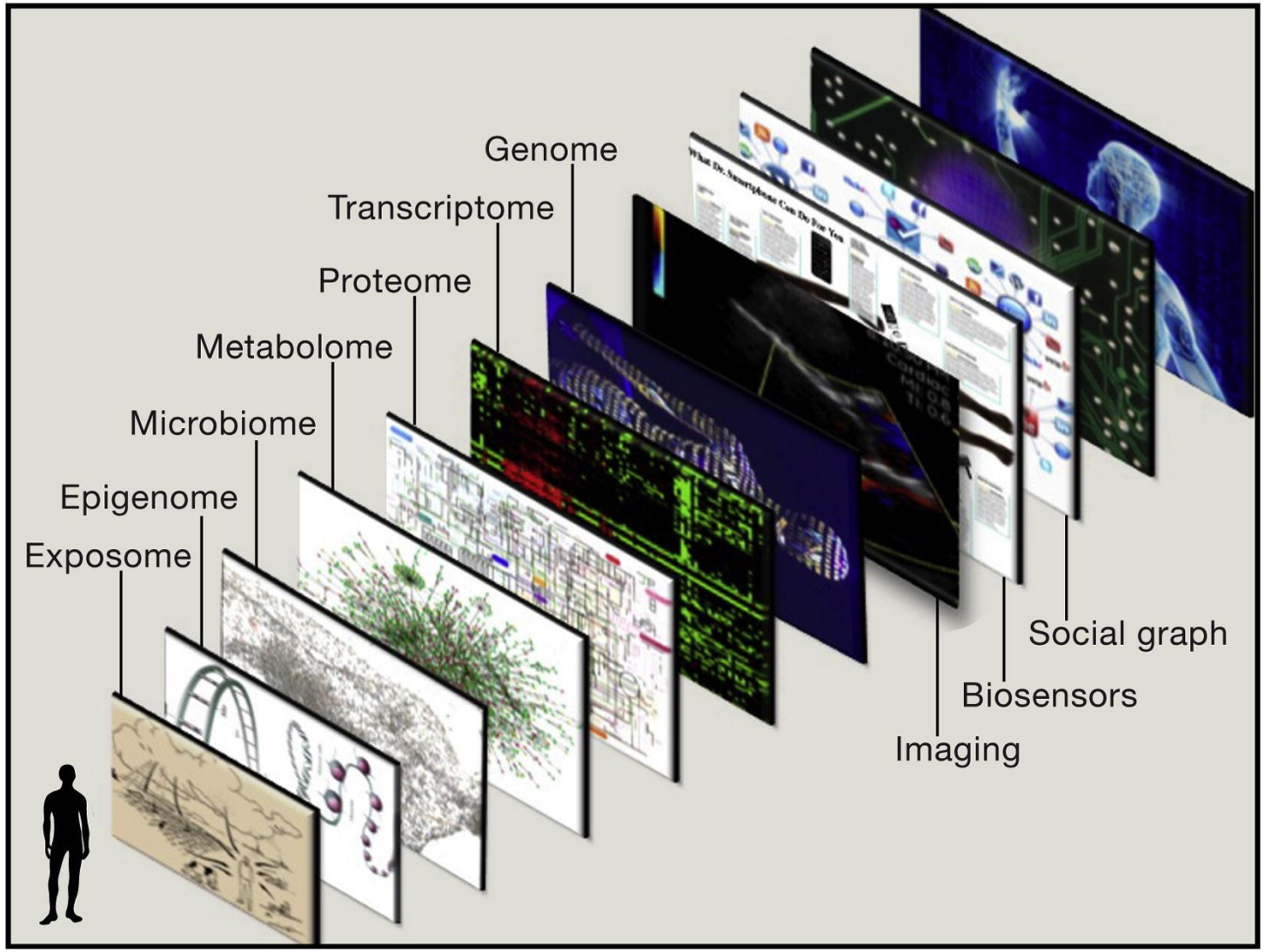
Bioinformatika 1

Uvodno predavanje

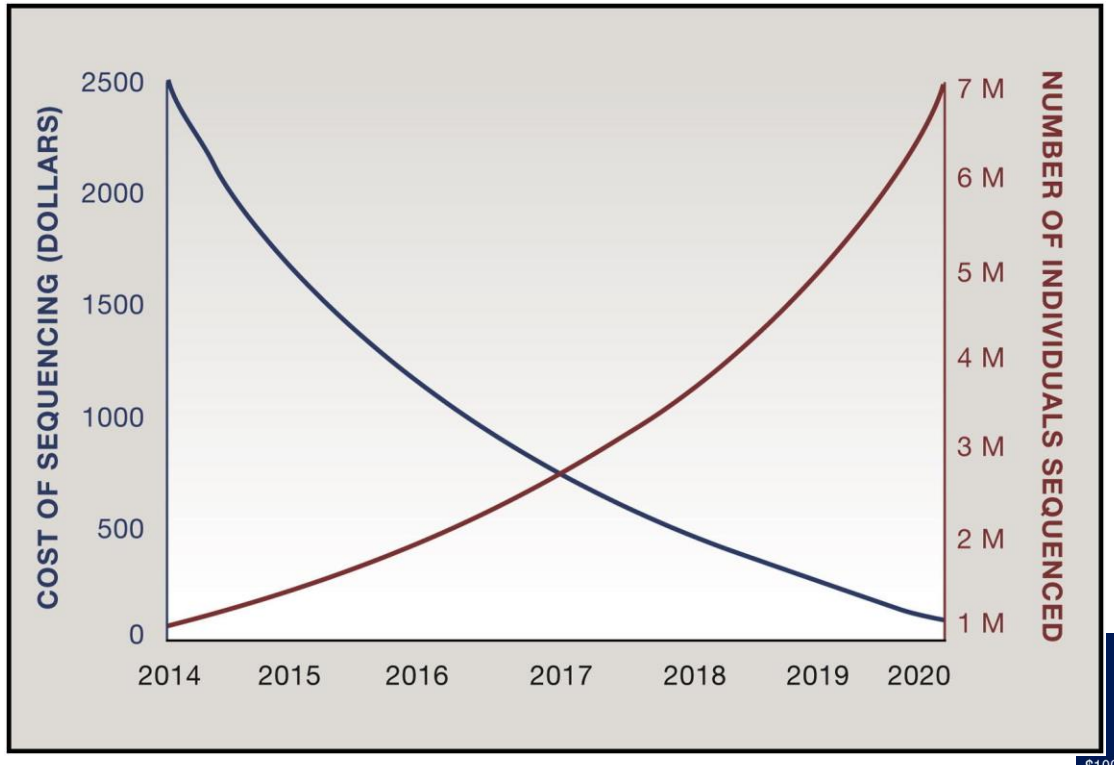
FER, 2020./2021.

Motivacijski primjeri

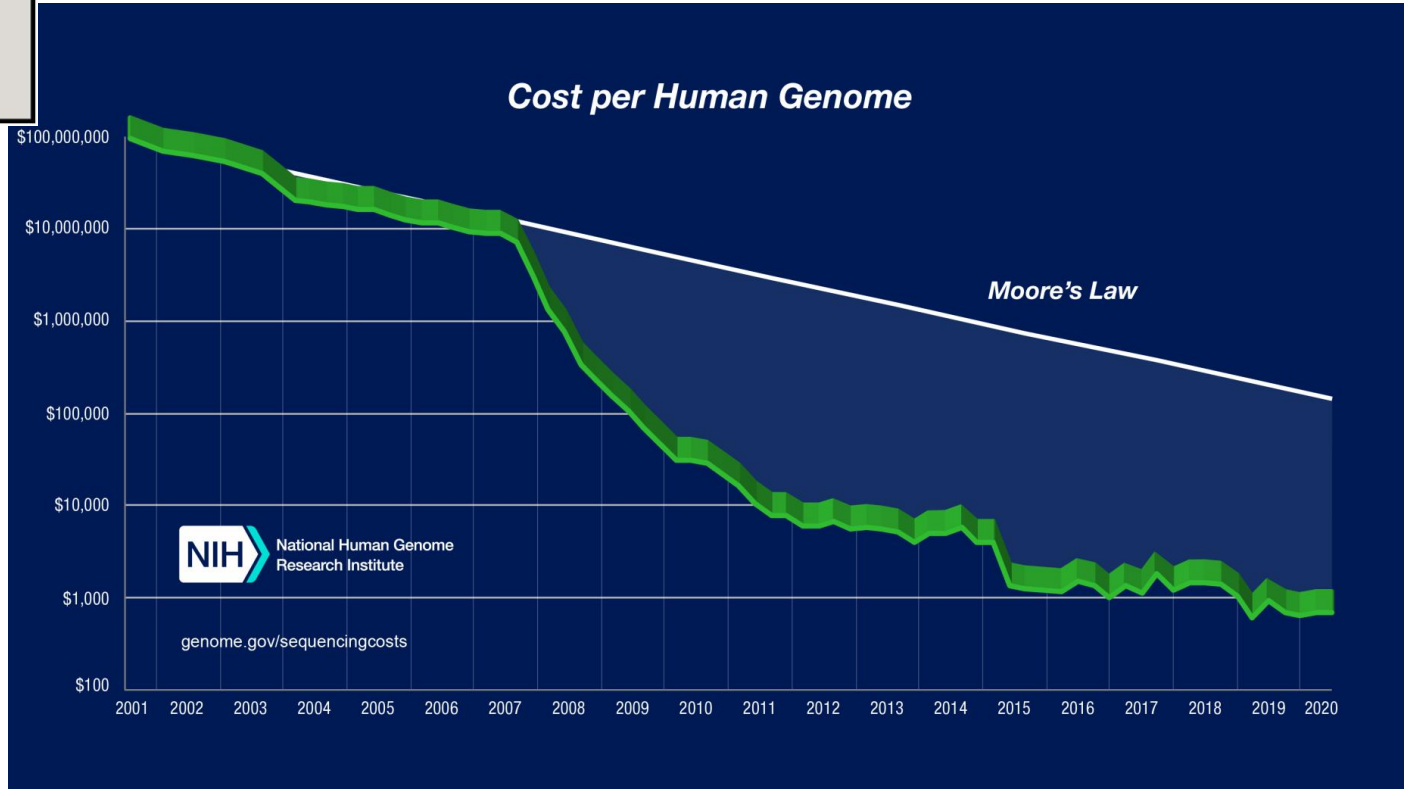
- Individualized medicine
- Epigenetics
- Future professions



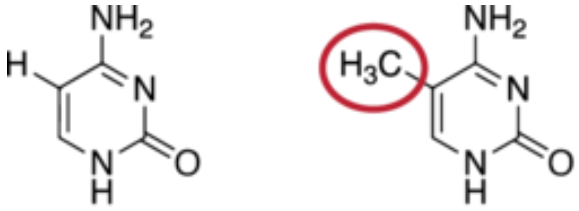
Topol, *Individualized Medicine from Prewomb to Tomb.* Cell 2014 157, 241-253.



Topol, *Individualized Medicine from Prewomb to Tomb.* Cell 2014 157, 241-253.



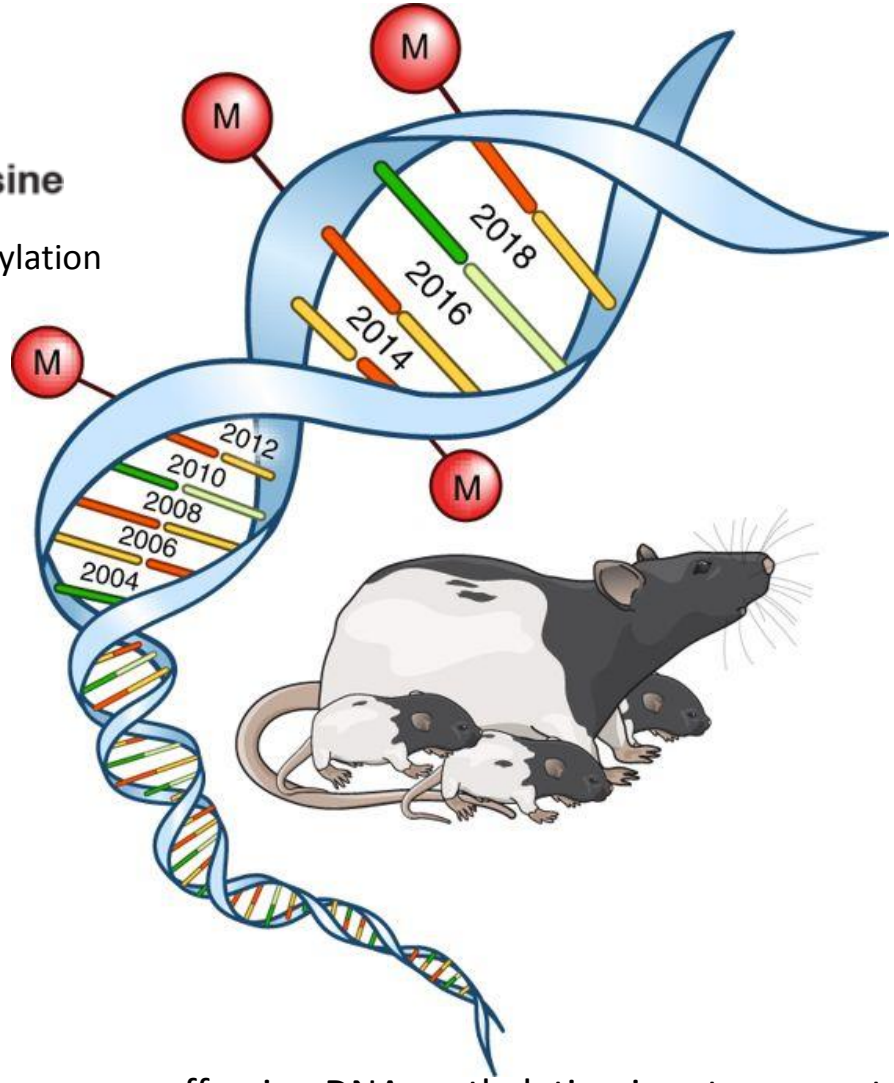
Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) www.genome.gov/sequencingcosts



Cytosine

methylated Cytosine

https://en.wikipedia.org/wiki/DNA_methylation



Genetically identical mice with different DNA methylation patterns causing kinks in the tail of one but not the other.
(Bradbury, 2003. *Human epigenome project – up and running*. PLoS Biology. **1** (3): E82)

Maternal influence on offspring DNA methylation in rats was a starting point for a more dynamic view of epigenetics that has expanded over time (Champagne, 2018. *Beyond the maternal epigenetic legacy*. Nature Neuroscience)

O predmetu

- Predznanje
  - Algoritmi i strukture podataka
  - poznavanje C programskog jezika
- Sadržaj
  - Genomika
  - Poravnavanje nizova – dinamičko programiranje
  - Heuristički algoritmi poravnanja
  - Sufiksna polja i stabla
  - Filogenetska stabla
  - Sastavljanje genoma

Literatura

- Predavanja
- Skripta
- Knjige
  - J. Pevsner, **Bioinformatics and Functional Genomics,** 3rd edition, Wiley-Blackwell (2015)
  - D. Gusfield, **Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology**, Cambridge University Press (1997)
  - N.C. Jones, P. J. Pevzner, **An Introduction to Bioinformatics Algorithms**, MIT Press (2004)
  - R. C. Deonier, S.Tavaré, M.S. Waterman, **Computational Genome Analysis**, Springer (2005)

Dodatni materijali

- Introduction to Biology - The Secret of Life (https://www.edx.org/course/mit/7-00x/introduction-biology-secret-life/1014)

- Rosalind – bioinformatički zadatci (http://rosalind.info/problems/locations/)

Predavači

- prof. dr. sc. Mile Šikić
- izv. prof. dr. sc. Mirjana Domazet-Lošo
- doc. dr. sc. Krešimir Križanović

Ocjenjivanje

- Kontinuirana provjera

Naziv provjere	Bodovi	Prag
MI (90 min)	25	0
ZI (90 min)	35	14
Projekt	40	0
Projekt*	100	0

- Ispitni rok

Naziv provjere	Bodovi	Prag
Pismeni ispit	60	15
Projekt	40	0

Projekt

- projekt za 40 bodova: rad u grupama do 5 studenata
- projekt za 100 bodova: rad u grupama do 2 studenta
- upute i prijedlozi tema na stranicama predmeta
- većina projekata se svodi na implementaciju algoritama u različitim programskim jezicima
- prezentacija projekta

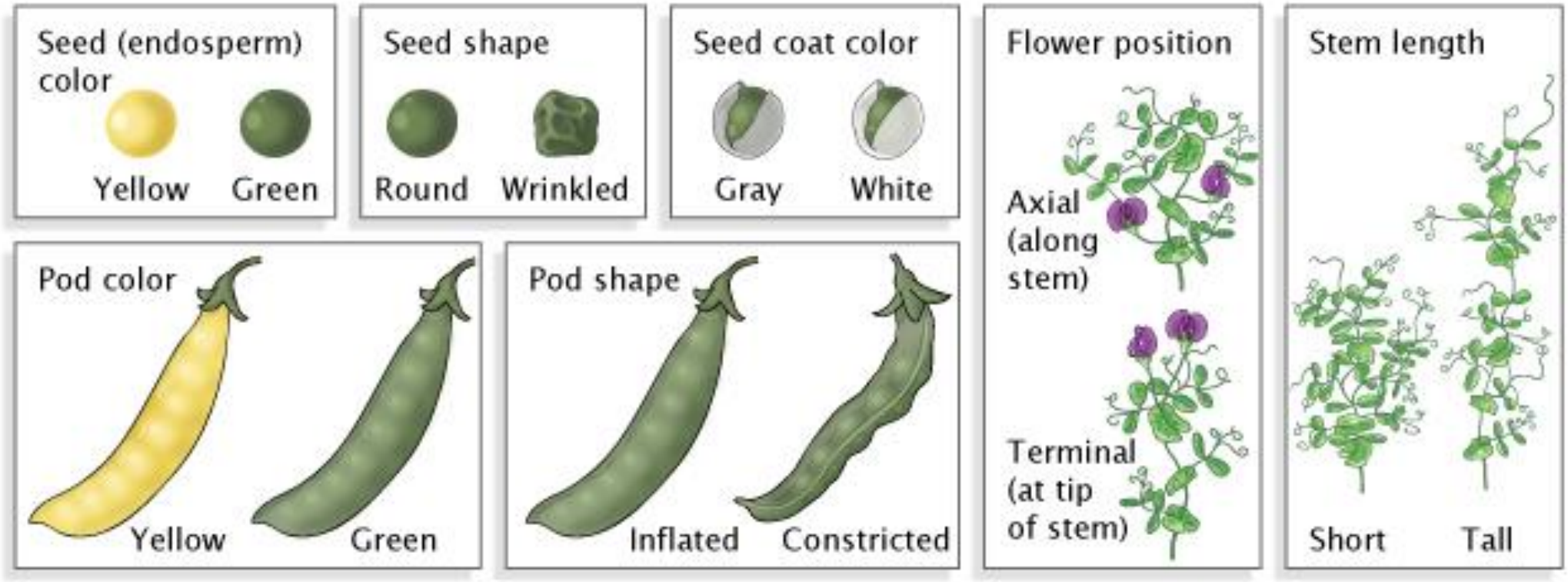
Ocjenjivanje

Ocjena	Minimalan broj bodova
2	50
3	60
4	75
5	90

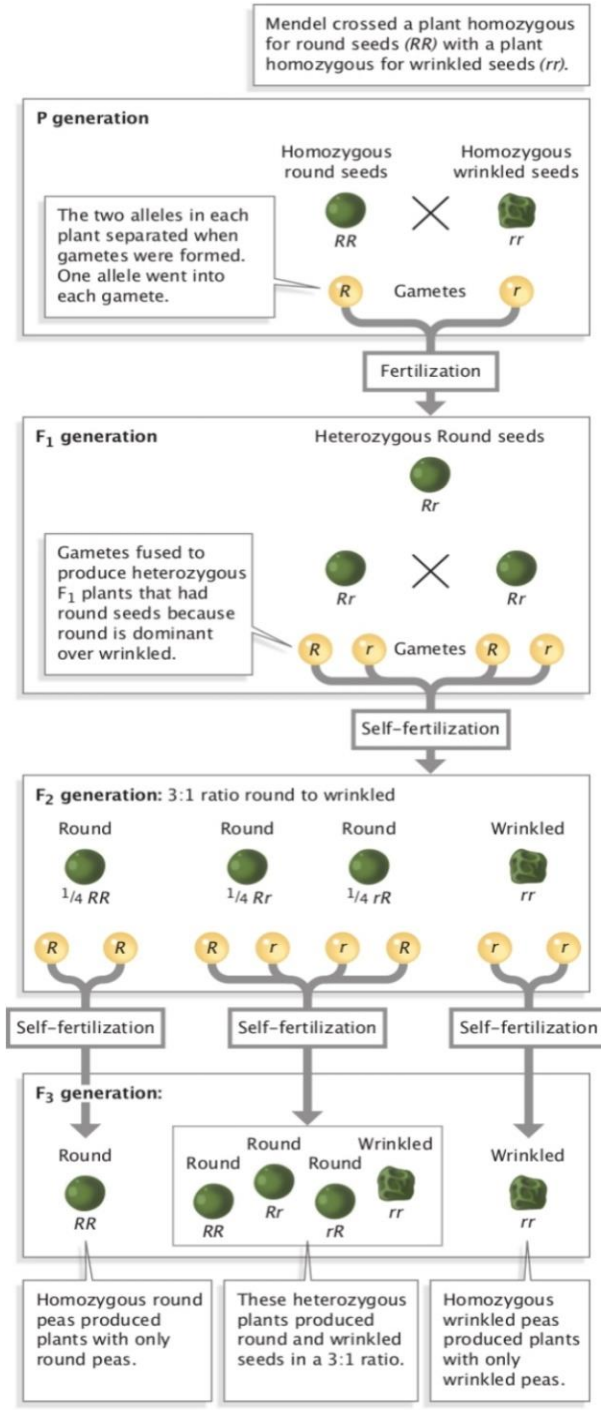
Uvod u genetiku

Klasična i moderna genetika

- sličnost potomaka i roditelja
- selektivno uzgajanje biljaka i životinja
- Mendel – principi nasljeđivanja
  - Objavio rad 1866.
  - Radio različita križanja graška
  - Sljedećih 35 godina skupio 3 citata ☺
  - Ponovno otkriven 1900.



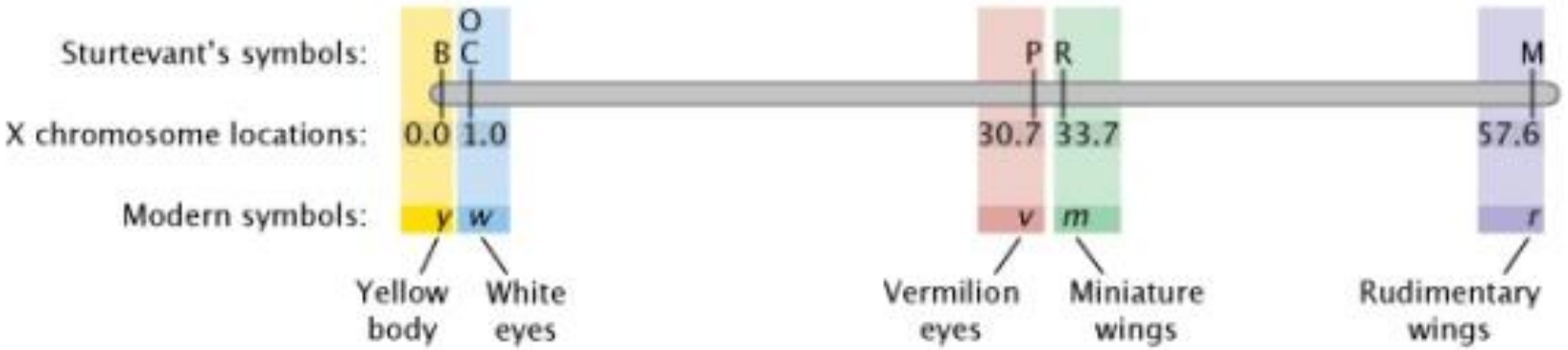
Miko, I. (2008) Gregor Mendel and the principles of inheritance. Nature Education 1(1):134



Miko, I. (2008) Gregor Mendel and the principles of inheritance. Nature Education 1(1):134

Klasična i moderna genetika

- U 20 st. genetika je postala važan biološki alat koristeći mutante u cilju razumijevanja procesa. Taj rad uključuje:
  - Analizu nasljeđivanje u populacijama.
  - Analiziranje evolucijskih procesa.
  - Identifikacija gena koji kontroliraju pojedine korake u procesima.
  - Mapiranje gena.
  - Utvrđivanje produkata gena.
  - Analiza molekularnih svojstava gene i regulacije ekspresije gena.



Sturtevant's *Drosophila* gene map. (Lobo and Shaw, 2008)

Klasična i moderna genetika

- Važni doprinosi sredinom i krajem 20. st.:
  - Određivanje strukture DNA (Wattson & Crick, 1953)
  - Konstrukcija prve rekombinirane DNA molekule (Berg, 1972)
  - Prvo kloniranje rekombinirane DNA molekule (Boyer & Cohen, 1973)
  - Metoda lančane reakcije polimerazom (engl. polymerase chain reaction, PCR) za umnažanje DNA molekula (Mullis, 1986)

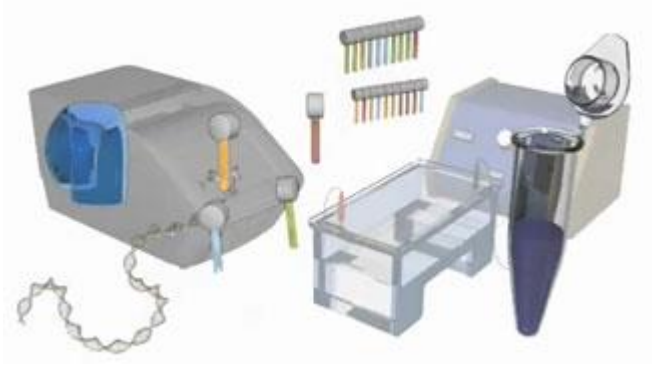


Figure 1: The various components required for PCR include a DNA sample, DNA primers, free nucleotides called ddNTPs, and DNA polymerase.

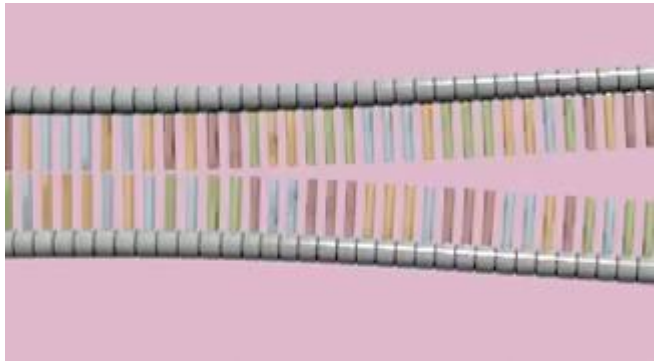


Figure 2: When heated, the DNA strands separate.

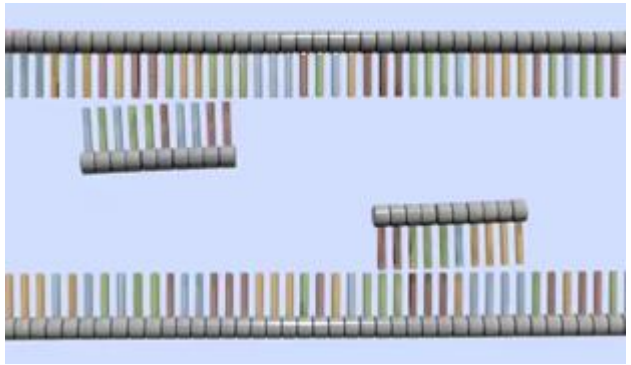


Figure 3: When the solution is cooled, the primers anneal.

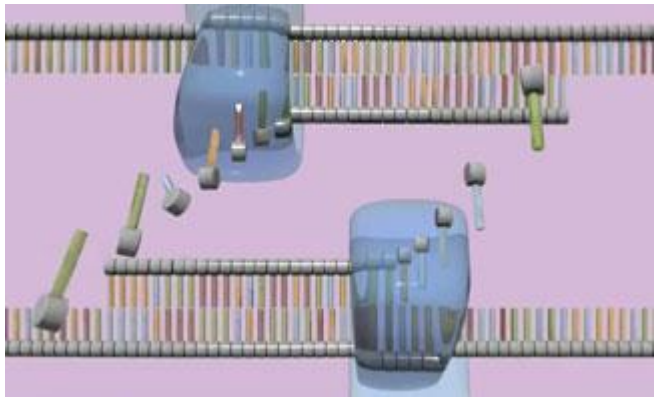


Figure 4: DNA polymerase attaches to each primer and assembles dNTPs to build a new strand.

http://www.nature.com/scitable/topicpage/scientists-can-make-copies-of-a-gene-6525968

Klasična i moderna genetika

- Najnovija otkrića - sekvenciranje:
  - Za sve veći broj organizama sekvenca je određena.
  - Poznavanje individualnih gena i njihove regulacije bit će važno za temeljna biološka istraživanja kao i medicinske primjene (medicinska genetika).
- Etička, pravna i socijalna pitanja

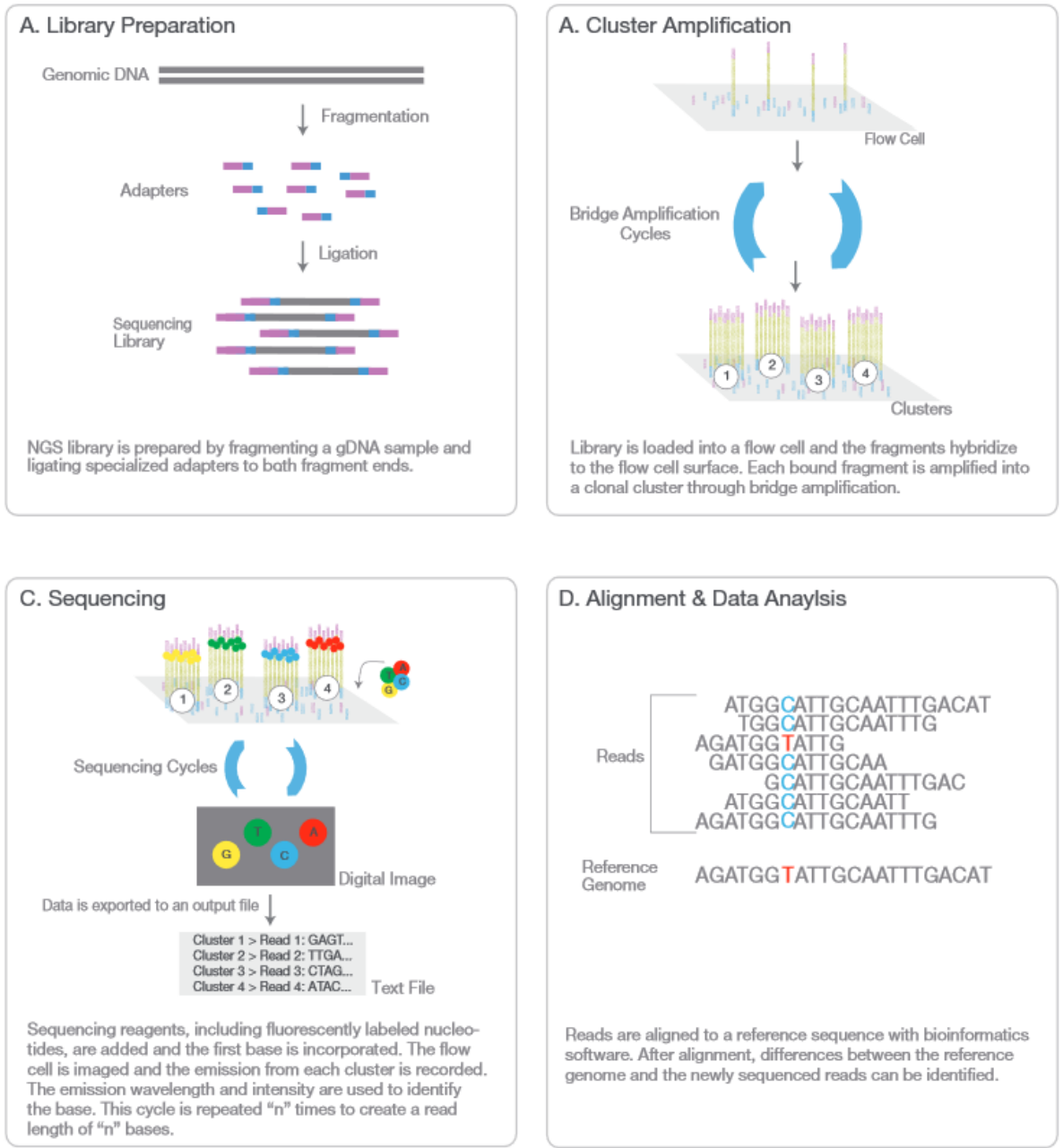


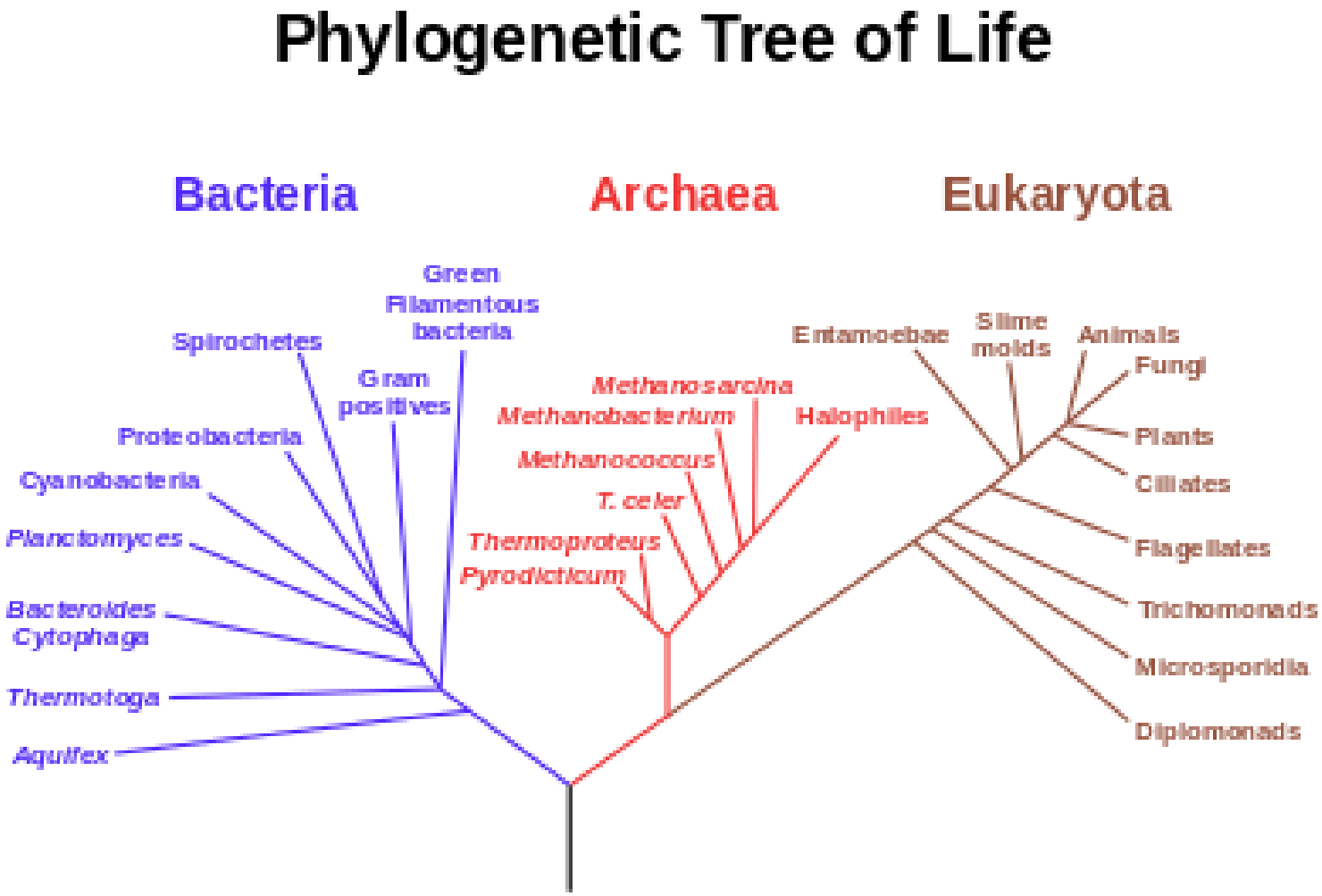
Figure 3: Next-Generation Sequencing Chemistry Overview.

http://www.illumina.com

DNA, geni i kromosomi

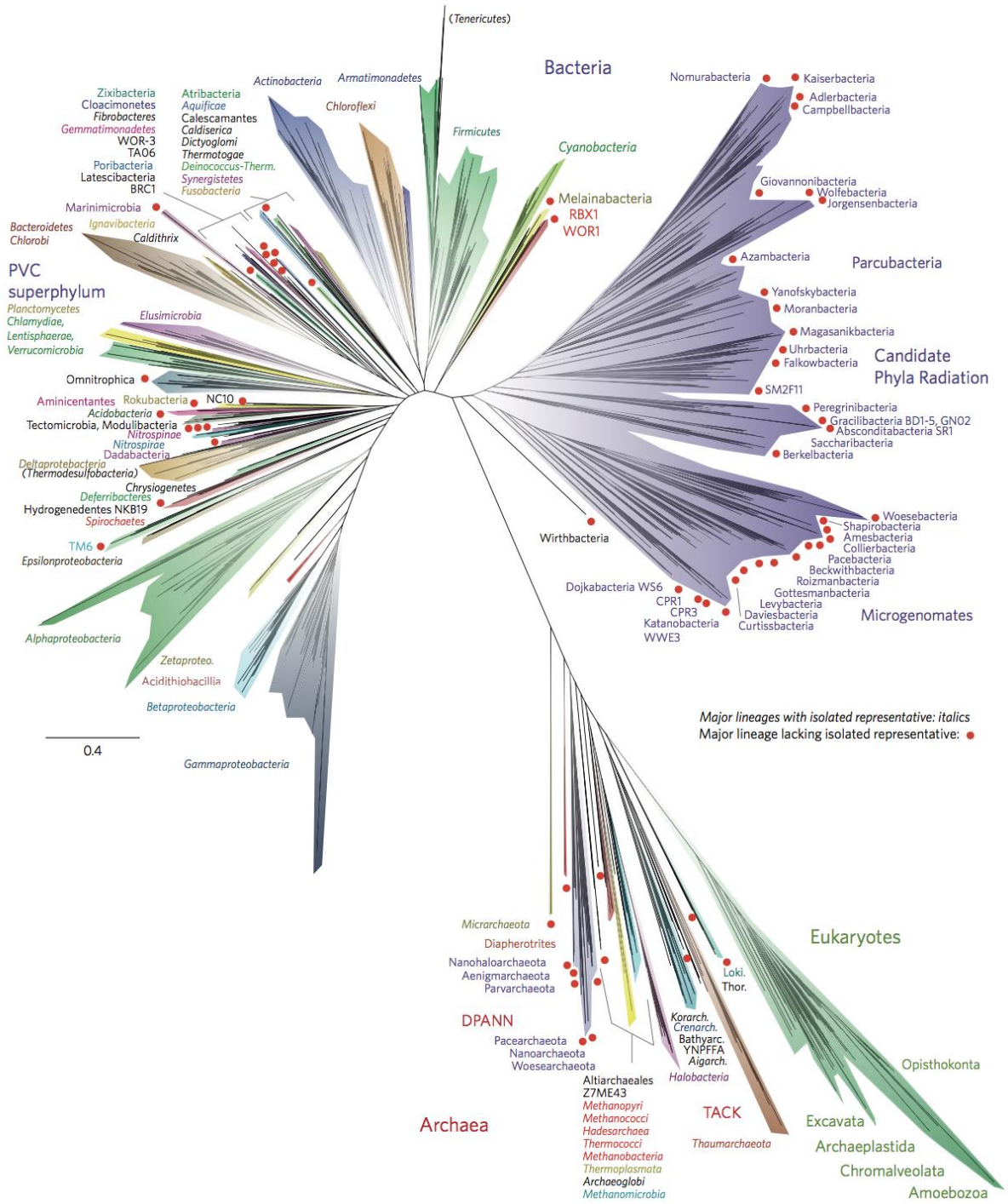
- Genetički materijal
  - DNA – eukarioti i prokarioti
  - DNA ili RNA - virusi
- DNA ima dva lanca tvorenih od nukleotida koji se sastoje od šećera deoksiriboze, fosfatne grupe i baze (adenin, timin, gvanin i citozin)

Filogenetsko stablo života - nekadašnji prikaz



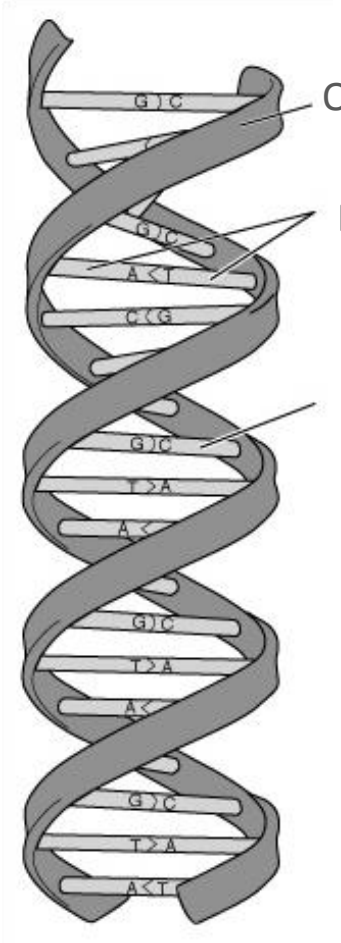
https://en.wikipedia.org/wiki/Three-domain_system

Filogenetsko stablo života



Hug et al. 2016. *A new view of the tree of life.* Nature Microbiology. 1: 16048.

DNA


Okosnica šećer - fosfat
Parovi dušičnih baza
Pojedina dušična baza

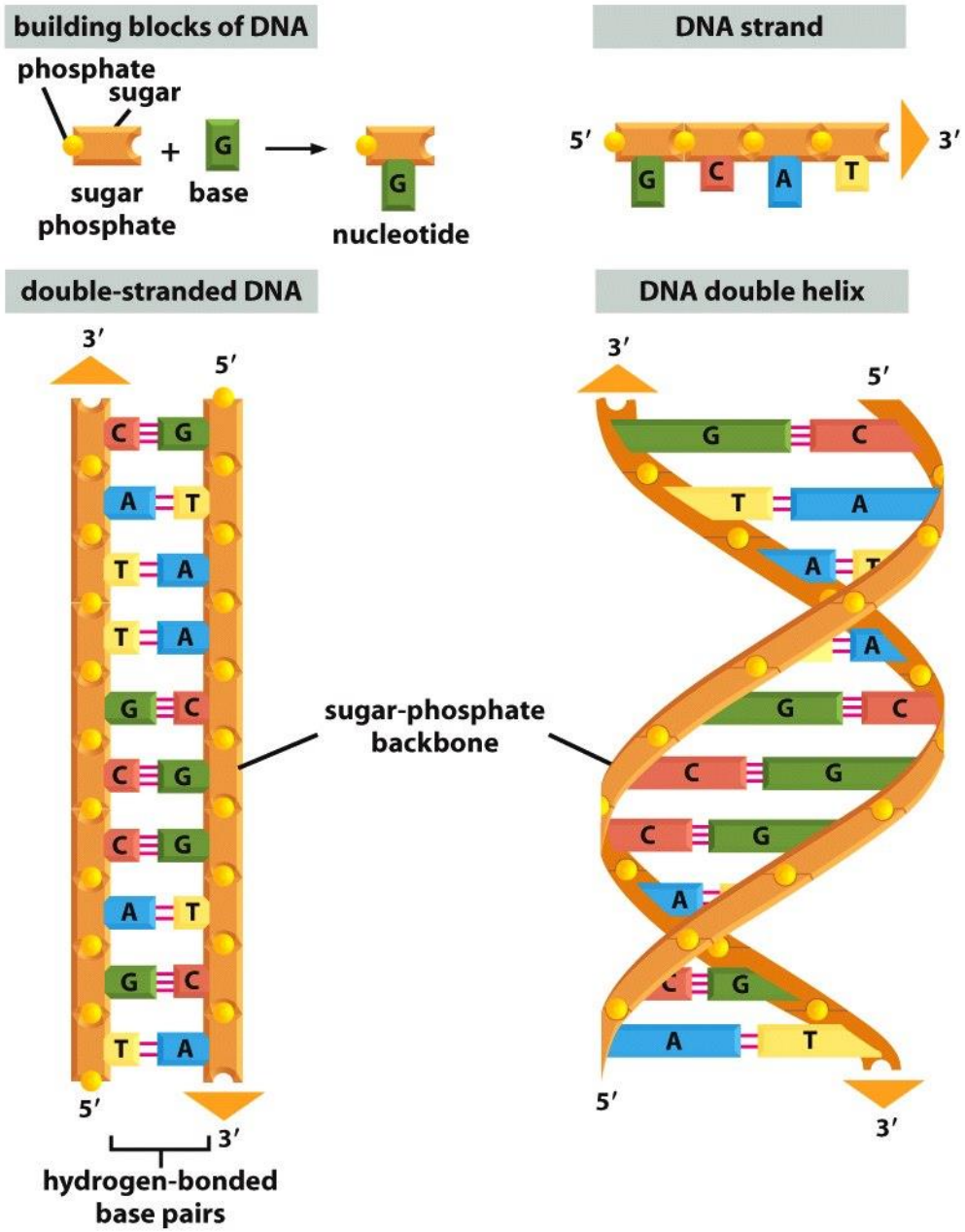


Figure 4-3 *Molecular Biology of the Cell* (© Garland Science 2008)

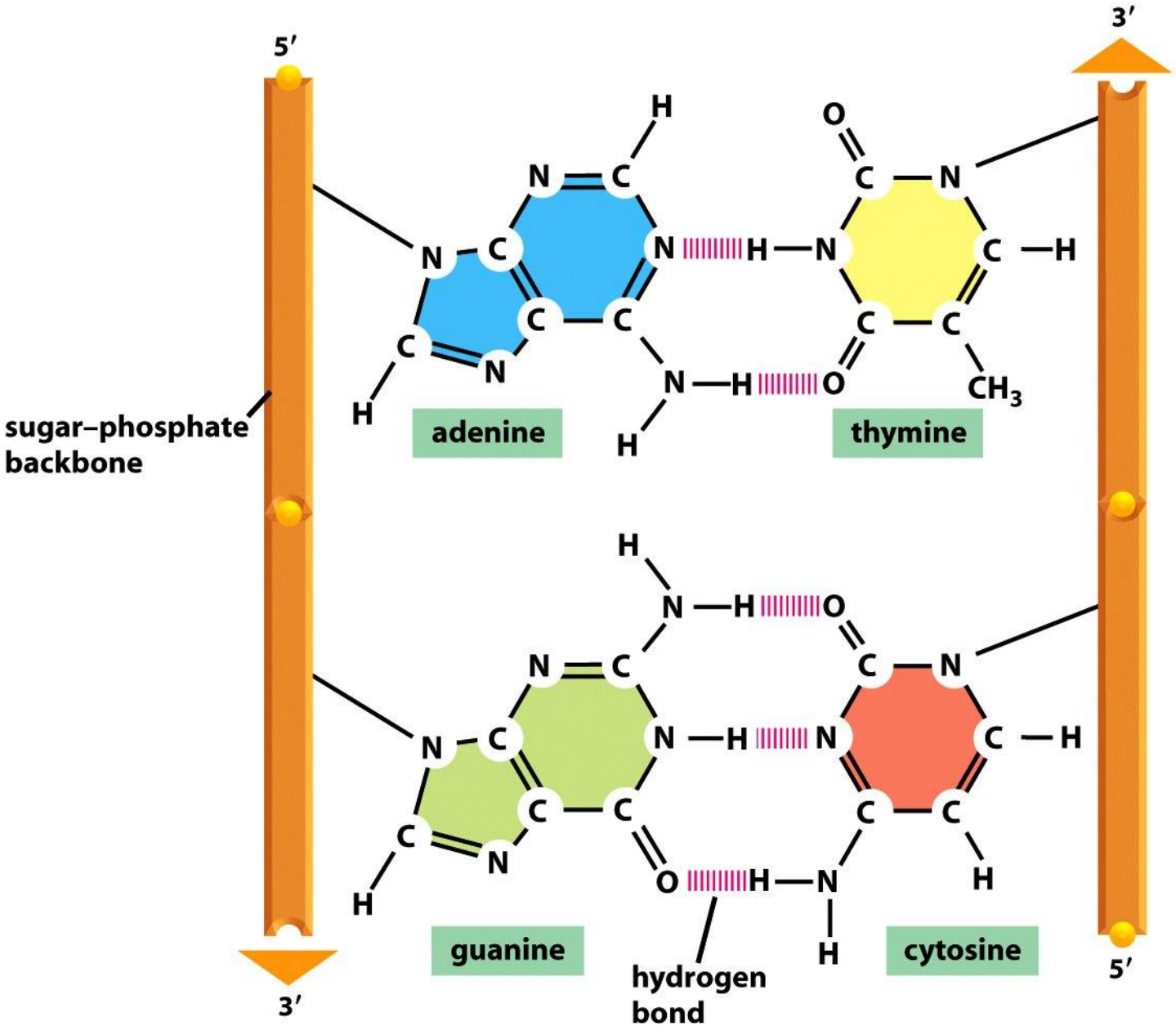


Figure 4-4 *Molecular Biology of the Cell* (© Garland Science 2008)

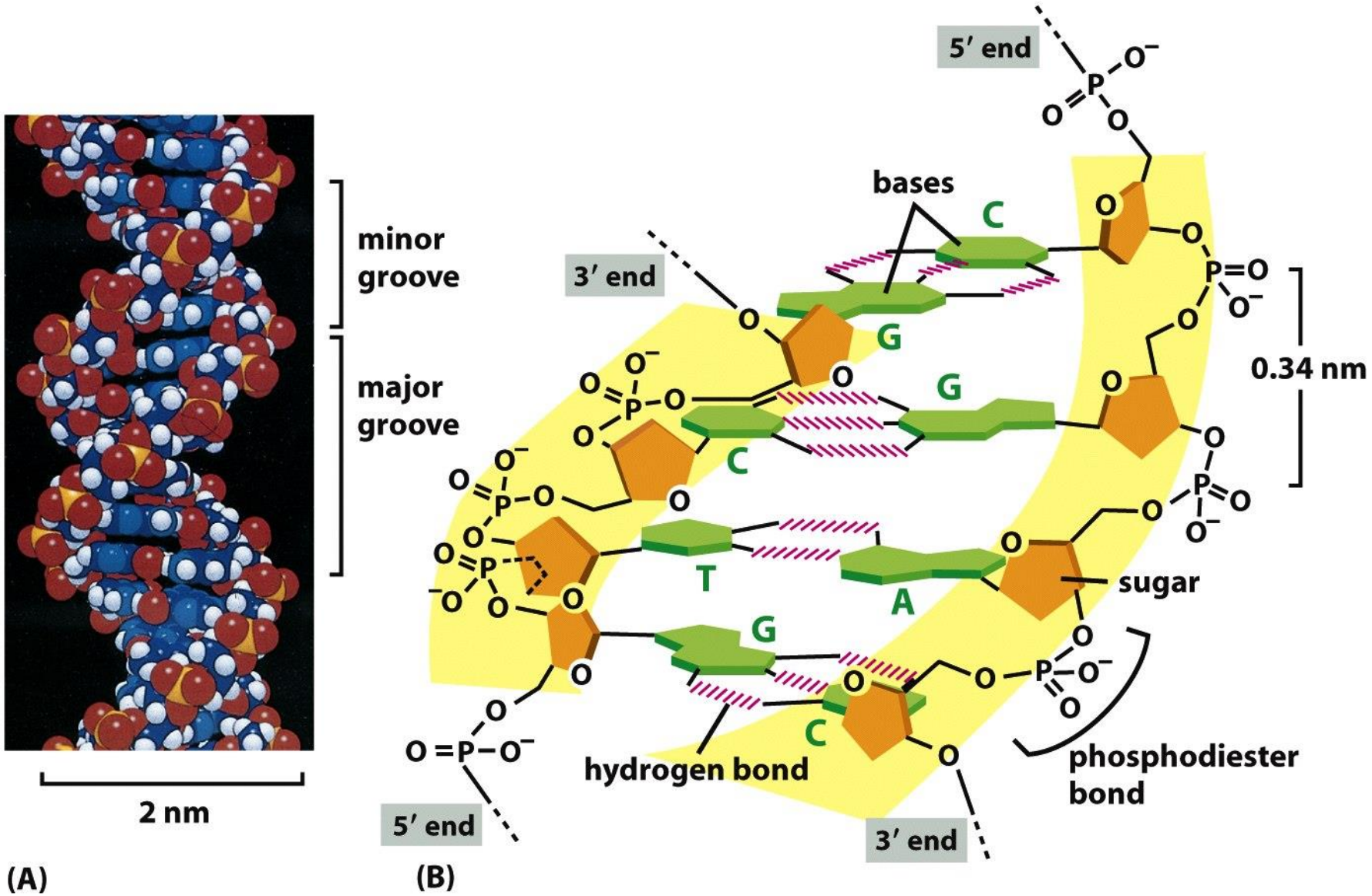


Figure 4-5 *Molecular Biology of the Cell* (© Garland Science 2008)

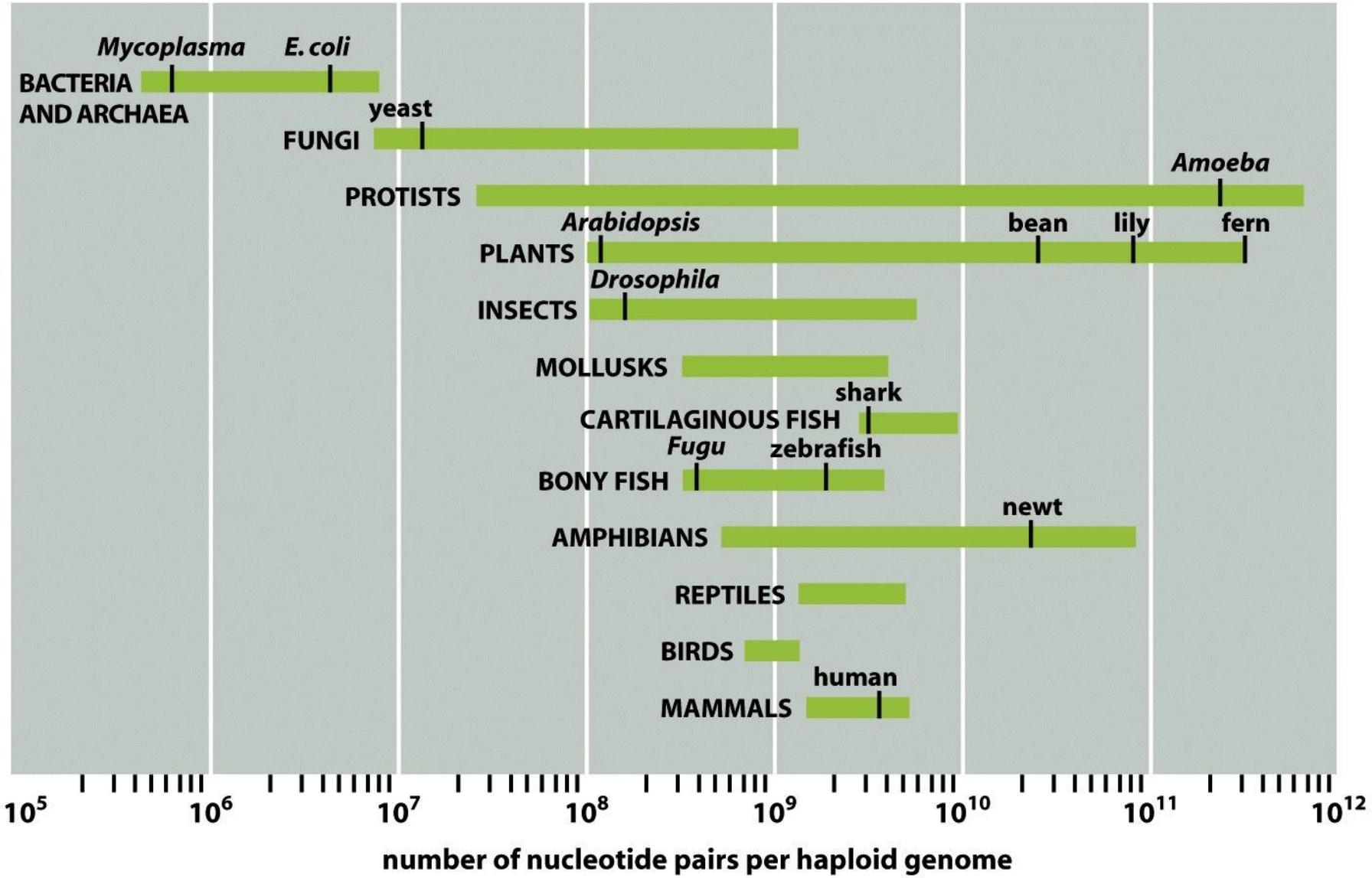
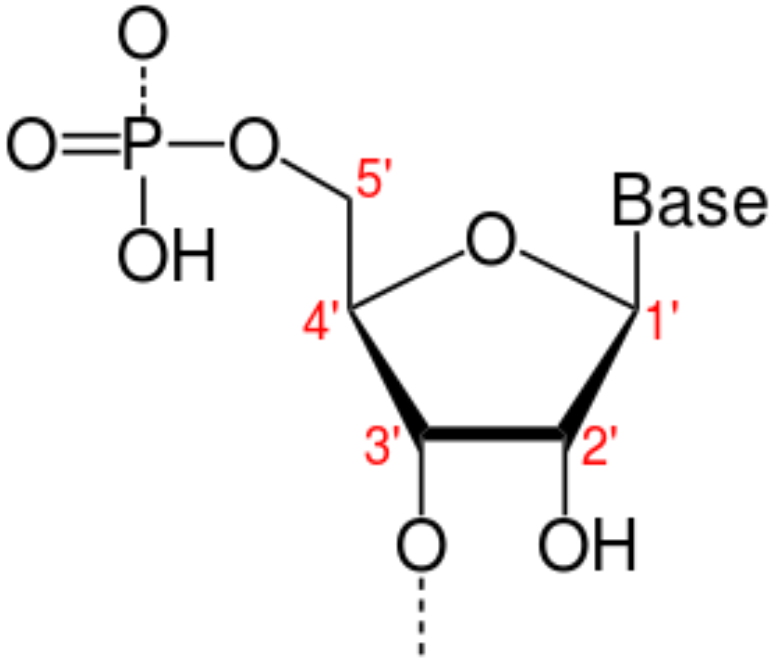


Figure 1-37 *Molecular Biology of the Cell*, Fifth Edition (© Garland Science 2008)

DNA smjer



- Imenovanje krajeva DNA molekule prema atomima ugljika u prstenu (šećer).
- Lanci idu u suprotnim smjerovima: 5’ – 3’ odnosno 3’ – 5’
- Čitanje DNA uvijek od 5’ prema 3’ kraju jednog lanca.

DNA, geni i kromosomi

- Četiri baze u DNA : A (adenin), G (gvanin), C (citozin) i T (timin).
  - U RNA U (uracil) zamjenjuje T (timin).
  - Sekvenca baza određuje genetičku informaciju.
  - Geni su specifične sekvence nukleotida koje prenose osobine s roditelja na potomstvo.

DNA, geni i kromosomi

- Genetički materijal u stanicama je organiziran u kromosome
  - Prokarioti uglavnom imaju jedan cirkularan kromosom.
  - Eukarioti uglavnom imaju:
    - Linearne kromosome u jezgri, pri čemu različite vrste imaju različit broj kromosoma.
    - DNA u organelima (npr. mitohondriji i kloroplasti) koji su obično cirkularne molekule.

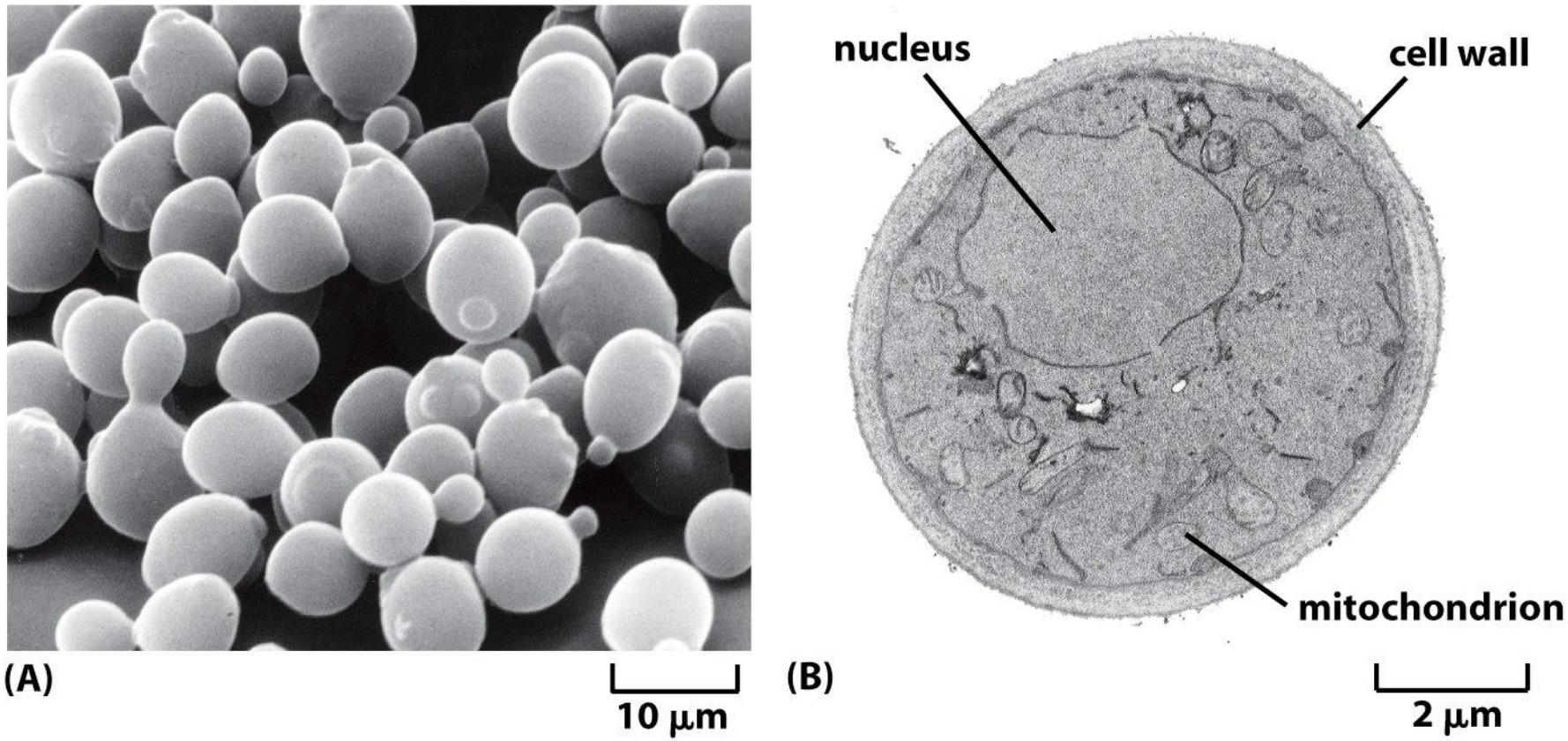
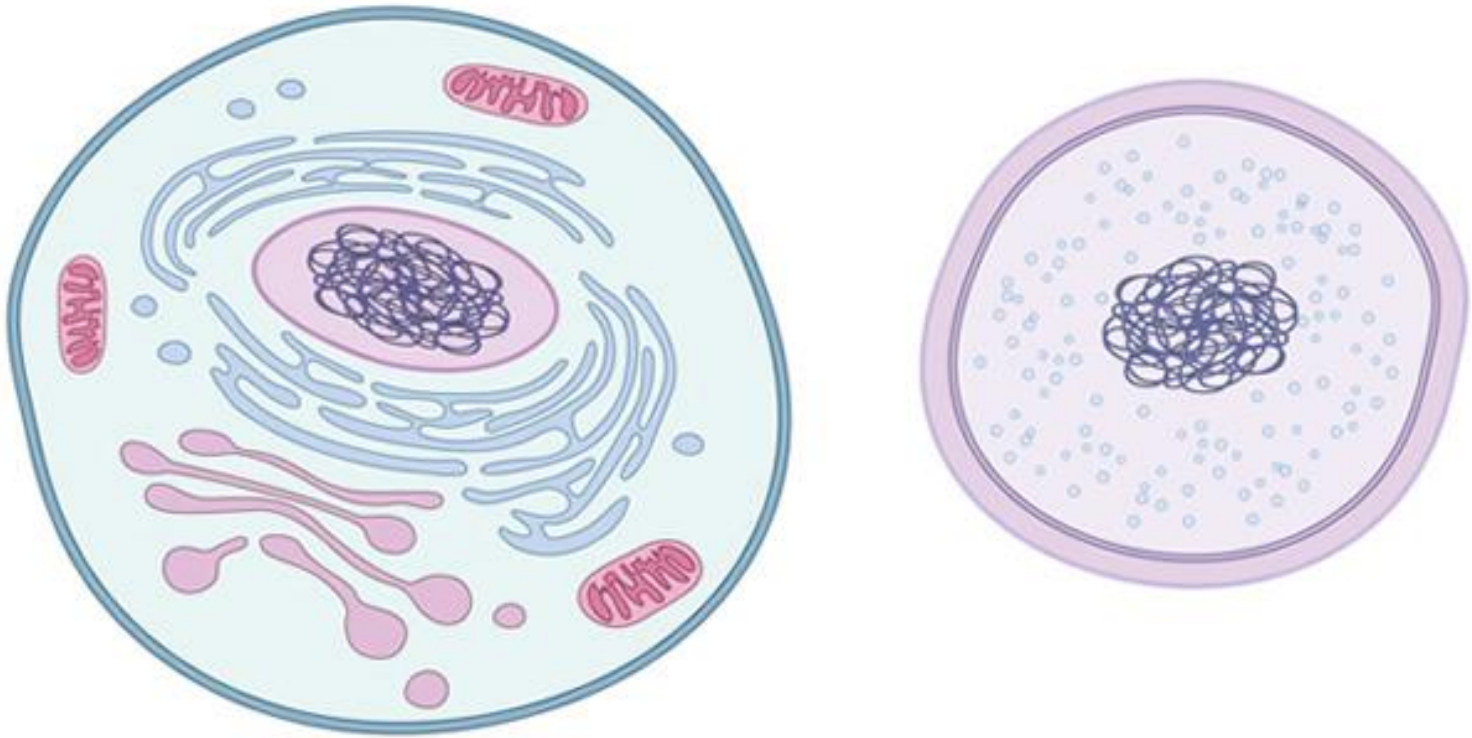


Figure 1-42 *Molecular Biology of the Cell*, Fifth Edition (© Garland Science 2008)



Comparing basic eukaryotic and prokaryotic differences (Fuerst, 2010)

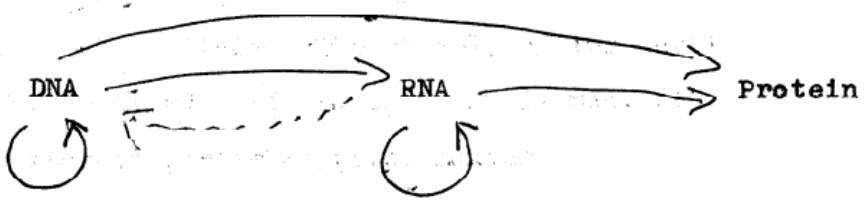
Centralna dogma – prijenos informacija

Ideas on Protein Synthesis (Oct. 1956)

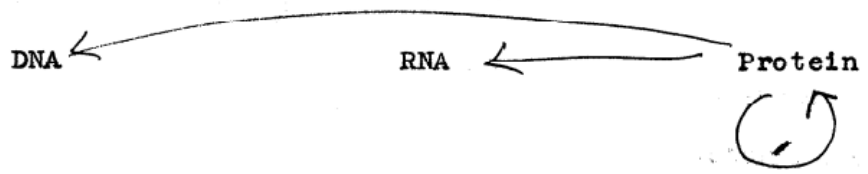
The Doctrine of the Triad.

The Central Dogma: "Once information has got into a protein it can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it.

That is, we may be able to have



but never



where the arrows show the transfer of information.

Crick FHC (1956). Ideas on protein synthesis.
Available at: http://profiles.nlm.nih.gov/SC/B/B/F/T/_/scbbft.pdf (acknowledged by Crick in 1958)

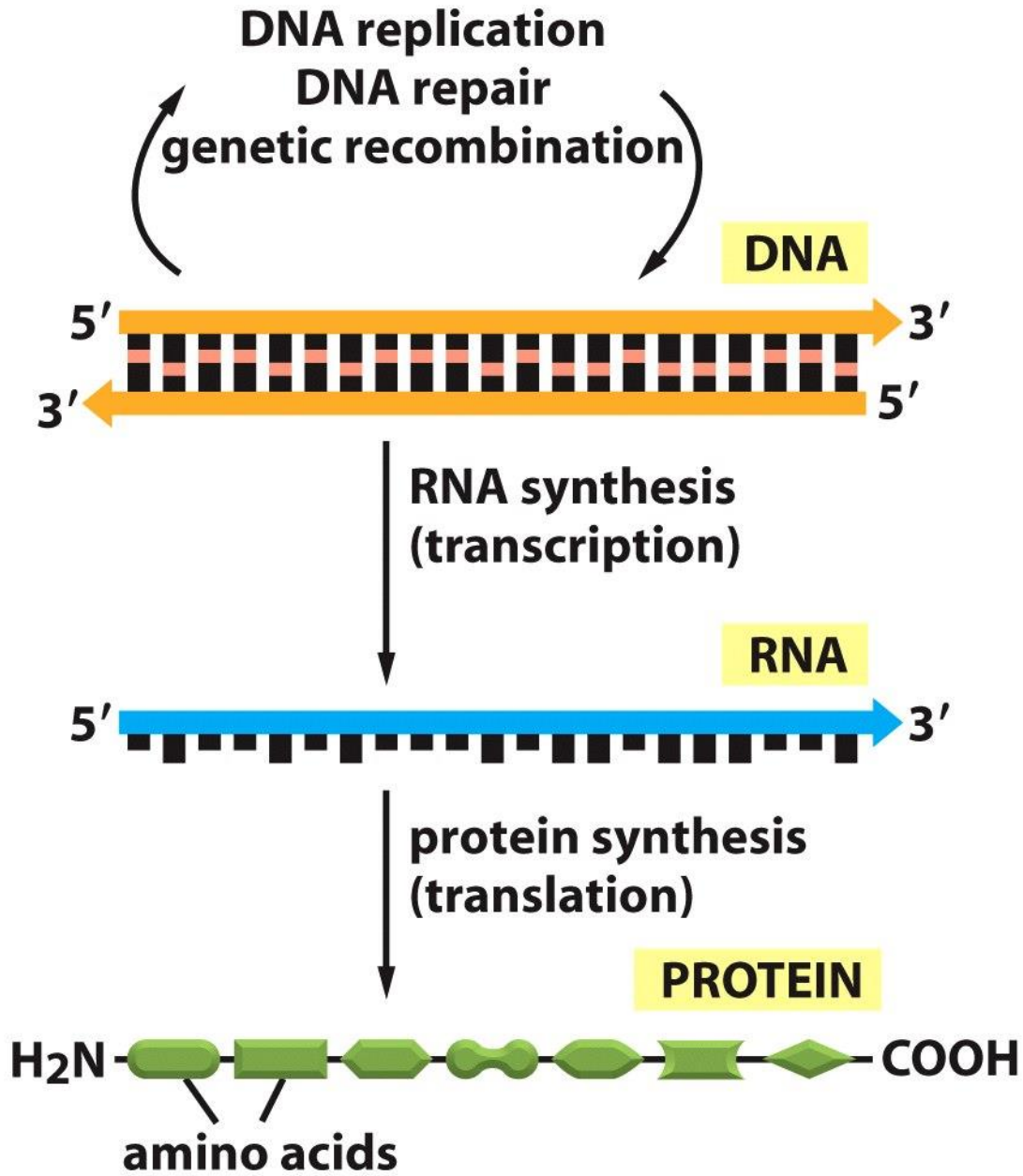


Figure 6-2 *Molecular Biology of the Cell* (© Garland Science 2008)

Centralna dogma – prijenos informacija

Opći	*Specijalni*	*Nepoznat*
DNA → DNA	RNA → DNA	protein → DNA
DNA → RNA	RNA → RNA	protein → RNA
RNA → protein	DNA → protein	protein → protein

Ekspresija gena

- Transkripcija: sinteza RNA (prepisivanje DNA)
  - Sinteza mRNA koristeći gene DNA molekule kao predložak.
  - U jezgri eukariota.
- Translacija sinteza proteina (prevođenje RNA)
  - Sinteza proteinskoga lanca koristeći genetički kod mRNA molekule kao uputu.

Ribonukleinska kiselina

- Nalazi se u cijeloj stanici:
  - jezgra
  - mitohondriji
  - kloroplast
  - ribosomi
  - citoplazma

Glavne vrste

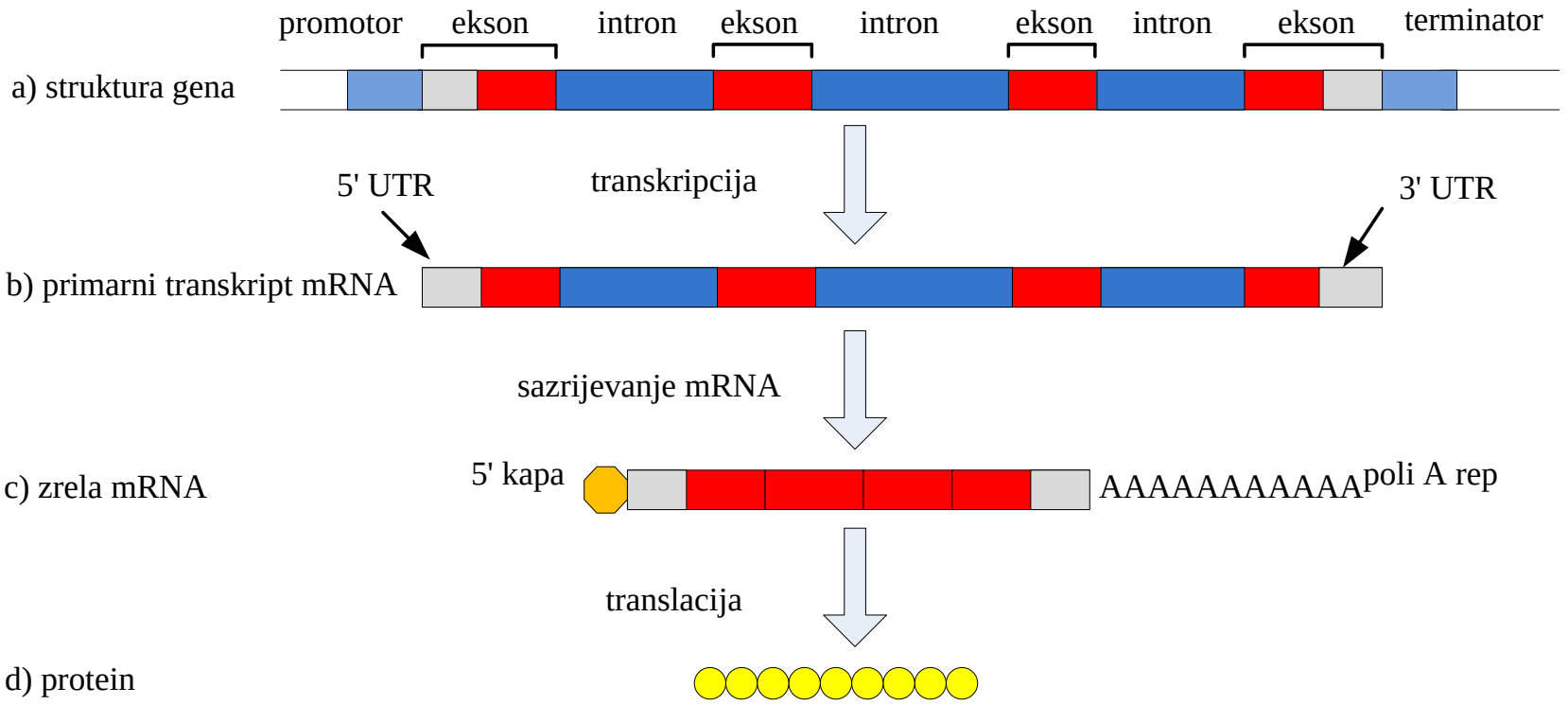
- Glasnička RNA (mRNA) <5%
- Ribosomska RNA (rRNA) do 80%
- Transportna RNA (tRNA) oko 15%
- Male RNA molekule u jezgri (snRNA): imaju ulogu u obradi mRNA u jezgri nakon transkripcije.
- Često svu RNA u stanici nazivamo transkriptomom.

Transkripcija : Sinteza mRNA (i drugih RNA)

- Koristi enzim RNA polimerazu.
- Tvori komplementarni lanac mRNA.
- Počinje na mjestu promotora koji signalizira blizinu gena (oko 20 do 30 nukleotida).
- Nakon što smo došli do kraja gena postoji terminirajuća sekvenca koja kaže RNA polimerazi da zaustavi prepisivanje.

Uređivanje mRNA

- U prokariotima prepisana mRNA ide direktno prema ribosomima u citoplazmi.
- U eukariotima prepisana mRNA se nalazi u jezgri i dugačka je oko 5000 nukleotida.
- Kada se ta ista mRNA prevodi u ribosomu dugačka je 1000 nukleotida.
- mRNA se uređuje.
- Dijelovi koji se zadržavaju za ekspresiju gena nazivaju se eksoni (exons = expressed).
- Dijelovi koji se uklanjanju se nazivaju introni.


promotor
ekson
intron
ekson
intron
ekson
intron
ekson
terminator
a) struktura gena
5' UTR
transkripcija
3' UTR
b) primarni transkript mRNA
sazrijevanje mRNA
c) zrela mRNA
5' kapa
AAAAAAAAAAA
poli A rep
translacija
d) protein

Genetski kod

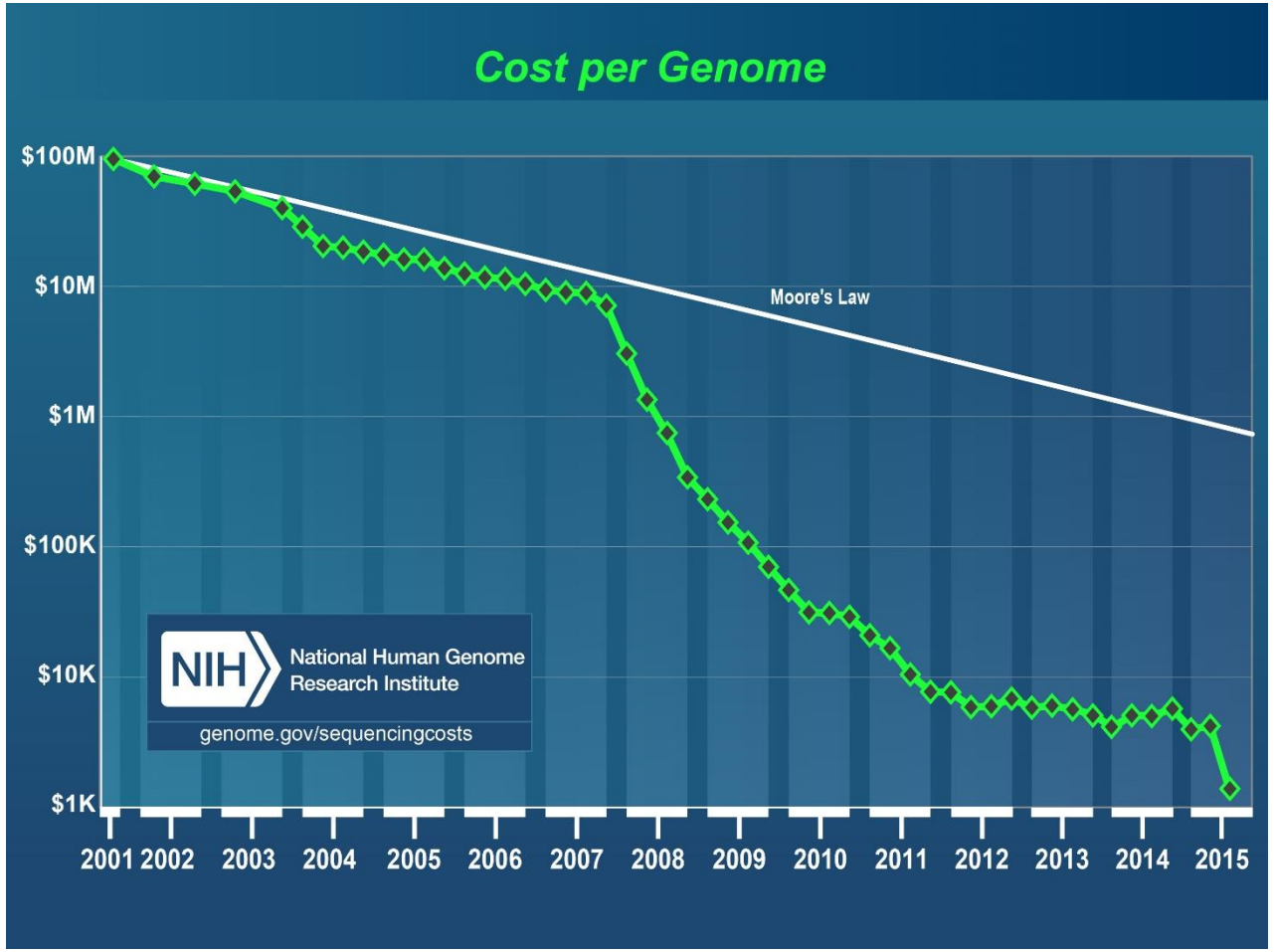
- Genetski kod se sastoji od slijeda baza nađenih uzduž mRNA molekule.
- Postoji samo 4 slova za taj kod (A, G, C i U).
- Kod mora biti dovoljno kompleksan da predstavlja 20 različitih aminokiselina koje se koriste za građu proteina.

Genetski kod

prva baza kodona (5')	druga baza kodona				treća baza kodona (3')
	U	**C**	**A**	**G**	
U	Phe	Ser	Tyr	Cys	**U**
	Phe	Ser	Tyr	Cys	**C**
	Leu	Ser	STOP	STOP	**A**
	Leu	Ser	STOP	Trp	**G**
C	Leu	Pro	His	Arg	**U**
	Leu	Pro	His	Arg	**C**
	Leu	Pro	Gln	Arg	**A**
	Leu	Pro	Gln	Arg	**G**
A	Ile	Thr	Asn	Ser	**U**
	Ile	Thr	Asn	Ser	**C**
	Ile	Thr	Lys	Arg	**A**
	Met	Thr	Lys	Arg	**G**
G	Val	Ala	Asp	Gly	**U**
	Val	Ala	Asp	Gly	**C**
	Val	Ala	Glu	Gly	**A**
	Val	Ala	Glu	Gly	**G**

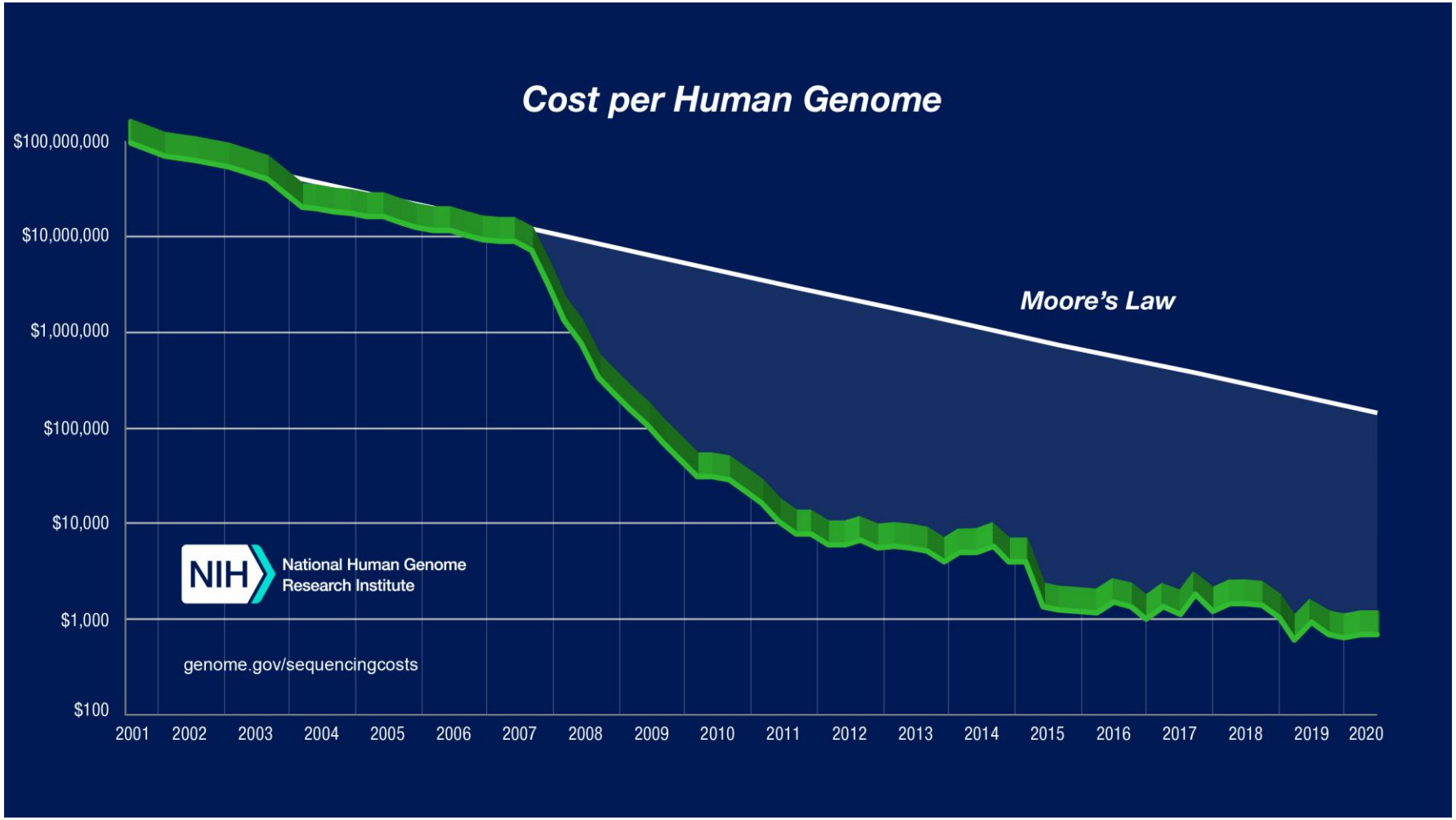
Podatci

Cijena sekvenciranja (do 2015.)



Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) www.genome.gov/sequencingcosts

Cijena sekvenciranja (danas)



Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) www.genome.gov/sequencingcosts

Utrka u sekvenciranju

Hayden, *Is the $1,000 genome for real?* (Nature News & Comment, 2014)

The HiSeq X Ten is composed of ten HiSeq X machines, and sells for at least $10 million.

http://www.illumina.com/systems/hiseq-x-sequencing-system/system.html

Oxford nanopores

FASTA format podataka

- Izgovara se kao “fast A”
- FASTA datoteka se sastoji od nekoliko blokova
- Može postojati i komentar (linije komentara moraju početi s ","); rijetko se koristi
- Svaki blok ima slijedeću strukturu
  - Liniju zaglavlja
  - Jednu ili više linija same sekvence

FASTA primjer

>FASTA blok primjer 1

AGCTAGCT-CATAT

FASTA format podataka

- Linija zaglavlja se razlikuje od linija sekvence time što počinje sa znakom veće od (“>”) u prvom stupcu.
- Riječ koja slijedi znak “>” je identifikator sekvence, a ostatak linije je opis (oboje je opcionalno).
- Ne smije biti razmaka između “>” i prvoga slova identifikatora.
- Preporuča se da sve linije teksta budu kraće od 80 znakova. Sekvenca (slijed) završava krajem datoteke ili sljedećom linijom koja počinje s “>”; ovo pokazuje početak sljedeće sekvence.

FASTA format podataka

Simbol*	Značenje
A	Adenin
C	Citozin
G	Gvanin
T	Timin
N	Adenin ili gvanin ili citozin ili timin
-	Procijep nepoznate duljine

*Postoje još i slova za kombinacije po dva i tri nukleotida, ali se rjeđe koriste.

FASTQ format podataka

- FASTQ format obično koristi 4 linije po sekvenci
  - Linija 1 počinje sa ‘@’ znakom, a nakon nje je identifikator sekvence i opcionalan opis.
  - Linija 2 je sama sekvenca (niz slova).
  - Linija 3 počinje s ‘+’ znakom, a opcionalno iza njega može biti isti identifikator sekvence (i bilo koji opis) ponovo.
  - Linija 4 predstavlja vrijednosti kvalitete za sekvencu iz linije 2; broj znakova mora biti jednak broju slova u sekvenci.

Primjer FASTQ zapisa

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

FASTQ format podataka

- Originalni FASTQ podaci omogućavaju da se sekvenca i niz znakova kvalitete prostiru u nekoliko linija.
- Ovo može zakomplicirati “parsiranje” podataka: “@” i “+” se također mogu nalaziti u nizu znakova kvalitete.

Formati kvalitete

- p – vjerojatnost da je očitana baza netočna
- Sangerov format: $Q_{sanger} = -10 \log_{10} p$
  - može kodirati rezultat kvalitete od 0 do 93 koristeći ASCII znakove 33 do 126

    !"#$%&'()*+,./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ

    [\]^_`abcdefghijklmnopqrstuvwxyz{|}~

  - **!** je oznaka najniže kvalitete, a **~** oznaka najviše

Formati kvalitete - primjer

- Vjerojatnost $p = 0.01$

$$Q_{sanger} = -10 \log_{10} p$$
$$= -10 \log_{10} 0.01$$
$$= 20$$