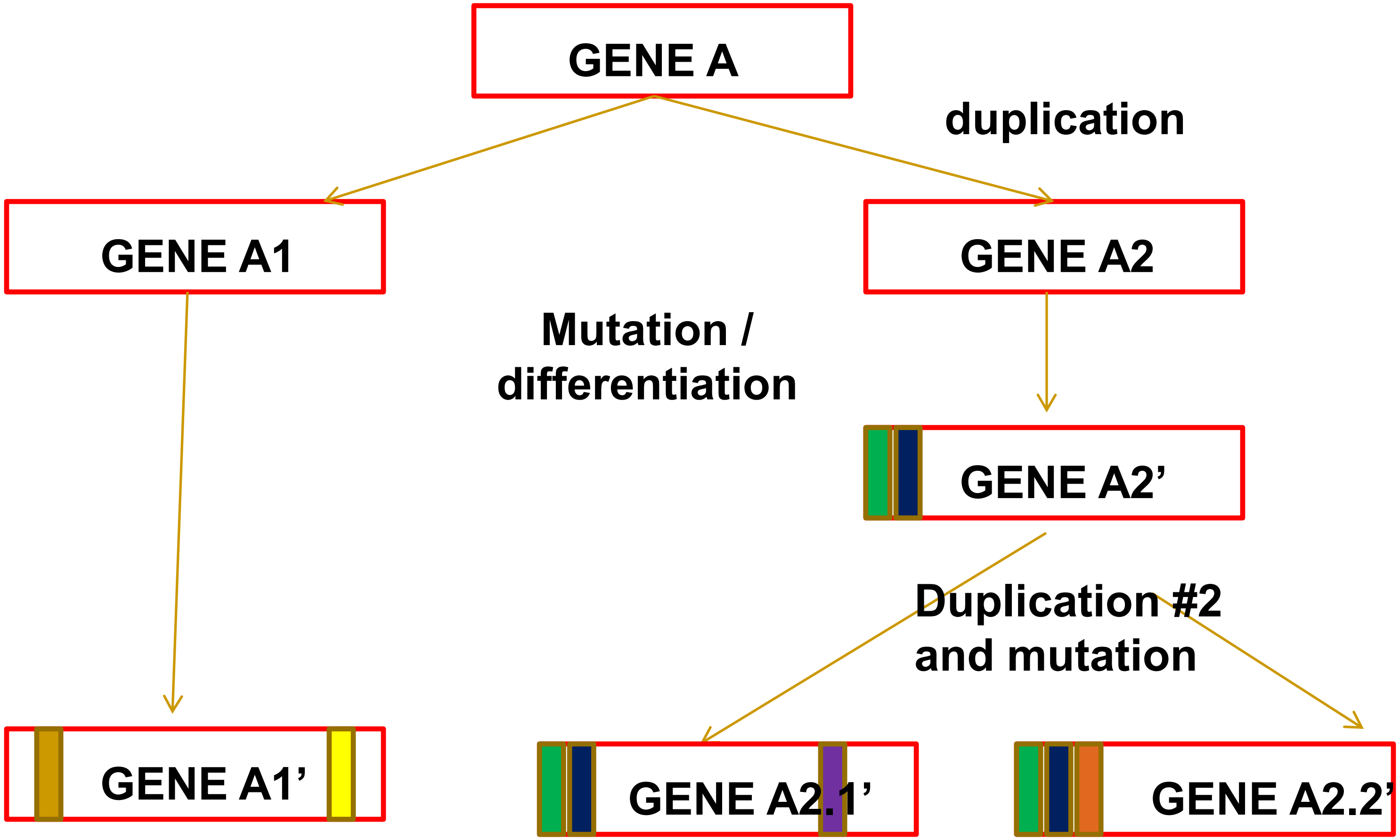
Satellites

- Microsatellites (STR=short tandem repeats) 1-10 bp
  - Used in population genetics, paternity tests and forensics
- Minisatellites (VNTR=variable number of tandem repeats): 10-60 bp
- Other satellites
  - Alpha satellites: centromeric/pericentromeric, 171bp in humans
  - Beta satellites: centromeric (some), 68 bp in humans
  - Satellite I (25-68 bp), II (5bp), III (5 bp)

Segmental duplications

- Low-copy repeats, >1 kbp & > 90% sequence identity between copies
- Covers ~5% of the human genome
  - Both tandem and interspersed in humans, about half inter chromosomal duplications
  - Tandem in mice, no inter chromosomal duplications
- Gene rich
- Provides elasticity to the genome:
  - More prone to rearrangements (and causal)
  - Gene innovation through duplication: Ohno, 1970

Gene innovation through duplication


GENE A
duplication
GENE A1
GENE A2
Mutation / differentiation
GENE A2'
Duplication #2 and mutation
GENE A1'
GENE A2.1'
GENE A2.2'