Computational Genomics

Irit Gat-Viks, Ron Shamir, Roded Sharan

Fall 2018-19

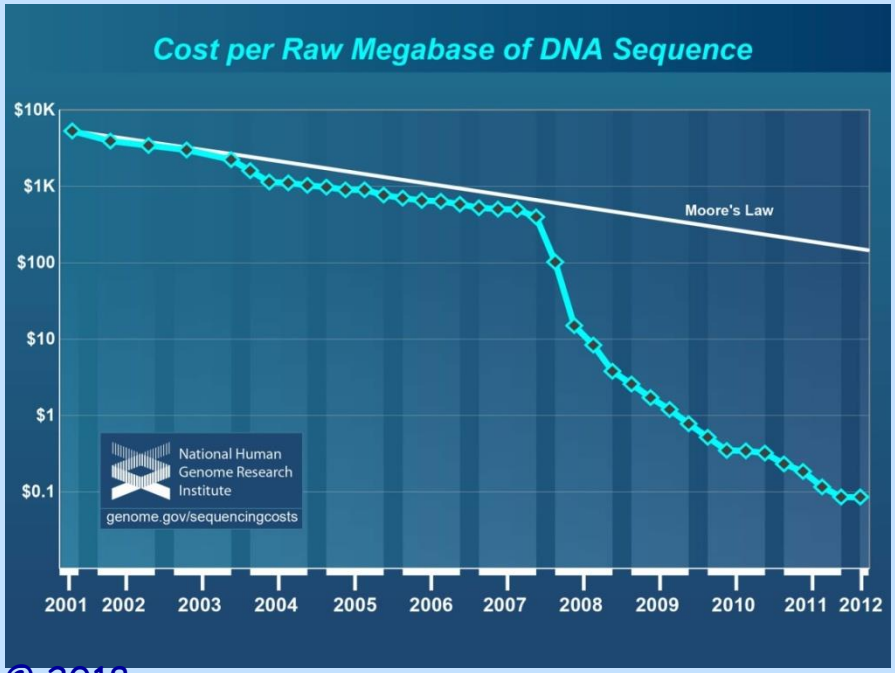
What's in class this week

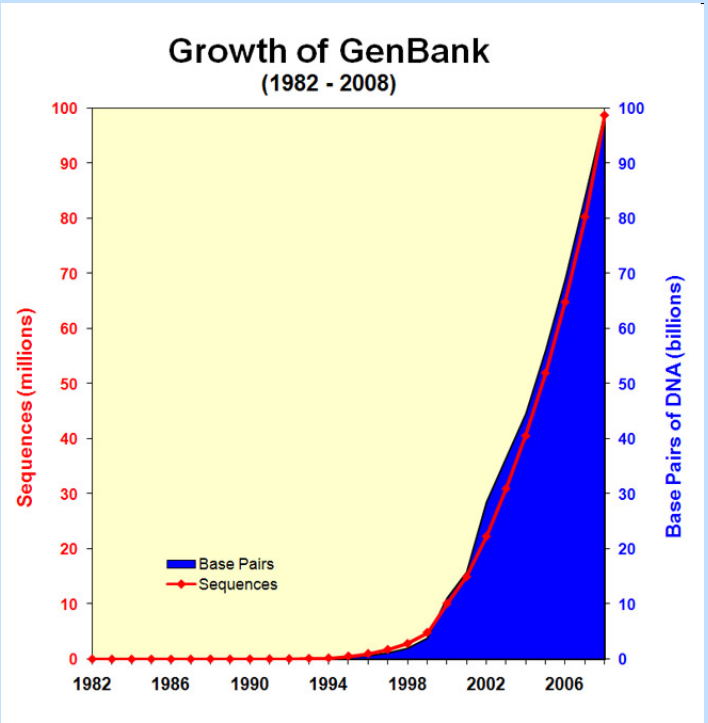
- Motivation
- Administration
- Some very basic biology & biotechnology, with examples of our type of computational problems
- Additional examples

Motivation

Bioinformatics

- The information science of biology: organize, store, analyze and visualize biological data
- Responds to the explosion of biological data, and builds on the IT revolution
- Use computers to analyze A LOT of biological data.





Paradigm shift in biological research

<u>Classical biology:</u> focus on a single gene or sub-system. *Hypothesis driven*

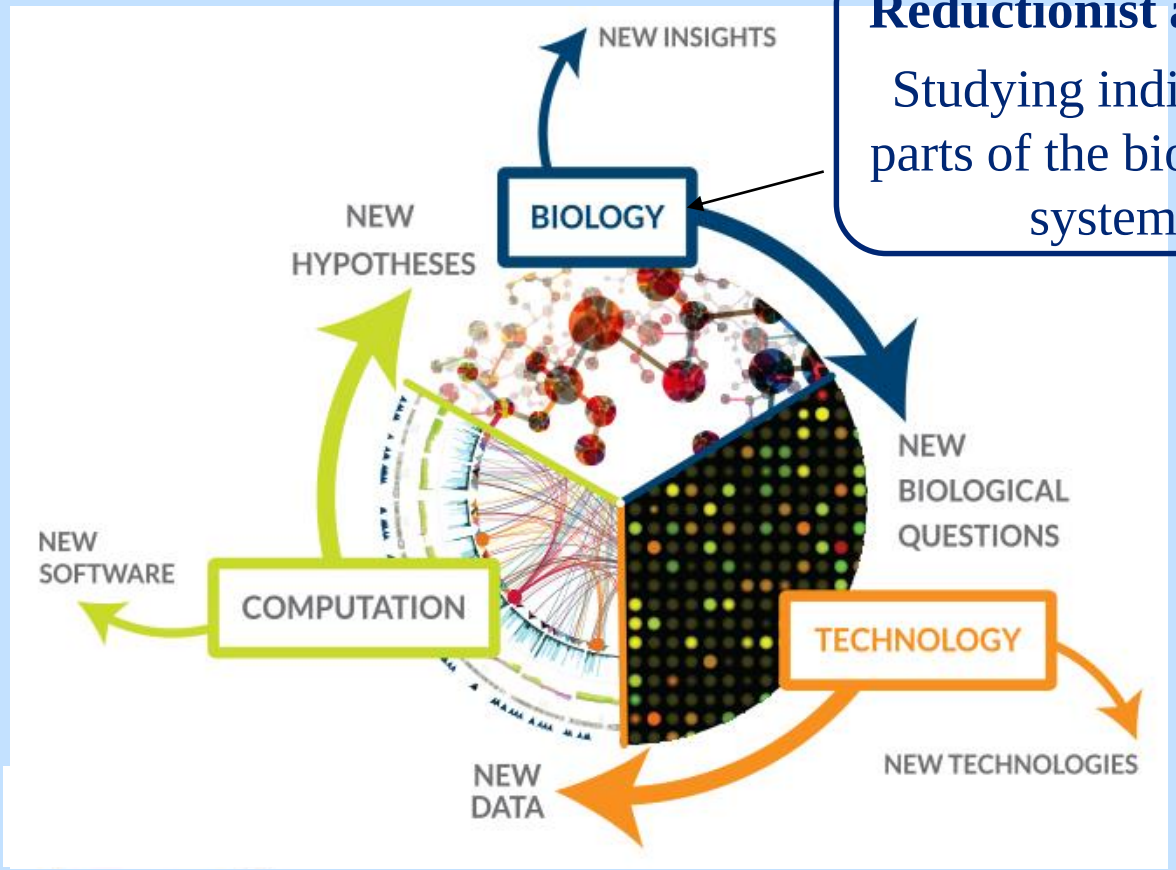
Large-scale data;
Bioinformatics

<u>Systems biology:</u> measure (or model) the behavior of numerous parts of an entire biological system. *Hypothesis generating*

What do bioinformaticians study?

- Bioinformatics today is part of almost every molecular biological research.
- It is also essential to the new era of precision / personalized medicine: using computational methods for improving disease prevention, diagnosis and treatment

Research in systems biology



Reductionist approach

Studying individual parts of the biological system

Systems approach: Unbiased analysis of numerous constituents of the biological system

Terminology

High throughput data נתונים רחבי היקף

Big data נתוני עתק

Bioinformatics tools/algorithms/methods

אלגוריתמים חישוביים בביואינפורמטיקה

ביולוגיה חישובית = ביואינפורמטיקה

The Bioinformatics Actors

- Biotechnology companies
- Academic biotechnology research

High throughput data
Bioinformatics tools

- Academic bioinfo research
- Medical informatics startups

- Big <u>Pharmas</u> and Big <u>Agri Biotechs</u>
- National and international <u>research centers</u>

Personalized medicine

The largest DNA ancestry service in the world

sign in | register kit | 0

23andMe

welcome | ancestry | how it works | buy | search | help

23andMe provides ancestry-related genetic reports and uninterpreted raw genetic data. We no longer offer our health-related genetic reports. If you are a current customer please go to the health page for more information. Close alert.

Find out what your DNA says about you and your family.

- Learn what percent of your DNA is from populations around the world
- Contact your DNA relatives across continents or across the street
- Build your family tree and enhance your experience with relatives

order now $99

Watch Greta and Stacy's Story.

Find out how these two women discovered their connection as sisters.

Course Administration

Administration

- ~5 home assignments as part of a home exam, to be done independently (40% of grade)
- Final exam (60%)
- Must pass the Final to pass the course (TAU rules)

- Classes: Tue **12:15-13:30**; Thu **14:30-15:45**
- TA: Nimrod Rappoport (Thu 16-17).

Administration (cont.)

- Web page of the course: http://www.cs.tau.ac.il/~rshamir/cg/18/

- Includes slides and full lecture scribes of previous years on each of the classes.

- Revised slide presentations will be posted in the website prior to each class

- Utilize these resources - Avoid taking notes in class!

Bibliography

- No single textbook covers the course :-(
- See the full bibliography list in the website (also for basic biology)
- Key sources:
  - Gusfield: Algorithms for strings, trees and sequences
  - Durbin et al.: Biological sequence analysis
  - Pevzner: Computational molecular biology
  - Pevzner and Shamir (eds.): Bioinformatics for Biologists
  - Pevzner and Compeau: Bioinformatics Algorithms: an active learning approach (also a Coursera MOOC)

Introduction

Introduction

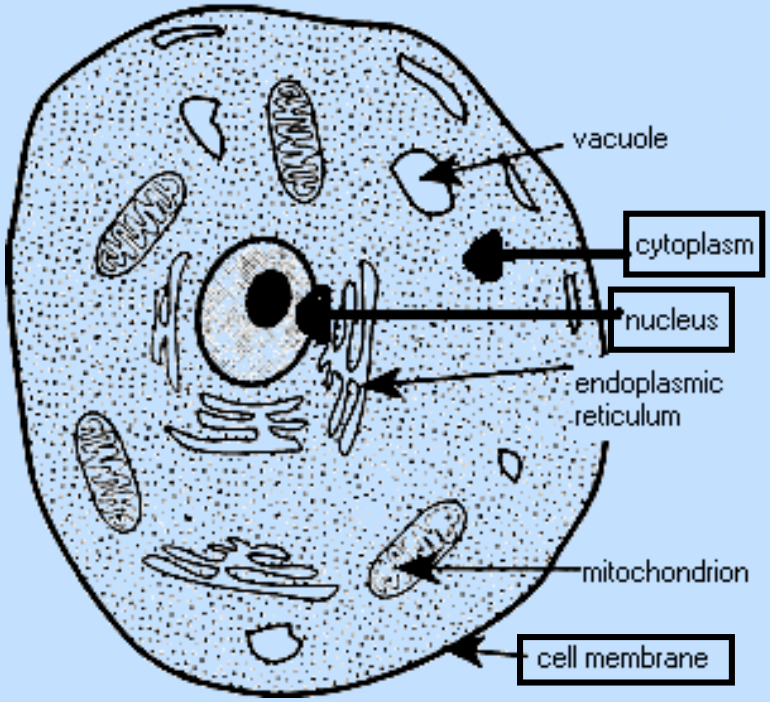
1. Basic biology
2. Basic biotechnology

+ some computational challenges arising along the way

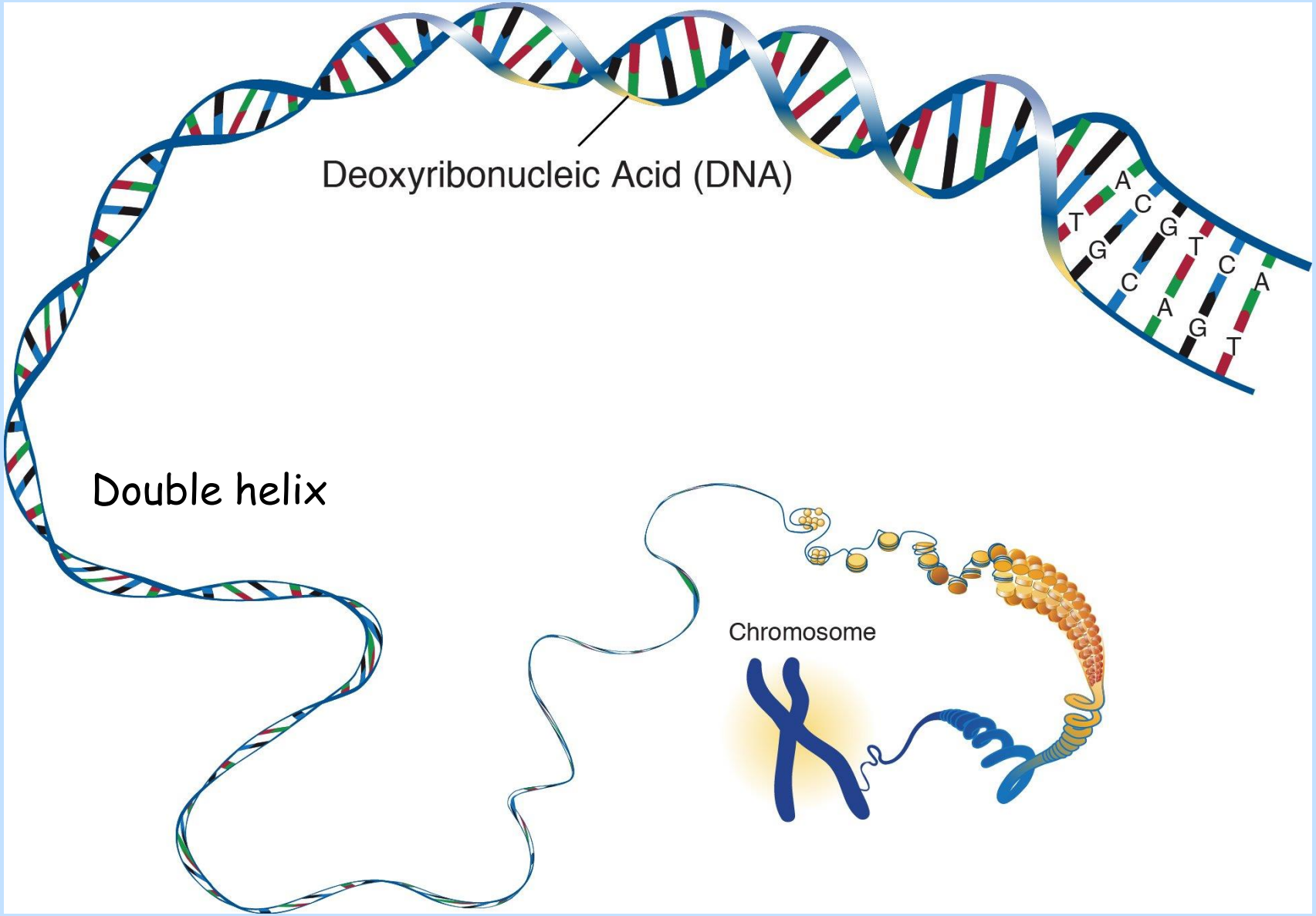
- Touches on Chapters 1-8 in "The Cell" by Alberts et al.

The Cell

- Basic unit of life.
- Carries complete characteristics of the species.
- All cells store hereditary information in DNA.
- All cells transform DNA to *proteins*, which determine cell's structure and function.
- Two classes: *eukaryotes* (with nucleus) and *prokaryotes* (without).



http://regentsprep.org/Regents/biology/units/organization/cell.gif

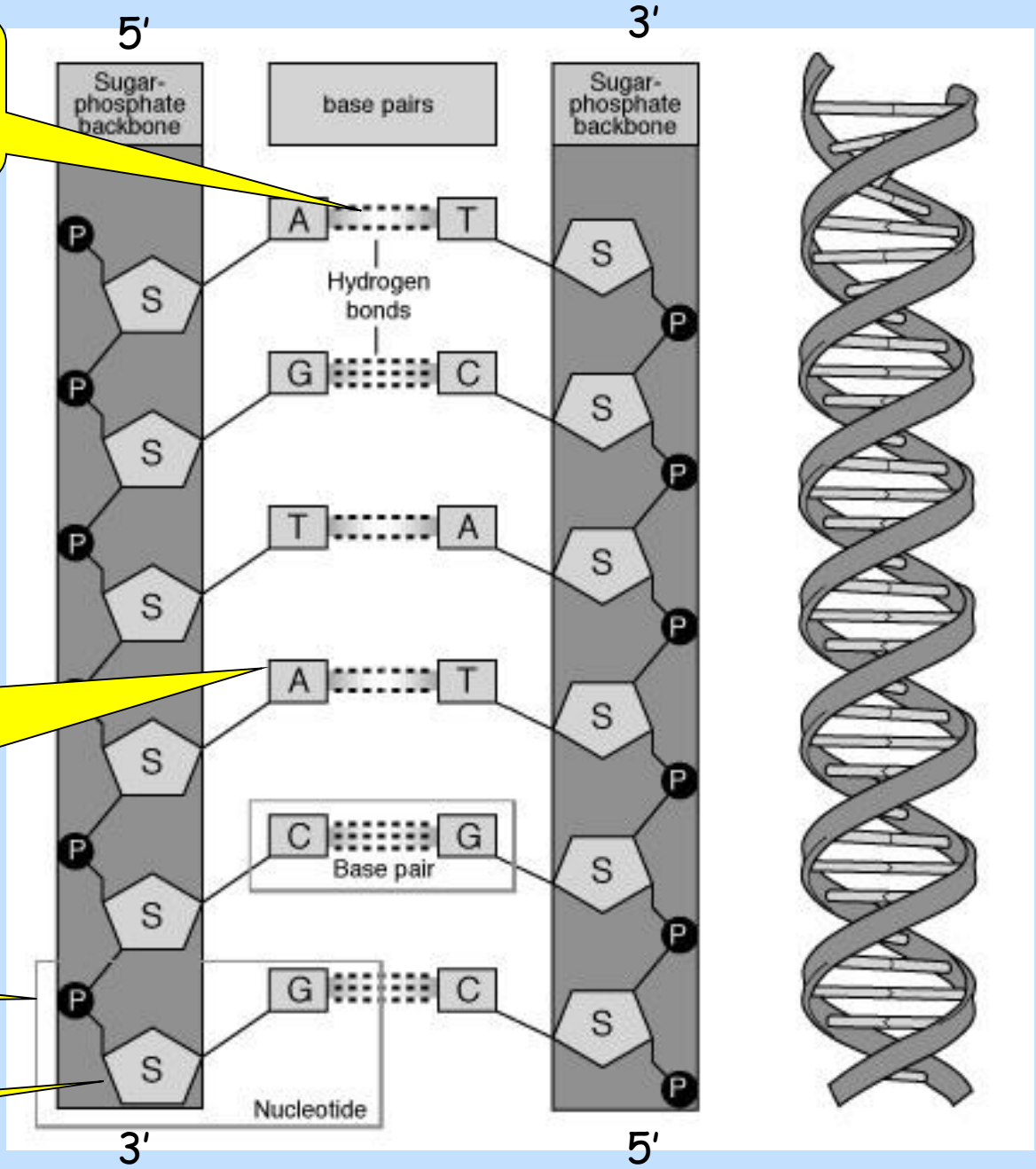

Deoxyribonucleic Acid (DNA)
A
C
T
G
G
T
C
C
A
A
G
T
Double helix
Chromosome

Weak hydrogen bonds between base pairs

Nucleotides/ Bases:
Adenine (A),
Guanine (G),
Cytosine (C),
Thymine (T).

phosphate

sugar



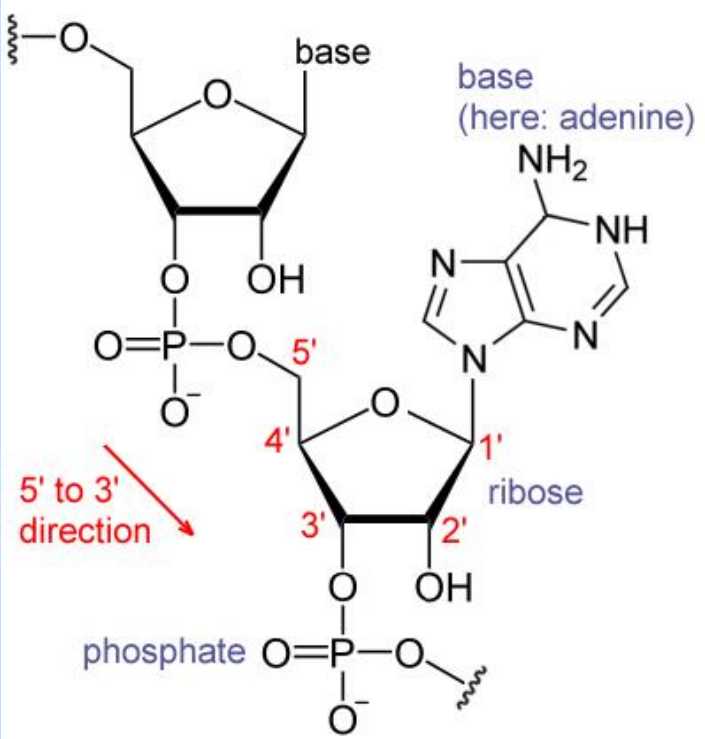
DNA (Deoxy-Ribonucleic acid)

- <u>Bases</u>:
  - Adenine (A)
  - Guanine (G)
  - Cytosine (C)
  - Thymine (T)

Purines

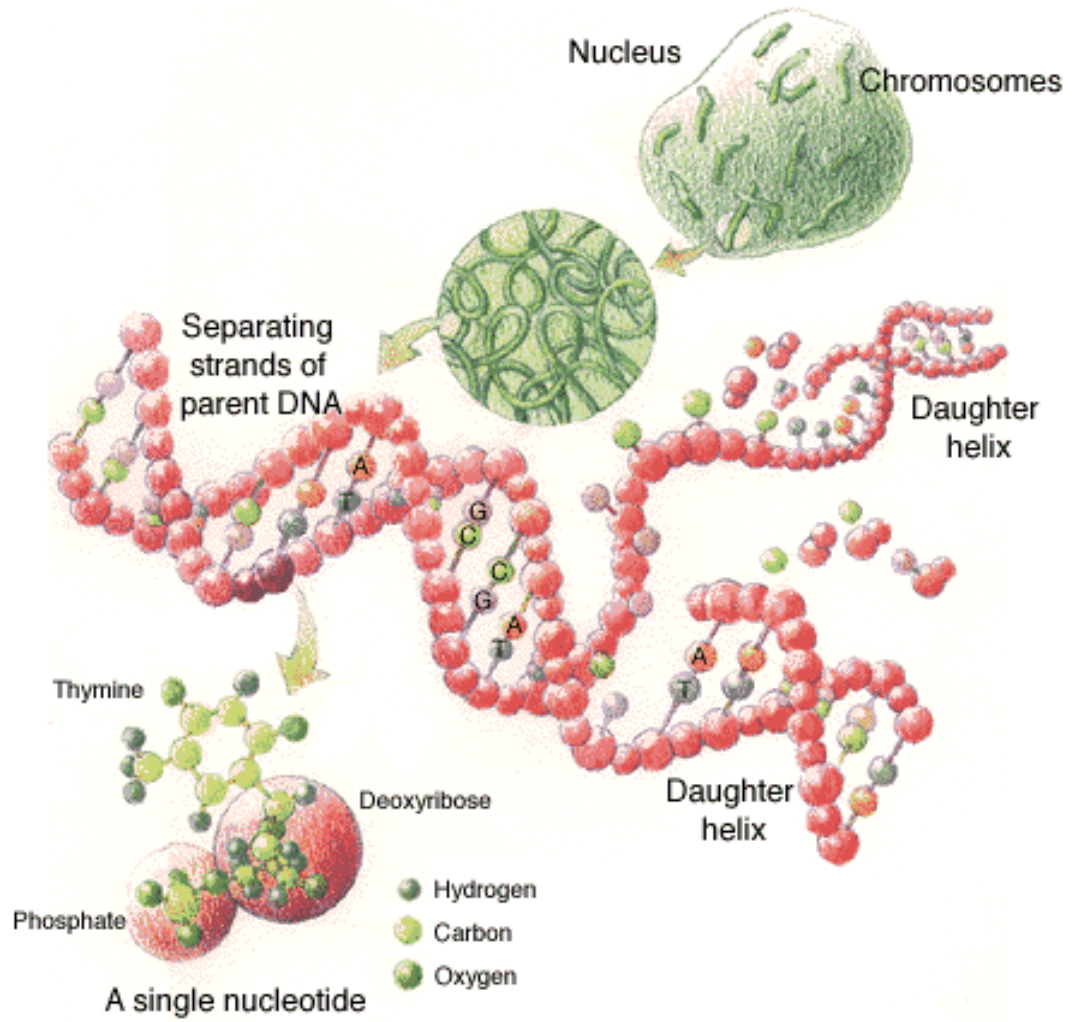
pyrimidines

- <u>Bonds:</u>
  - G - C
  - A - T
- Oriented from 5′ to 3′.
- Located in the cell nucleus

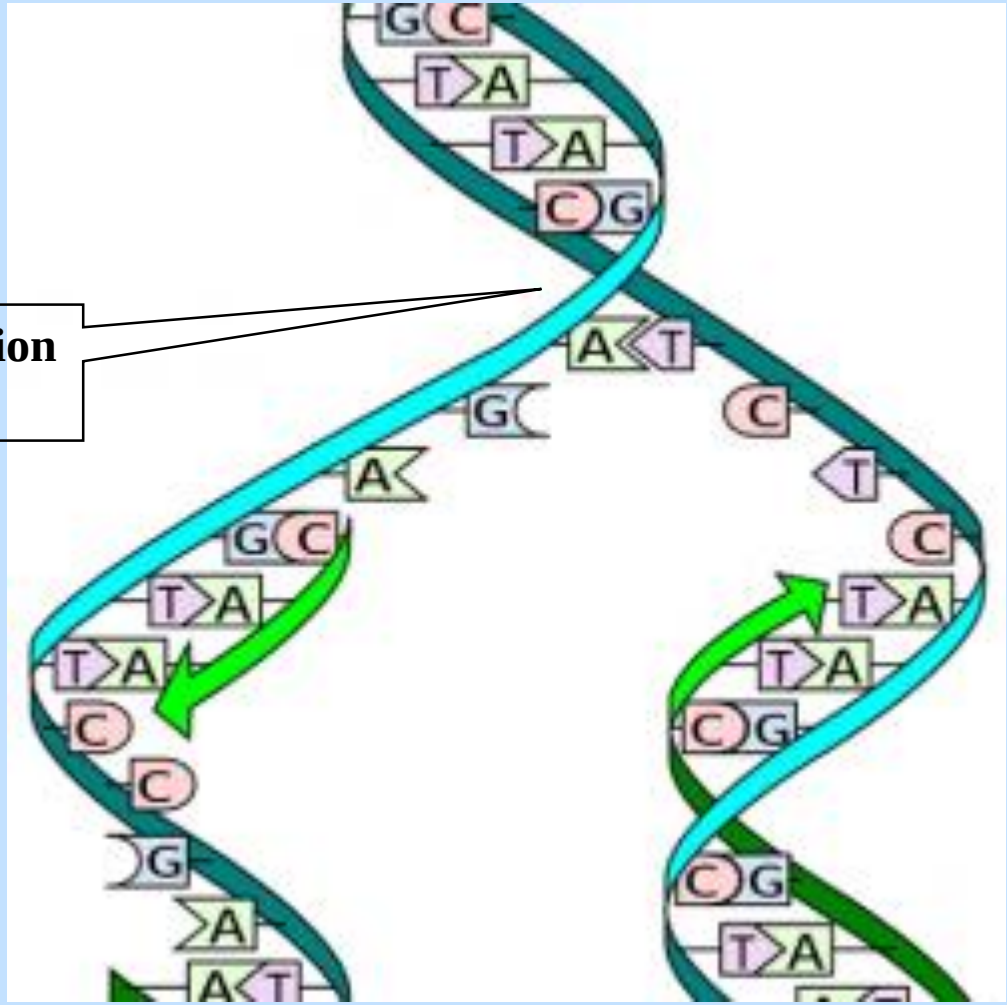


DNA and Chromosomes

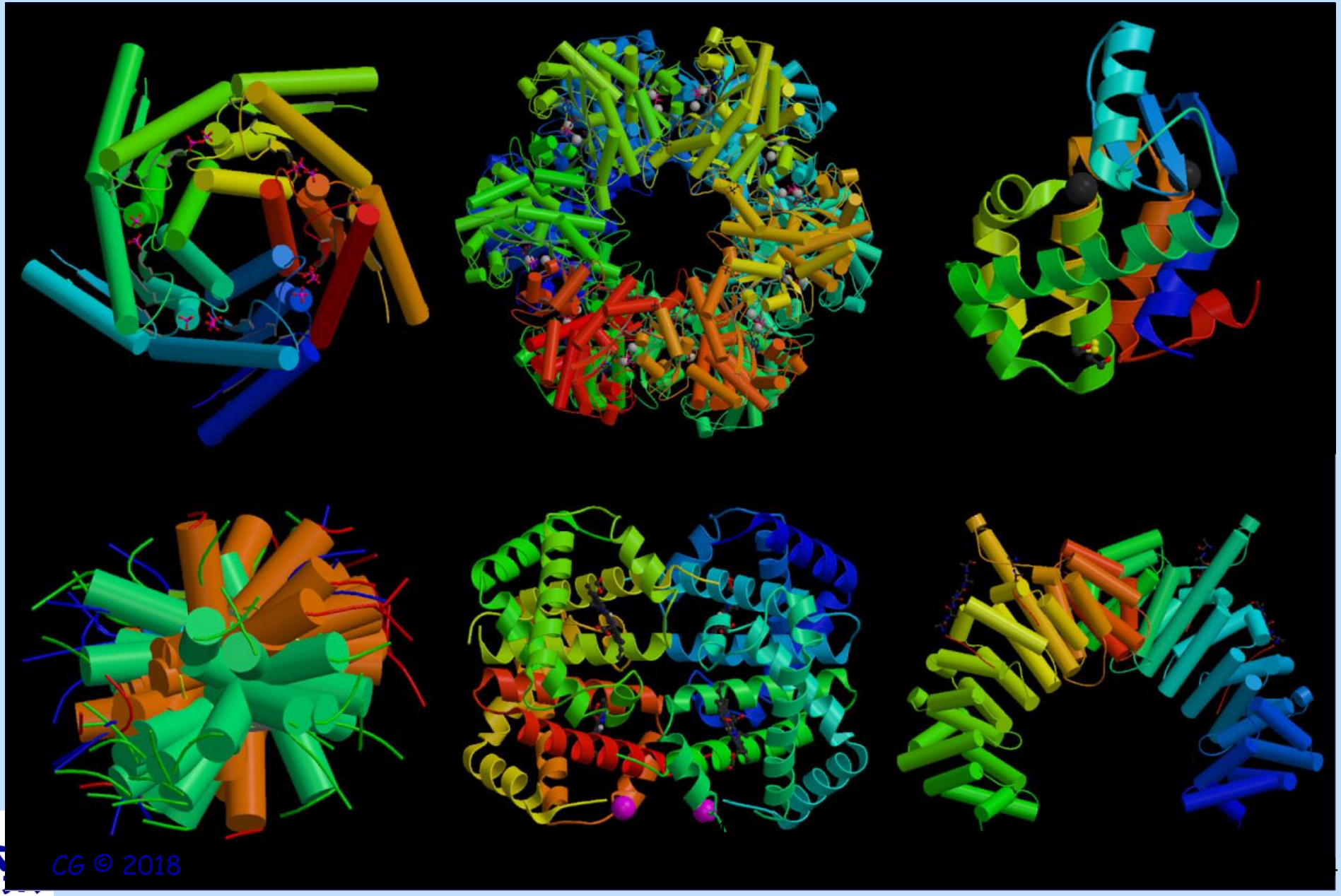
- DNA is packaged
Chromatin: complex of DNA and proteins that pack it (histones)
- *Chromosome*: contiguous stretch of DNA
- *Genome*: totality of DNA material
- *Diploid genome*: two homologous chromosomes, one from each parent



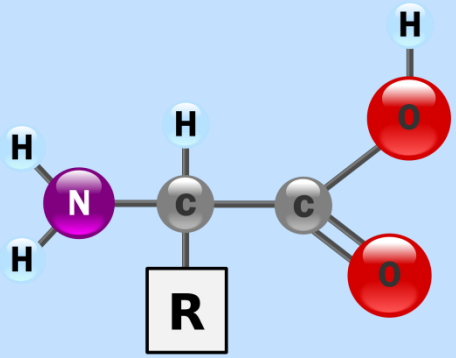
Replication



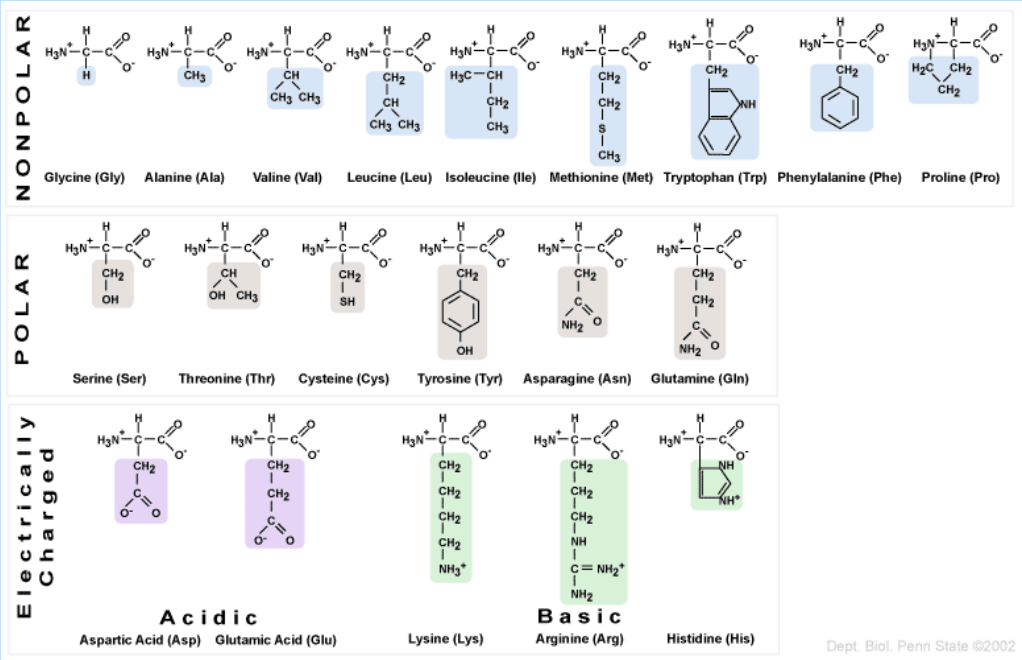
Proteins: The Cellular Machines

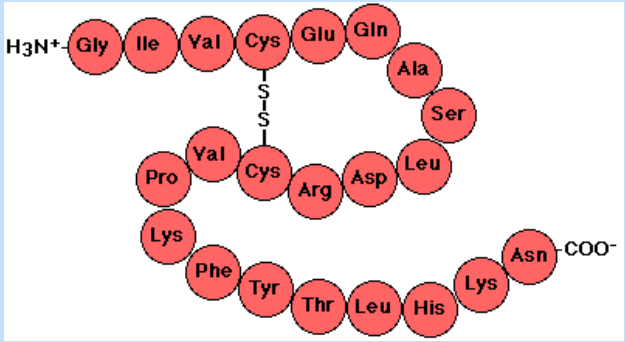


Proteins

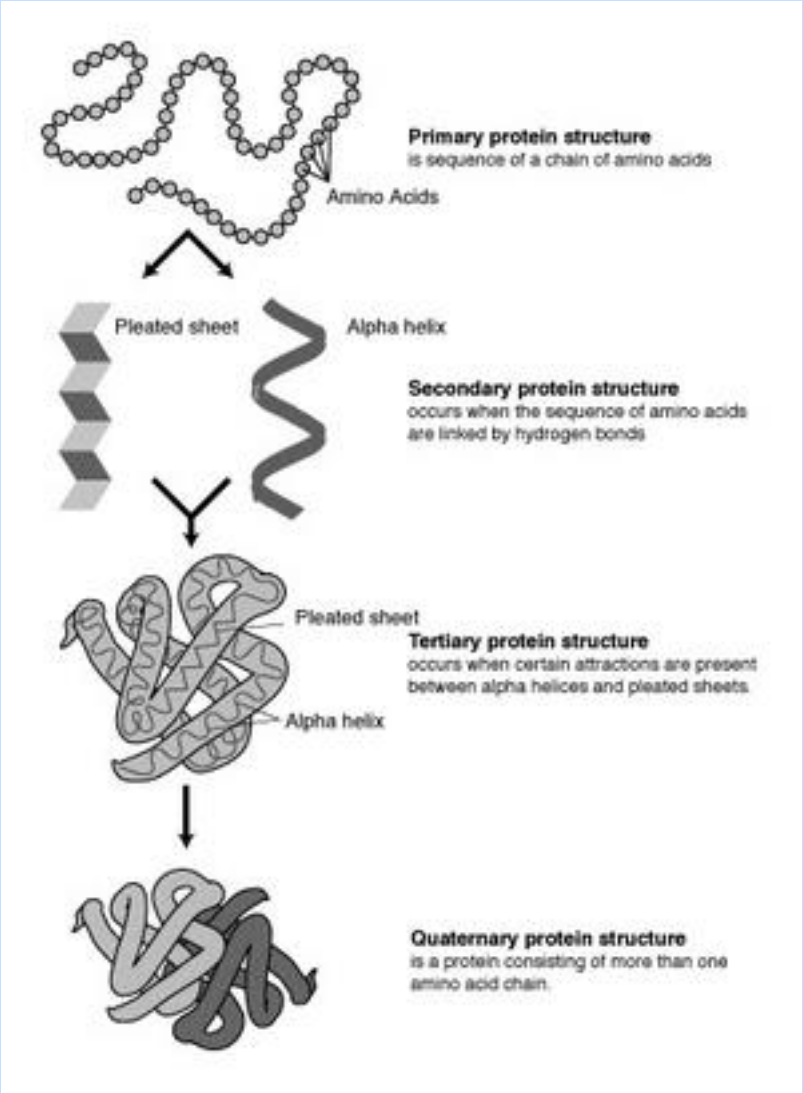


- Build the cell and drive most of its functions.
- Polymers of **amino-acids** (20 types), linked by peptide bonds.
- Oriented (from amino to carboxyl group).
- Fold into 3D structure of lowest energy.



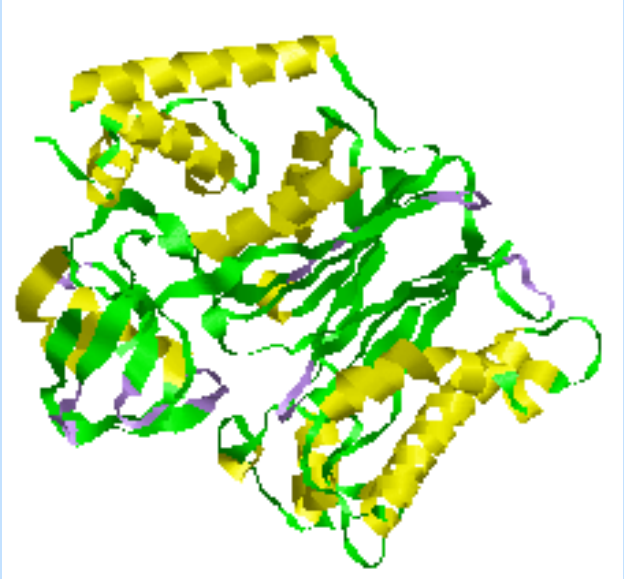


Protein structure



The Protein Folding Problem

PQITLWQRPLVTIKVGGQLKEALLDTGADDTVLKDIELPGRWKPKIIGGIGGFVKVR



The Protein Folding Problem

- Given a sequence of amino acids, predict the 3D structure of the protein.
- Motivation: functionality of protein is determined by its 3D structure.
- Solution Approaches:
  - de novo / ab initio (=from scratch): extremely hard
  - Homology
  - Threading

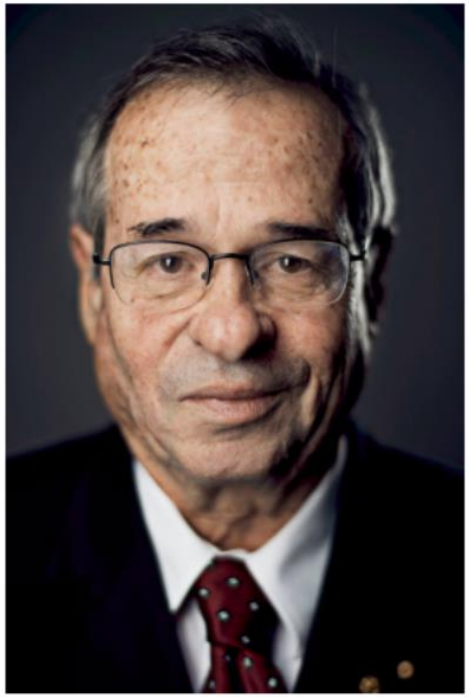
The Nobel Prize in Chemistry 2013

© Nobel Media AB. Photo: A. Mahmoud

Martin Karplus

© Nobel Media AB. Photo: A. Mahmoud

Michael Levitt



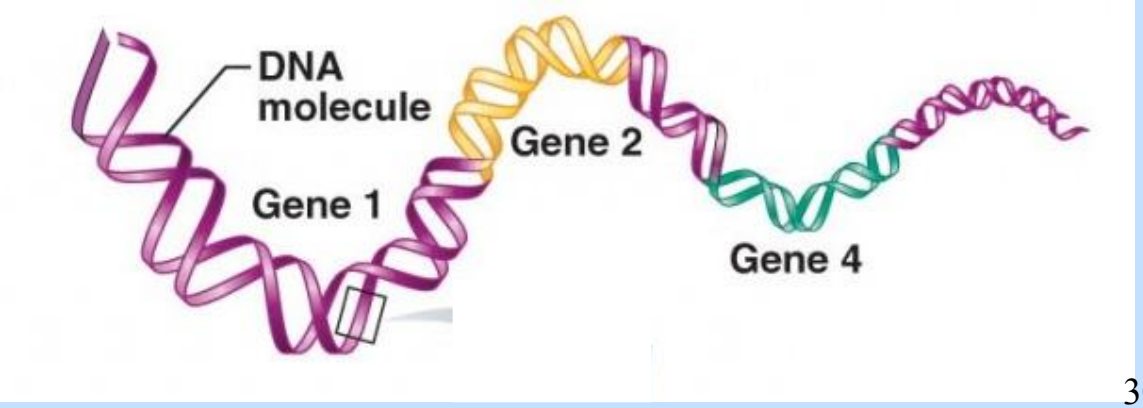
© Nobel Media AB. Photo: A. Mahmoud

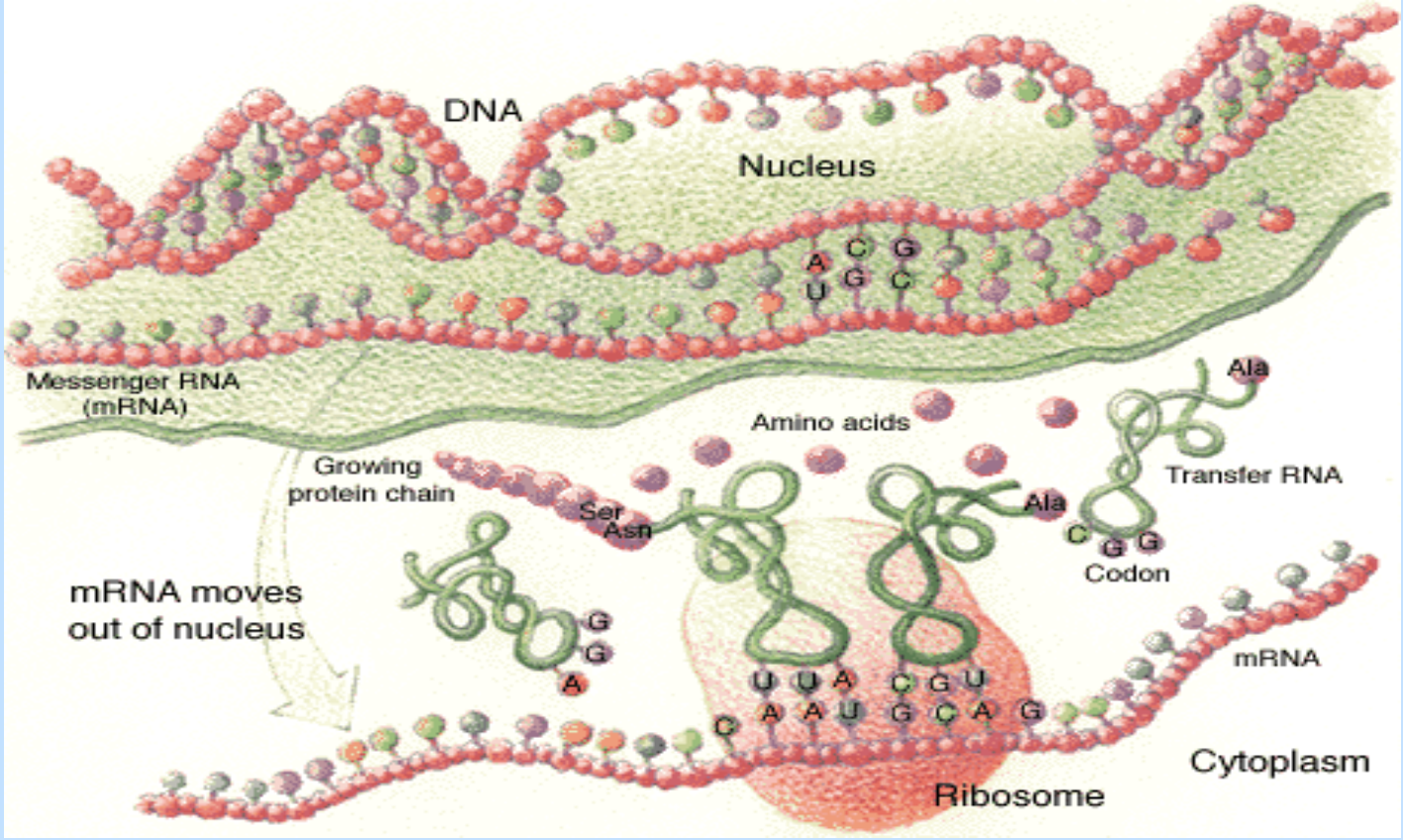
Arieh Warshel

"for the development of multiscale models for complex chemical systems."

Genes

- Gene: a segment of DNA that specifies a protein.
- Genes are < 3% of human DNA
- The rest - non-coding (used to be called "junk DNA")
  - RNA elements
  - Regulatory regions
  - Retrotransposons
  - Pseudogenes
  - and more…





http://www.ornl.gov/hgmis/publicat/tko/index.htm

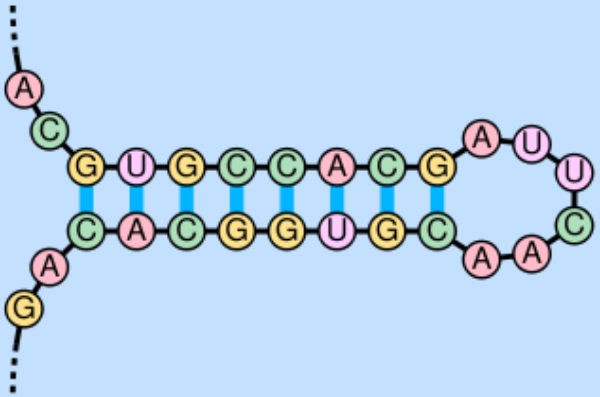
DNA → RNA → protein

The hard disk ↑ **One program** ↑ **Its output**

transcription translation

RNA (Ribonucleic acid)

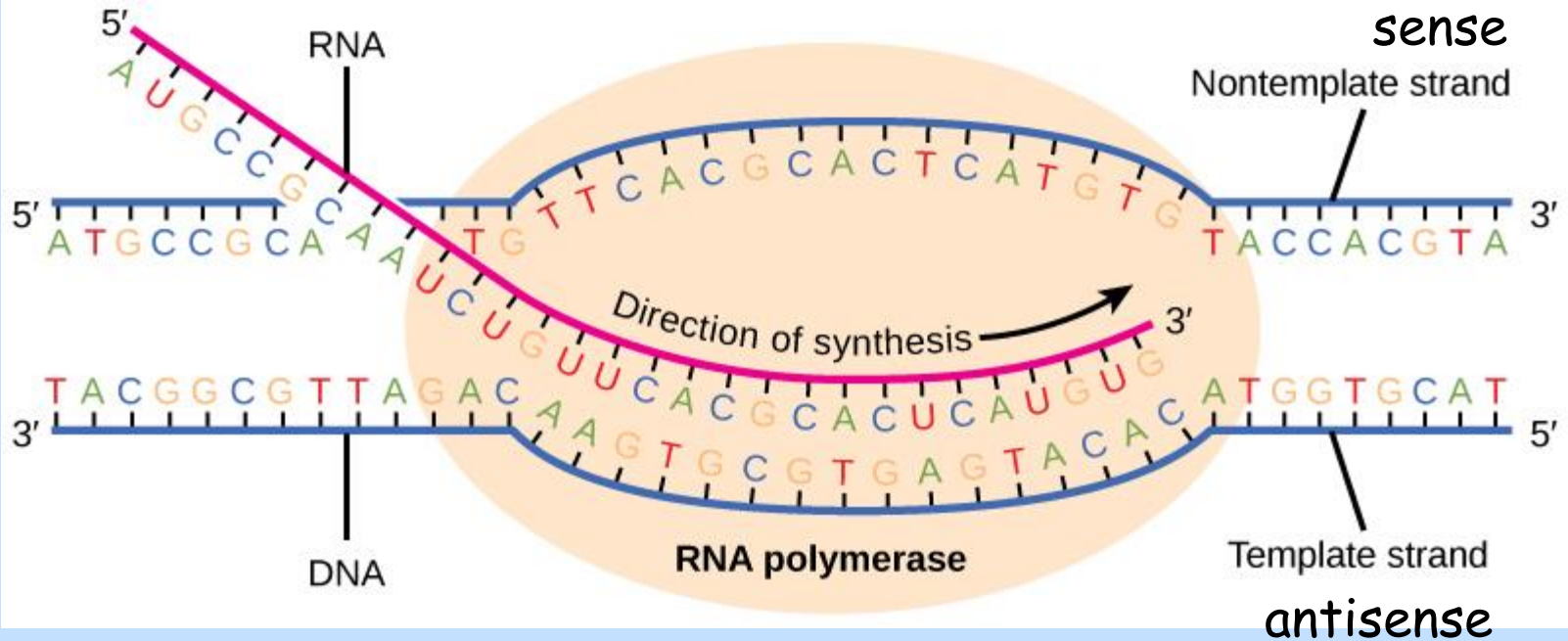
- <u>Bases</u>:
  - Adenine (A)
  - Guanine (G)
  - Cytosine (C)
  - Uracil (U); replaces T
- Oriented from 5′ to 3′.
- Single-stranded => flexible backbone => secondary structure => catalytic role.



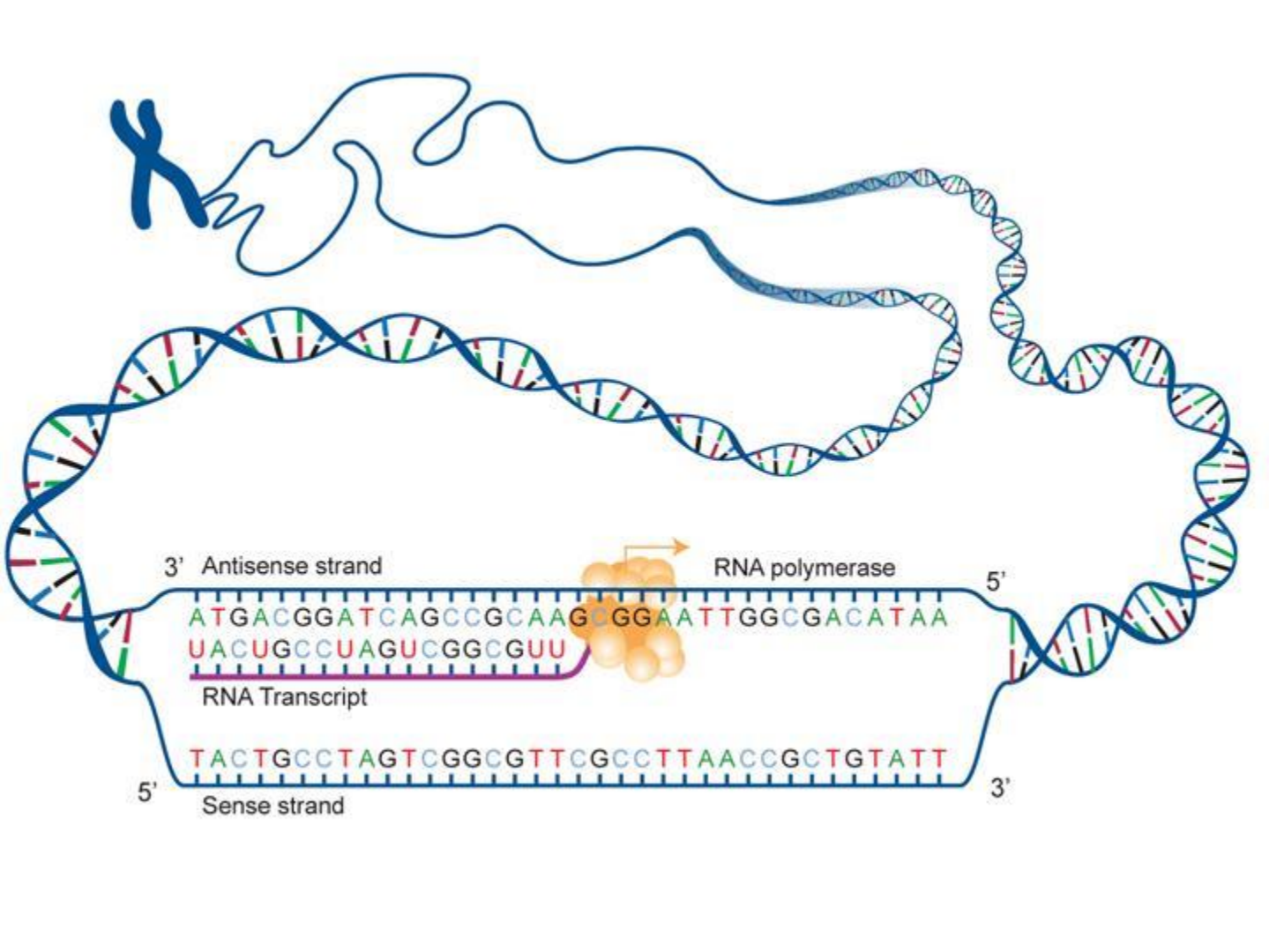
Transcription of DNA into RNA

Complementarity:

A-U; C-G



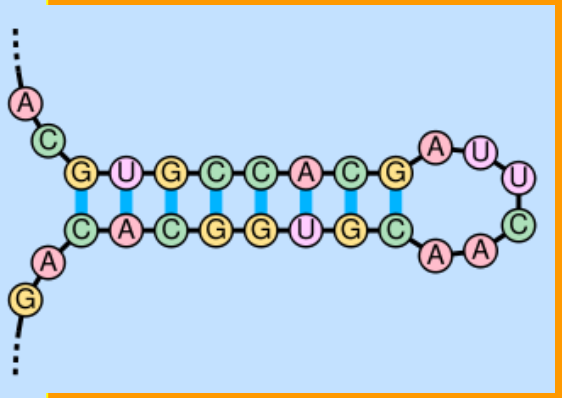
Transcription of DNA into RNA

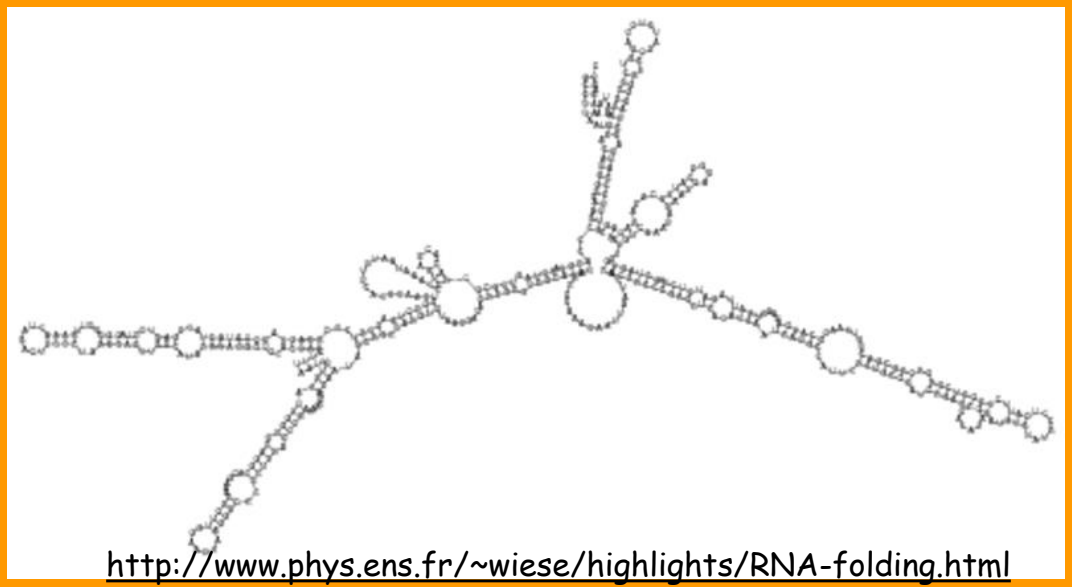


The RNA Folding Problem

Given an RNA sequence, predict its folding = the one that creates a maximum number of matched pairs
Motivation: RNA function is determined by its 2D structure.

GCCUUAAUGCACAUGGGCAAGCCCACGUAGCUAGUCGCGCGACACCAGUCCCAAAUAUGUUCACCCAACUCGCCUGACCGUCCCGCA
GUAGCUAUACUACCGACUCCUACGCGGUUGAAACUAGACUUUUCUAGCGAGCUGUCAUAGGUAUGGUGCACUGUCUUUAAUUUUGU
AUUGGGCCAGGCACGAAAGGCUUGGAAGUAAGGCCCCGCUUGACCCGAGAGGUGACAAUAGCGGCCAGGUGUAACGAUACGCGGGU
GGCACGUACCCCAAACAAUUAAUCACACUGCCCGGGCUCACAUUAAUCAUGCCAUUCGUUGCCGAUCCGACCCAUAAGGAUGUGUA
UGCCUCAUUCCCGGUCGGGGCGGCGACUGUUAACGCAUGAGAACUGAUUAGAUCUCGUGGUAGUGCUUGUCAAAUAGAAUGAGGCC
AUUCCACAGACAUAGCGUUUCCCAUGAGCUAGGGGUCCCAUGUCCAGGUCCCCUAAAUAAAAGAGUCUCAC





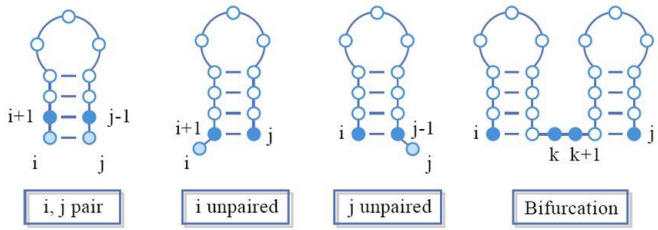
http://www.phys.ens.fr/~wiese/highlights/RNA-folding.html

CONFERENCE ON INTELLIG

JOIN ISCB

The Nussinov algorithm

- Given sequence $X = x_1 \ldots x_N$,
- Define DP matrix: $F(i, j)$ = maximum number of base-pairs if $x_i \ldots x_j$ folds optimally
  - Matrix is symmetric, so let $i < j$



ISMB 2018

Chicago, United States

July 6 - July 10

GY

KEY DATES

ISCB Accomplishments by a Senior Scientist Award Keynote

Ruth Nussinov

Senior Principal Investigator, National Cancer Institute, National Institutes of Health, United States;
Professor, School of Medicine, Department of Human Genetics, Tel Aviv University, Israel

Presentation Title: A woman's computational biology journey
Time: Tuesday July 10, 5:00 pm - 6:00 pm
Room: Grand Ballroom C-F

The ISCB Accomplishments by a Senior Scientist Award recognizes leaders in the fields of computational biology and

The Genetic Code

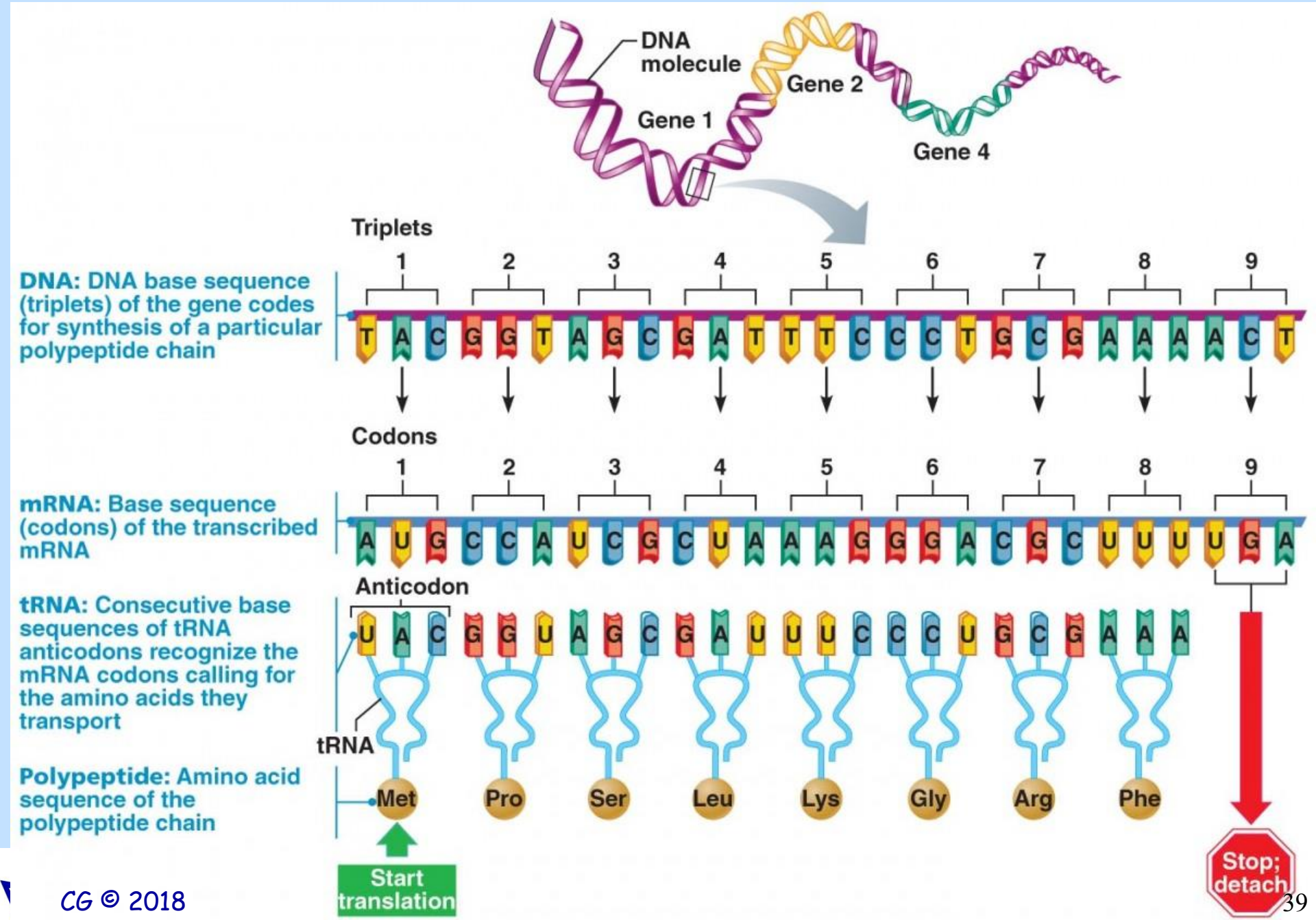
- Codon - a triplet of bases, codes a specific amino acid (except the stop codons)
- Stop codons - signal termination of the protein synthesis process
- Different codons may code the same amino acid

Second base of codon

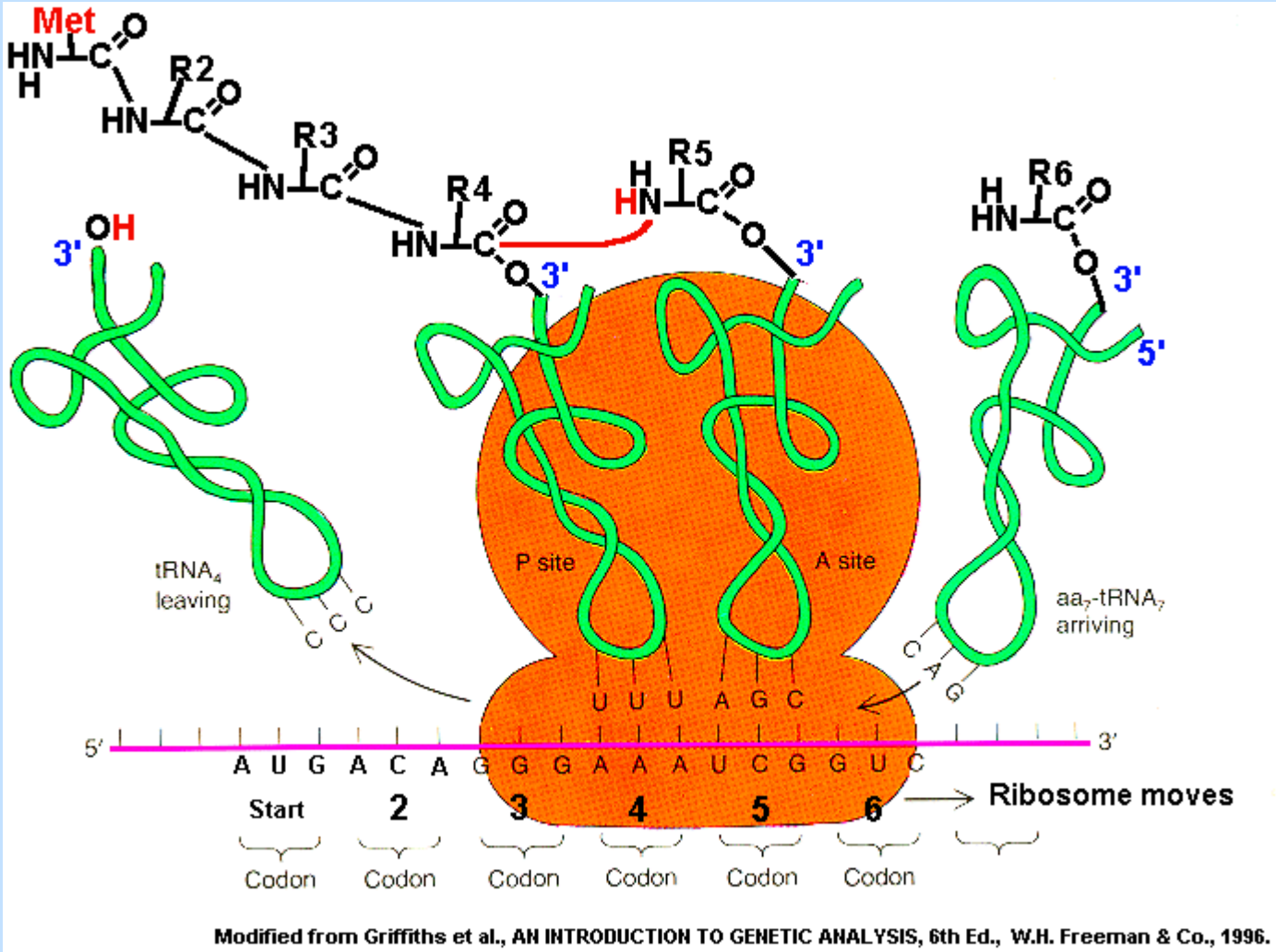
First base of codon	U	C	A	G	Third base of codon
U	UUU, UUC Phe; UUA, UUG Leu	UCU, UCC, UCA, UCG SER	UAU, UAC Tyr; UAA, UAG	UGU, UGC Cys; UGA; UGG Trp	U, C, A, G
C	CUU, CUC, CUA, CUG Leu	CCU, CCC, CCA, CCG Pro	CAU, CAC His; CAA, CAG Gln	CGU, CGC, CGA, CGG Arg	U, C, A, G
A	AUU, AUC, AUA Ile; AUG Met	ACU, ACC, ACA, ACG Thy	AAU, AAC Asn; AAA, AAG Lys	AGU, AGC Ser; AGA, AGG Arg	U, C, A, G
G	GUU, GUC, GUA, GUG Val	GCU, GCC, GCA, GCG Ala	GAU, GAC Asp; GAA, CAG Glu	GGU, GGC, GGA, GGG Gly	U, C, A, G

The genetic code, written by convention in the form in which the Codons appear in mRNA. The three terminator codons, UAA, UAG, and UGA, are boxed in red; the AUG initiator codon is shown in green.

http://ntri.tamuk.edu/cell/ribosomes.html



Translation



Modified from Griffiths et al., AN INTRODUCTION TO GENETIC ANALYSIS, 6th Ed., W.H. Freeman & Co., 1996.

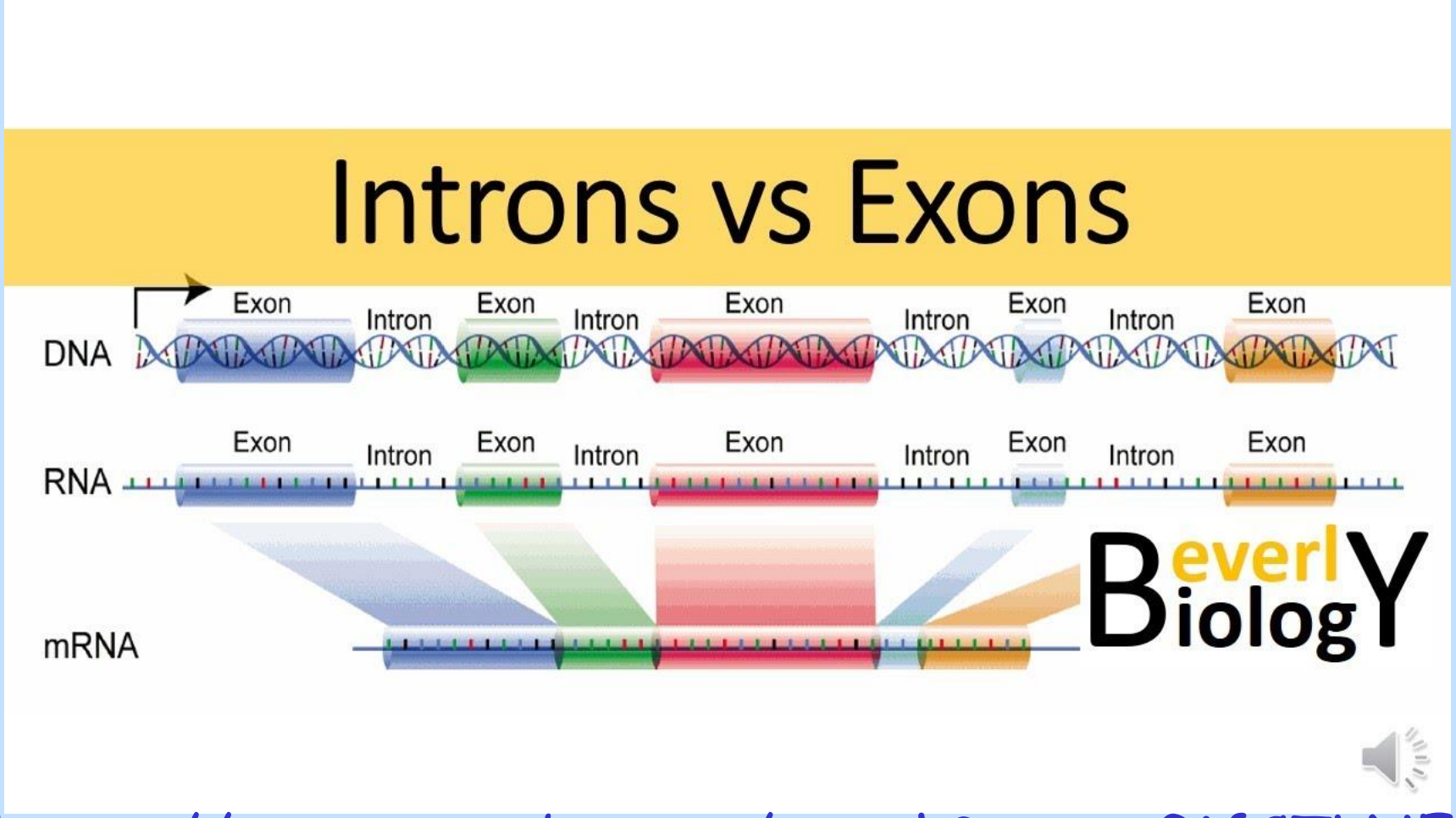
http://biology.kenyon.edu/courses/biol114/Chap05/Chapter05.html#Protein

The Gene Finding Problem

Given a DNA sequence, predict the location of genes (open reading frames) exons and introns.

- A simple solution: seeking stop codons.
- 6 ways of interpreting DNA sequence
- In most cases of eukaryotic DNA, a segment encodes only one gene.
- Difficulty in Eukaryotic DNA: introns & exons

Gene Structure



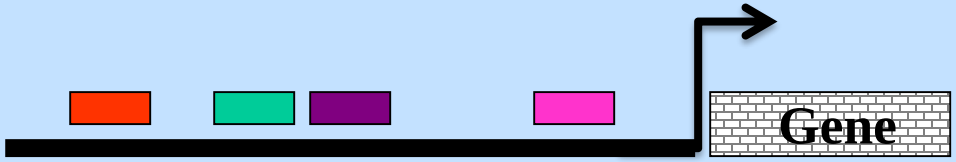
https://www.youtube.com/watch?v=_asGjfCTLNE

Expression and Regulation

DNA → RNA → Protein

transcription translation

Transcription factors (TFs) : proteins that control transcription by binding to specific DNA sequence motifs in the gene's promoter.

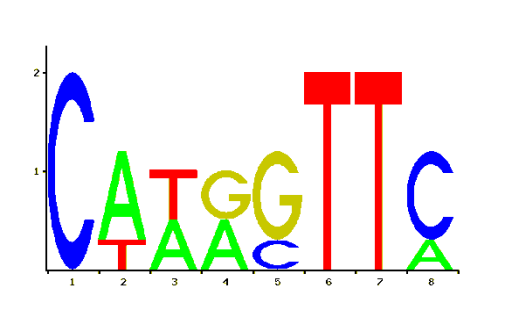


promoter

The Motif Discovery Problem

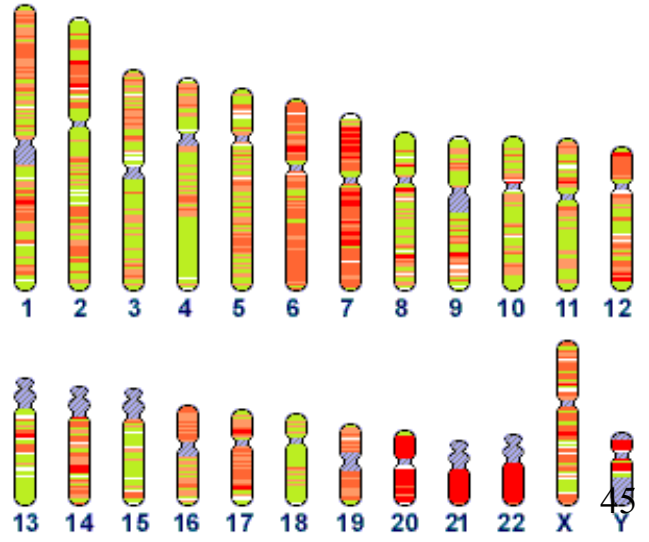
Given a set of DNA sequences that are expected to be co-regulated, find the TF binding motif(s) that are regulating these genes

- Short motifs, probabilitstic
- Long promoters
- Needle in a haystack!



The Human Genome: numbers

- 23 pairs of chromosomes
- ~3,000,000,000 bases
- ~20,000 genes
- Gene length: 1000-3000 bases, spanning 30-40K bases

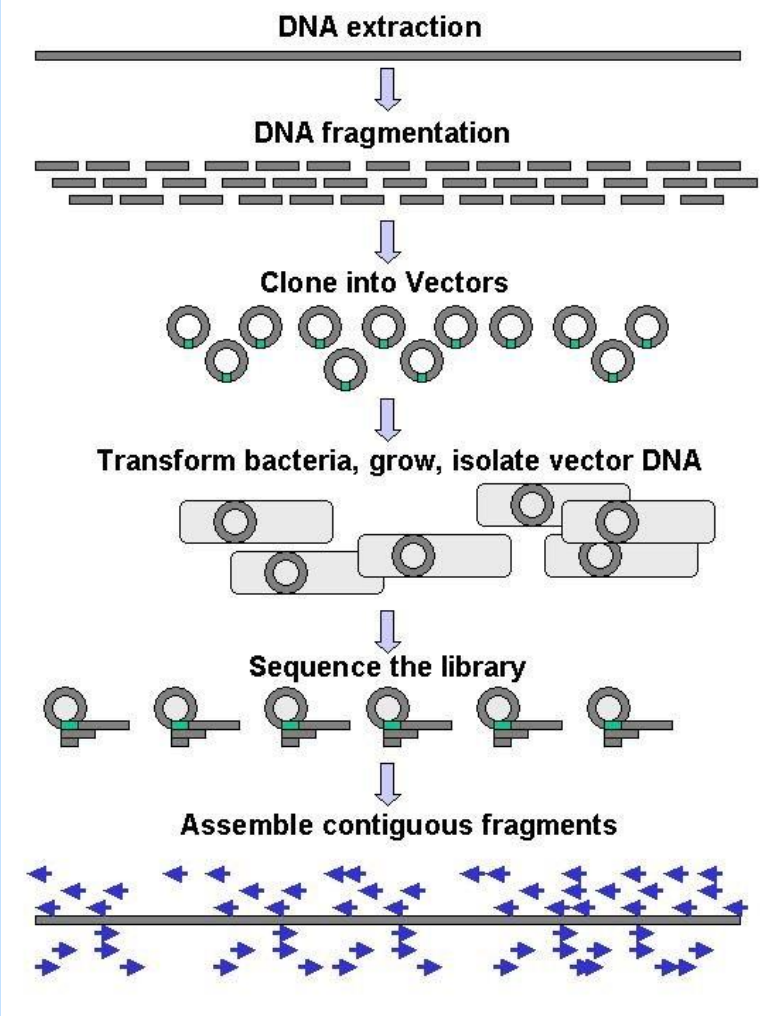


Sequencing the human genome

1990	2000	2006
Project initiation	First draft	"Full sequence"

The Sequence Assembly Problem

- Given a set of sequences, find the shortest (super)string containing all of them.

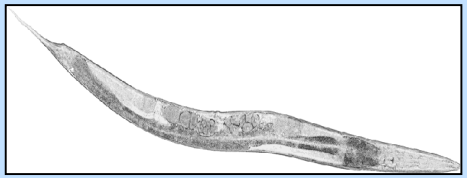


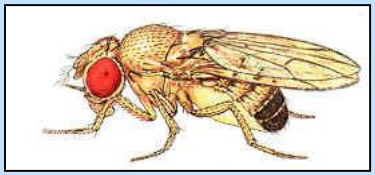
http://www.ornl.gov/hgmis/graphics/slides/images1.html

Now that we have the human genome sequence, what are Computational problems?

Model Organisms

- Eukaryotes; increasing complexity
- Easy to grow, manipulate.





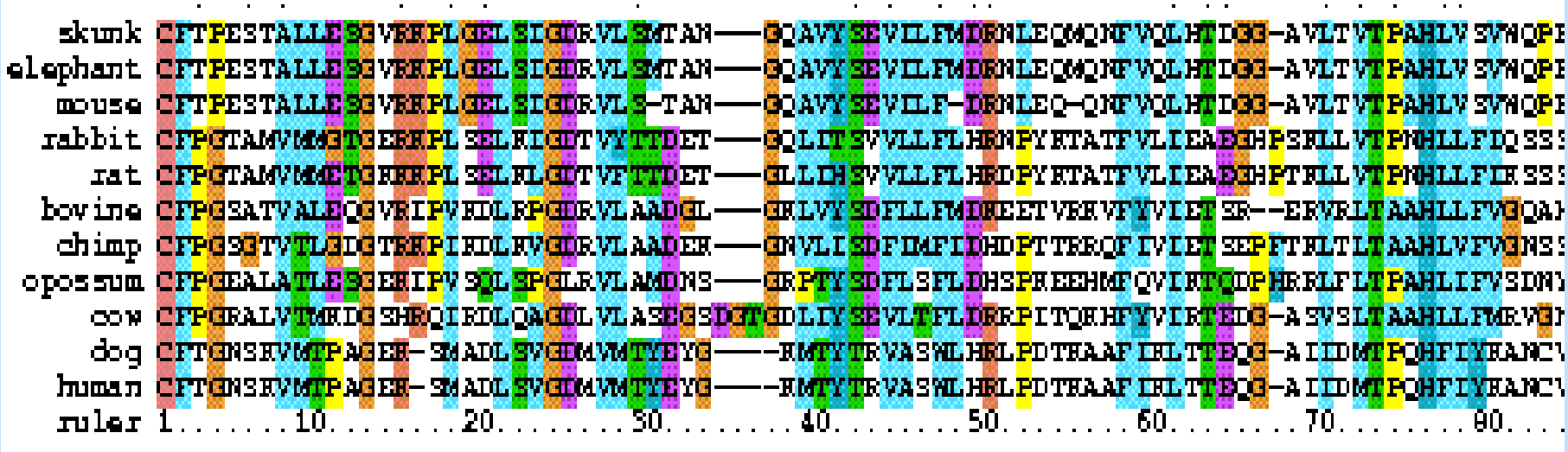
Budding yeast
- 1 cell
- 6K genes

Nematode worm
- 959 cells
- 19K genes

Fruit fly
- vertebrate-like
- 14K genes

mouse
- mammal
- 30K genes

Compare proteins with similar sequences and understand what the similarities and differences mean.



The Rosetta stone

Writing: Ancient Egyptian hieroglyphs, Demotic script, and Greek script

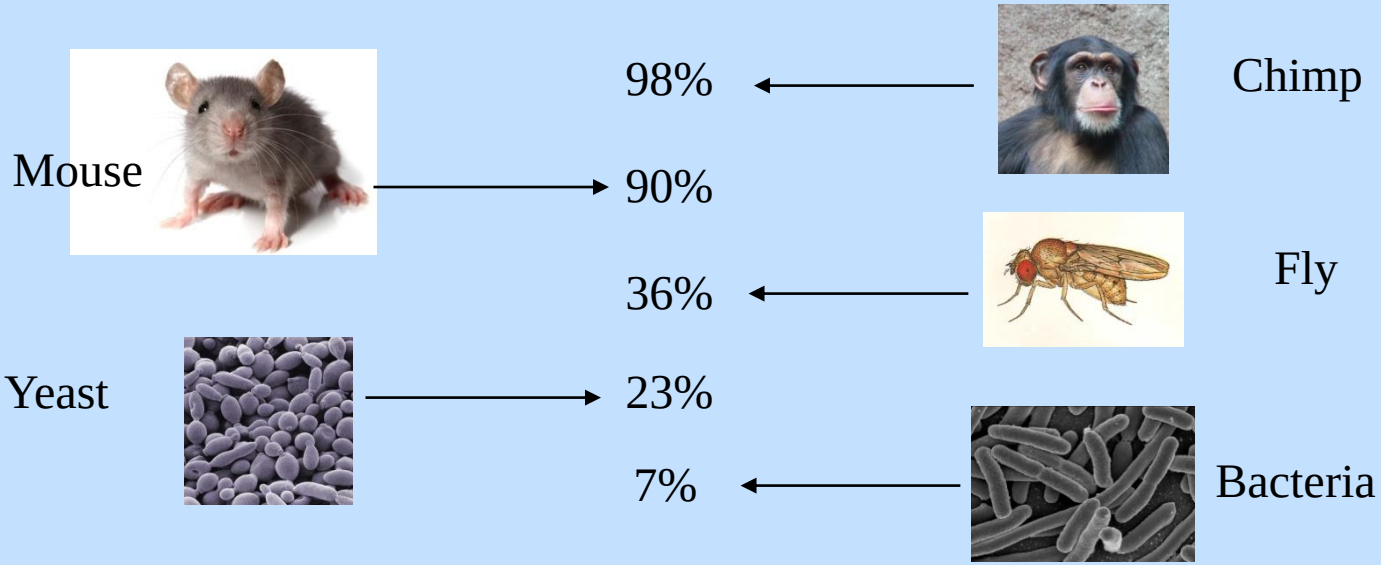
Sequence Alignment problems

- Given two sequences, find their best alignment: Match with insertion/deletion of min cost.
- Same for several sequences

- "Workhorse" of Bioinformatics!
- Key challenge: huge volume of data (more on this later)

Understanding differences

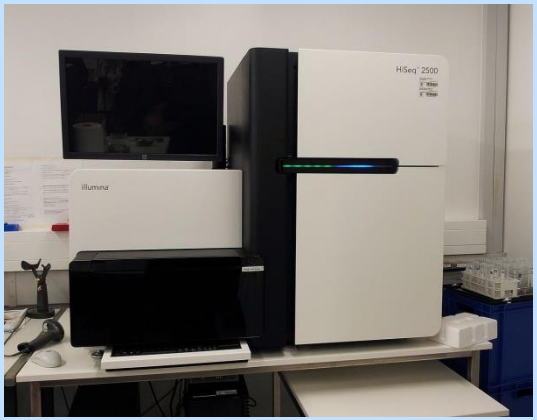
2 persons: 99.9% similarity



- Lots of common ground of model organisms with humans: many / most genes are common – but with mutations

Sequencing

- Sequencing: reading the sequence of bases in a given DNA or RNA molecule.
- To be sequenced, long sequences must be broken into short segments called "reads"
- Classical approach: gel electrophoresis; produces 10-100 longish reads (~1000nt) per run
- Next-Generation Sequencing: the modern sequencing techniques, producing many millions of short reads (100-300 nt) per run



One of Many NGS analysis problems

READ MAPPING: Given 10^8 reads, each 100bp long, and a reference genome of length 10^7 – 10^9 bp, quickly find all the matches of each read in the genome, with differences

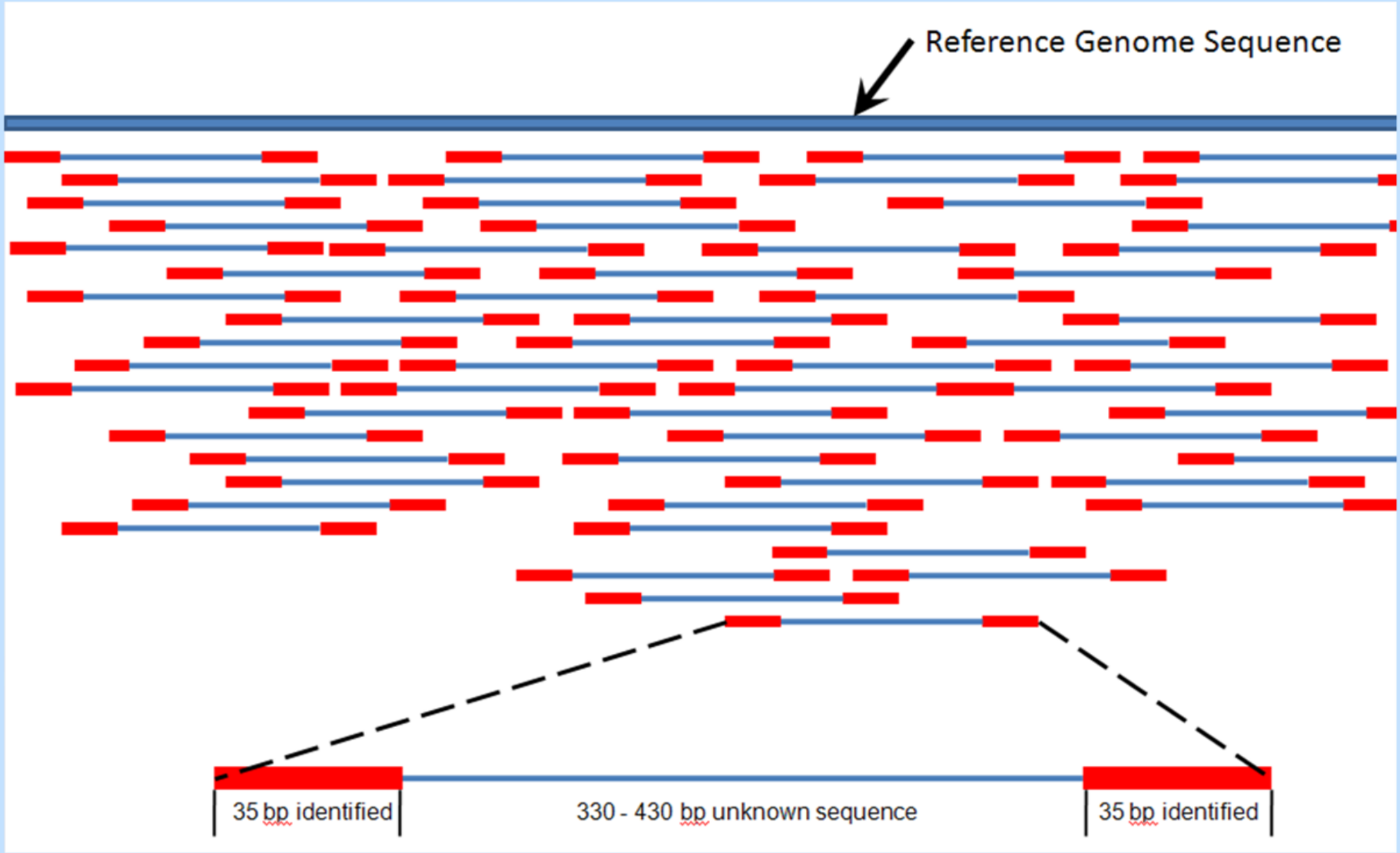
- The simple alignment solution: way too slow

- Need better algorithms, sacrificing as little accuracy as possible for far higher speed and smaller space

- An ongoing challenge: By 2025 the amount of DNA sequences is expected to reach 10^{21} bp…

55

Utilize RNA-sequencing and alignment to evaluate RNA levels



Gene Expression analysis

- We can measure the amount of expression of every gene of a person quickly and cheaply, producing her expression profile

- A working assumption: Expression ~ activity
- => compare many profiles and infer biology from the commonalities and differences!

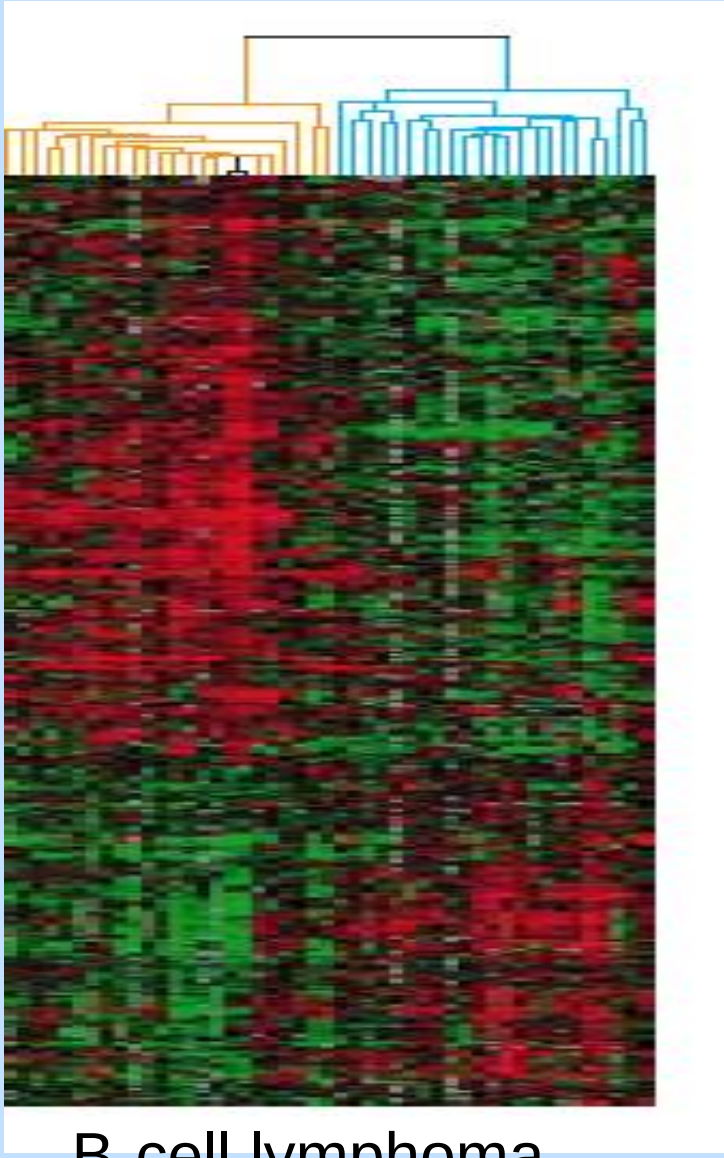
Clustering problem

Given the expression profiles of many individuals, partition the profiles into groups such that
-Within each group profiles are similar
-Between different groups profiles are dissimilar

Example: Clustering of B cell lymphoma samples, no known subtypes

Output: Two molecularly distinct forms of B-cell lymphoma which had distinct gene expression patterns

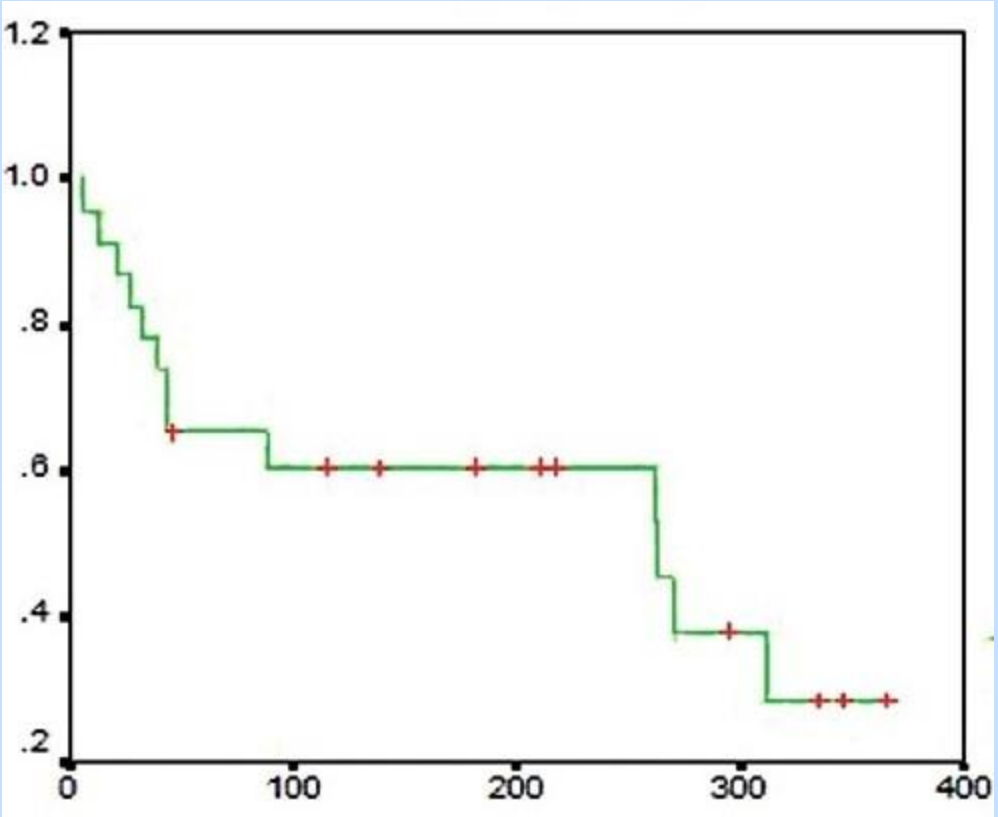
Question: What is the clinical relevance of these distinct forms?



Evaluate clinical relevance
Kaplan-Meier plot

Fraction of surviving subjects

("survival probability")

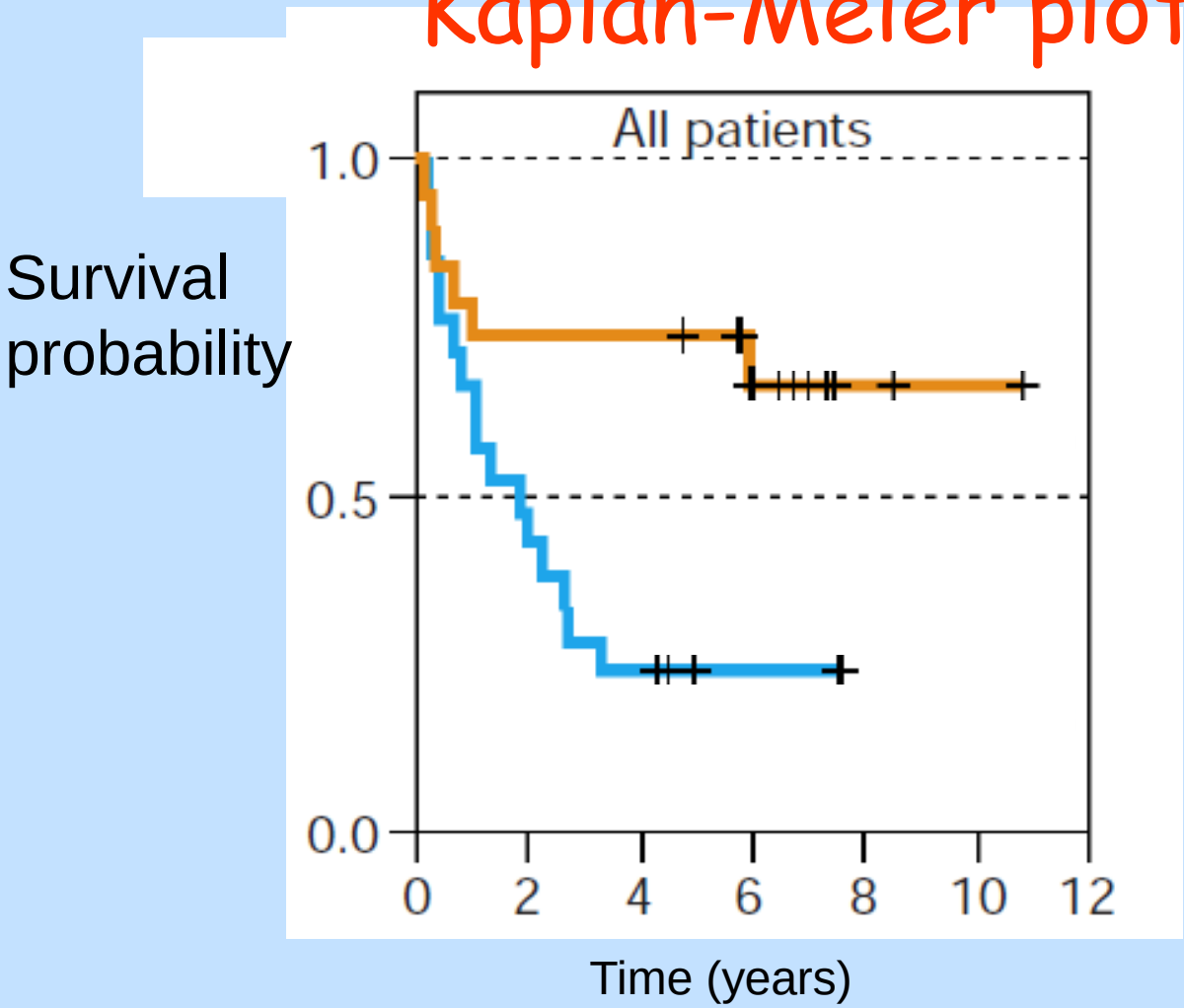


Time (weeks)

The plot presents the fraction of subjects surviving until a certain time

Evaluate clinical relevance

Kaplan-Meier plot



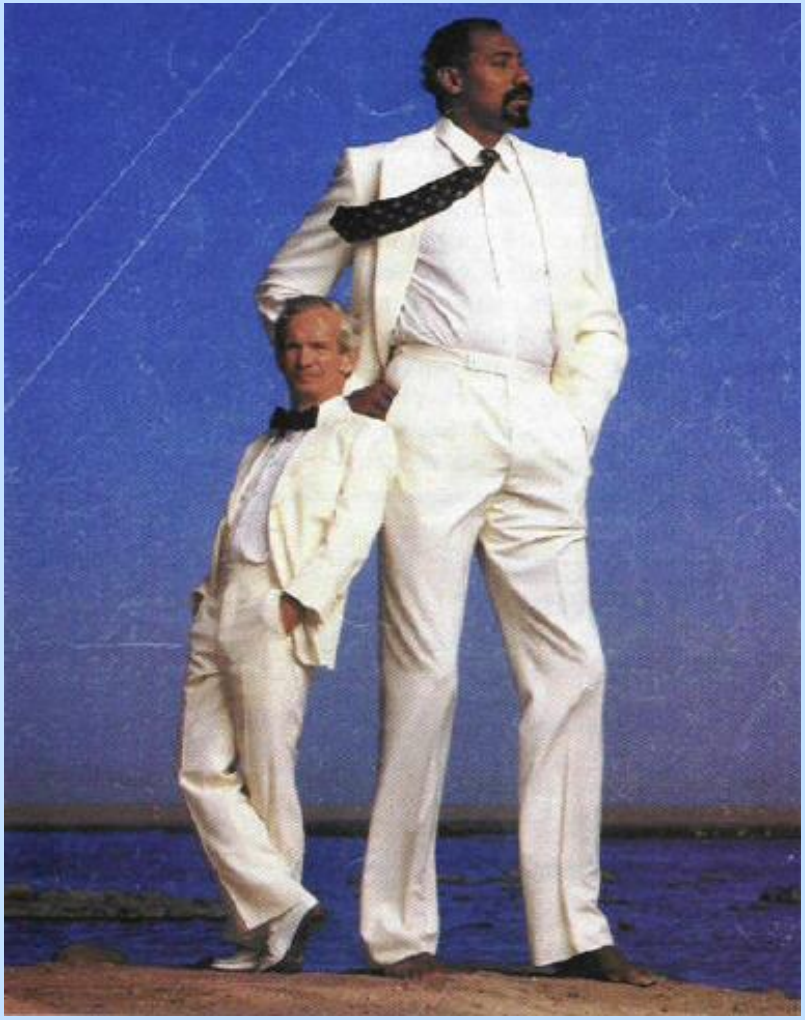
Kaplan-Meier plot of overall survival of B-cell lymphoma patients clustered on the basis of gene expression profiling.

ADDITIONAL EXAMPLES

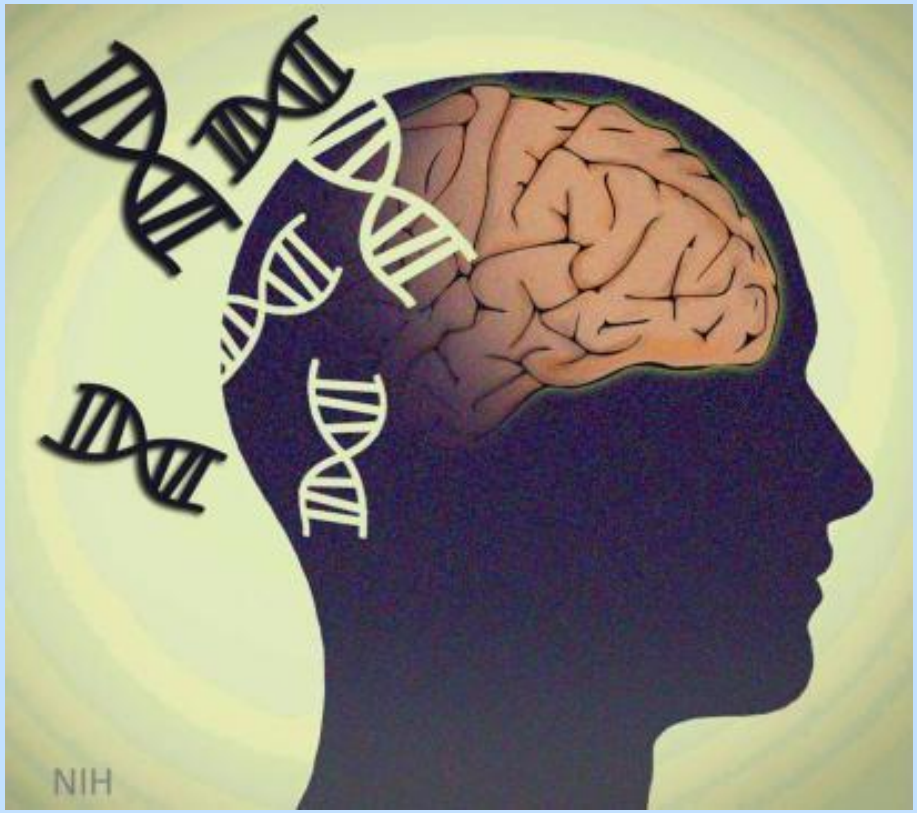
§ Computational genetics

- DNA of two human beings is ~99.9% identical
- Phenotype and disease variation is due these 1/1000 mutations

Challenges:
- Associate mutations to specific disease
- Deal with huge datasets (noise and statistics)



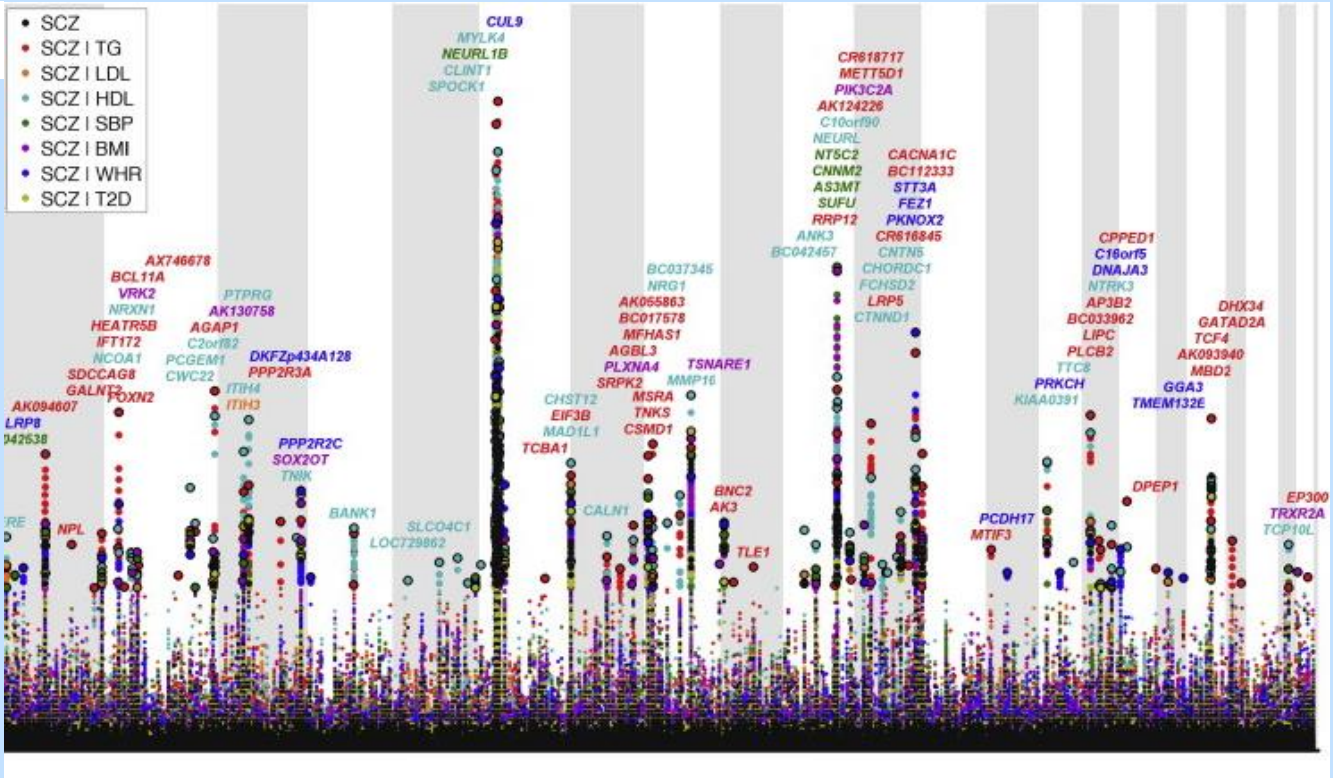
Schizophrenia



Schizophrenia is one of the most prevalent, tragic, and frustrating of all human illnesses, affecting about 1% of the human population.
Decades of research have failed to provide a clear cause in most cases, but family clustering has suggested that inheritance must play some role.

Searching for the genetic basis of Schizophrenia

Association score



Genomic position

Exome sequencing: 2K USD per patient (at the time of the study).

Broad institute: 2000 patients per week!

Data here: 2500 healthy & 2500 Schizophrenia patients

Searching for a gene that confers submergence tolerance to rice

• Most rice strains die within a week of complete submergence – a major constraint to rice production in south and southeast Asia.

• Some strains are highly tolerant and survive up to two weeks of complete submergence (no aerobic respiration, no photosynthesis) and renew growth when the water subsides

→The bioinformatics field of 'computational genetics'
→ found a region near the centromere of chromosome 9 , called *sub1*.

Confirming the submergence tolerance *sub1* region

submergence-intolerant strain "Swarna"

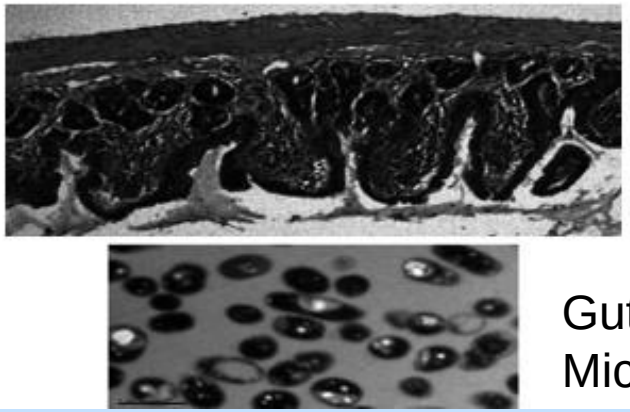
"Swarna"-*sub1*

submergence-tolerant strain, *Sub1* donor

Xu et al. 2006

§ Metagenomics

Sampling the human gut



Gut Microbiome

Metagenomic analysis

- Noval genes
- Antibiotic resistant genes
- Functional dysbiosis
- Microbial diversity

Metagenomics: sampling the human gut

ARTICLE

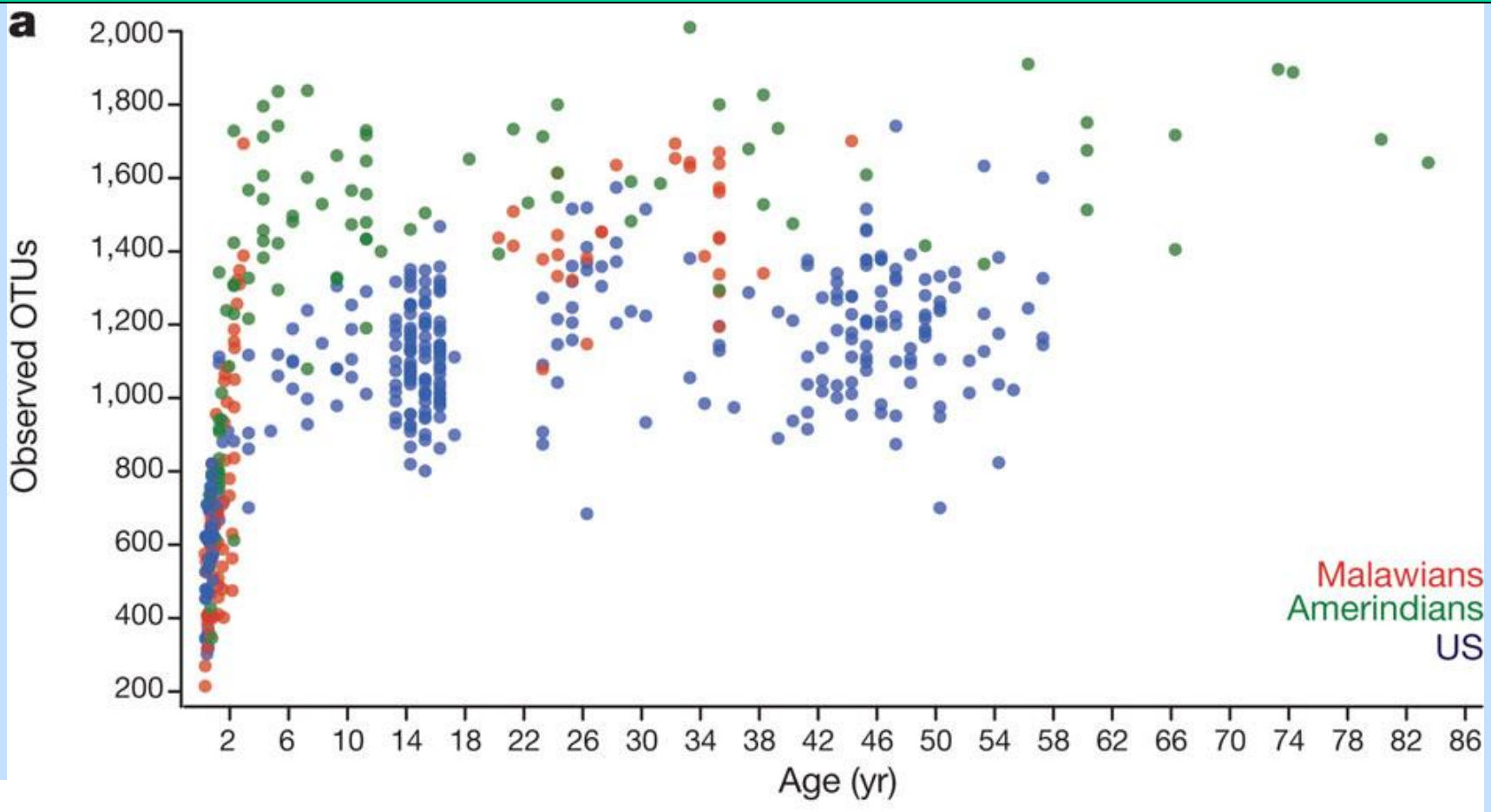
doi:10.1038/nature11053

Human gut microbiome viewed across age and geography

Tanya Yatsunenko[1], Federico E. Rey[1], Mark J. Manary[2,3], Indi Trehan[2,4], Maria Gloria Dominguez-Bello[5], Monica Contreras[6], Magda Magris[7], Glida Hidalgo[7], Robert N. Baldassano[8], Andrey P. Anokhin[9], Andrew C. Heath[9], Barbara Warner[2], Jens Reeder[10], Justin Kuczynski[10], J. Gregory Caporaso[11], Catherine A. Lozupone[10], Christian Lauber[10], Jose Carlos Clemente[10], Dan Knights[10], Rob Knight[10,12] & Jeffrey I. Gordon[1]

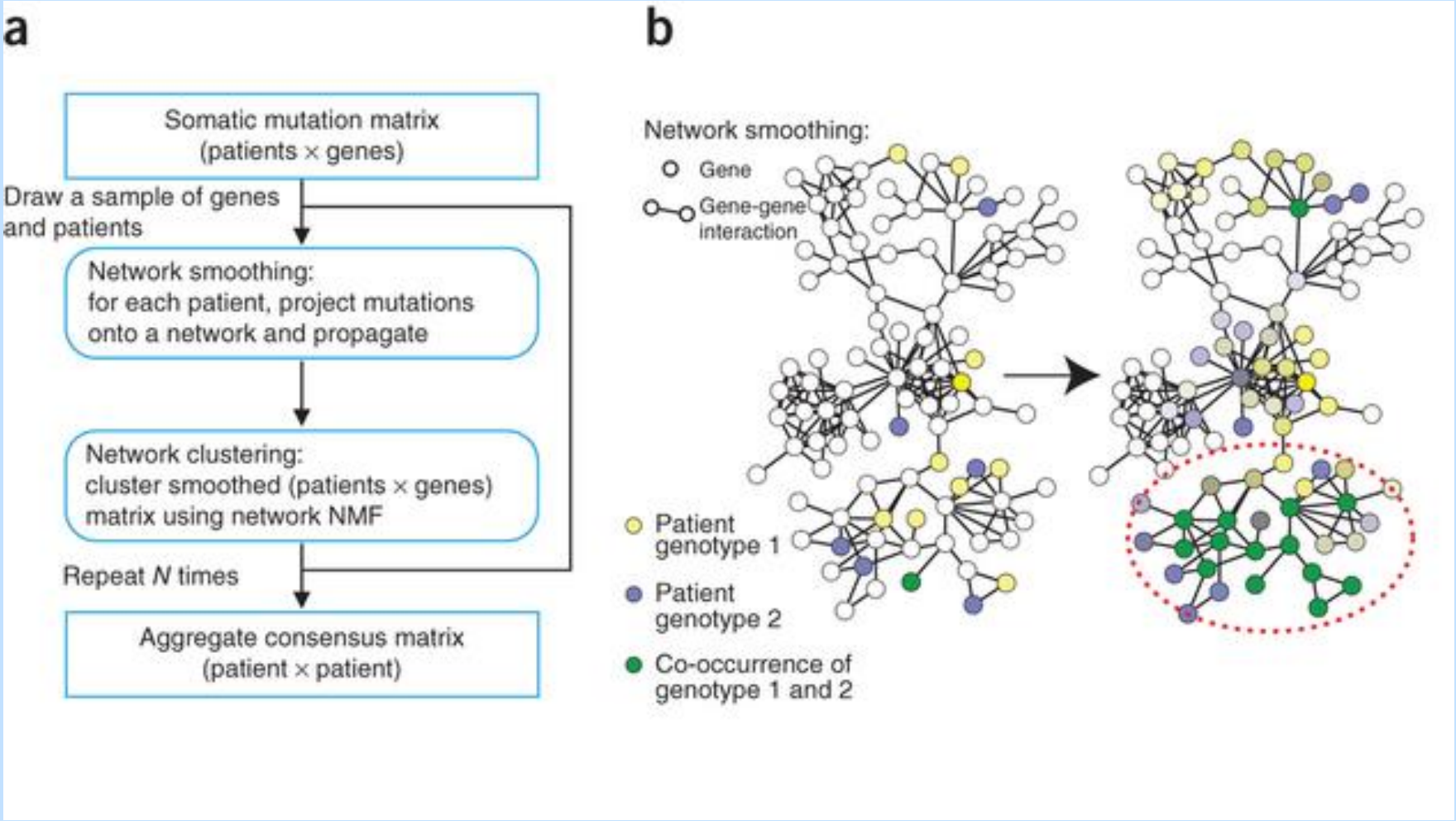
Gut microbial communities represent one source of human genetic and metabolic diversity. To examine how gut microbiomes differ among human populations, here we characterize bacterial species in fecal samples from 531

Bacterial diversity increases with age (based on NGS of fecal samples from 531 individuals)



§ cancer genomics

Network-based analysis of tumor mutations



Hofree et al. Nature methods 2013

§ Pathogenomics

revolutionizing HIV treatment

There are very efficient drugs for HIV

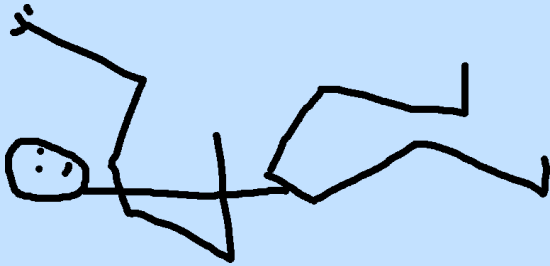
Many viruses in blood

DRUG,
+a few days

A few viruses in blood

DRUG,
+more days

Many viruses in blood



Explanation: the virus mutates and some viruses become resistant to the drug.

Solution: combination of drugs (cocktail).
But: do not give drugs for which the virus is already resistant. For example, if one was infected from a person who receives a specific drug.

The question: how does one know to which drugs the virus is already resistant?

Sequences of HIV-1 from patients who were treated with drug A:

AAGACGCATCGATCGATCGATCGATCGTACG
ACGACGCATCGATCGATCGATCGATCGTACG
AAGACACATCGATCGTTCGATCGTACG

Sequences of HIV-1 from patients who were never treated with drug A:
AAGACGCATCGATCGATCGATCTTACG
AAGACGCATCGATCGATCGATCTTACG
AAGACGCATCGATCGATCGATCTTACG

drug A+

AAGACGCATCGATCGATCGATCGATCGTACG

ACGACGCATCGATCGATCGATCGATCGTACG

AAGACACATCGATCGTTCGATCGTACG

drug A-

AAGACGCATCGATCGATCGATCGATCTTACG

AAGACGCATCGATCGATCGATCGATCTTACG

AAGACGCATCGATCGATCGATCGATCTTACG

This is an easy example.

drug A+

AAGACGCATCGATCGATCGATCGTACG

ACGACGCATCGATCGATCGATCGTACG

AAGACACATCGATCATTCGATCATACG

drug A-

AAGACGCATCGATCTATCGATCTTACG

AAGACGCATCGATCTATCGATCTTACG

AAGACGCATCGATCAATCGATCGTACG

This is NOT an easy example. This is an example of **a classification problem**.

Genotypic predictors of human immunodeficiency virus type 1 drug resistance

Soo-Yon Rhee*, Jonathan Taylor†, Gauhar Wadhera*, Asa Ben-Hur‡§, Douglas L. Brutlag‡, and Robert W. Shafer*¶

Division of Infectious Diseases, Departments of *Medicine, †Statistics, and ‡Biochemistry, Stanford University, Stanford, CA 94305

Communicated by Bradley Efron, Stanford University, Stanford, CA, August 28, 2006 (received for review December 5, 2005)

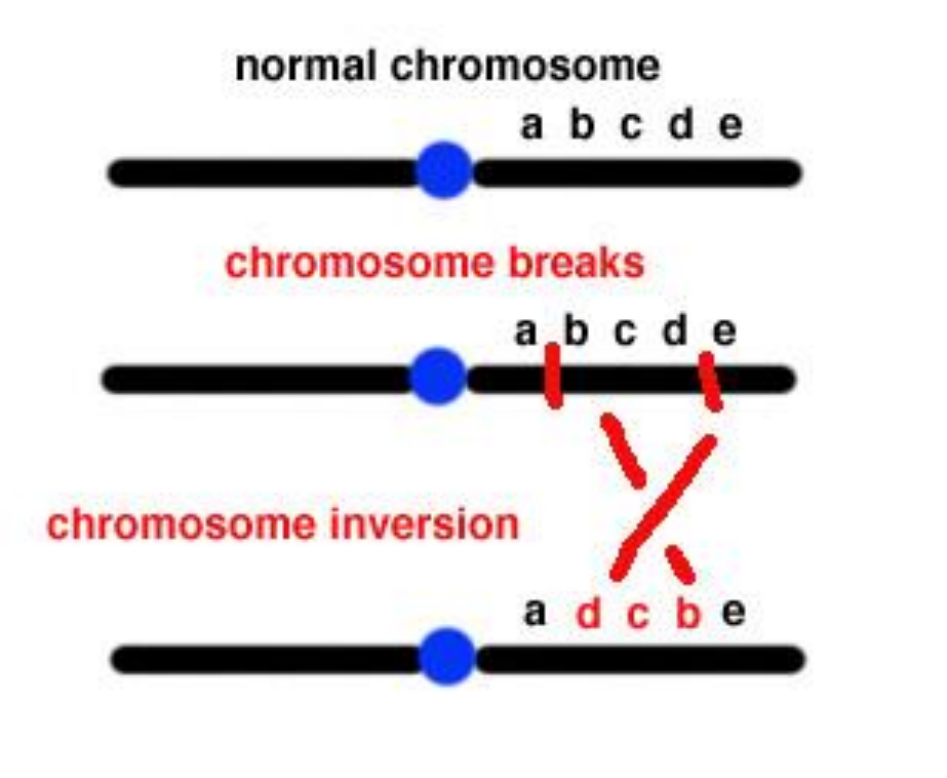
Understanding the genetic basis of HIV-1 drug resistance is essential to developing new antiretroviral drugs and optimizing the use of existing drugs. This understanding, however, is hampered by the large numbers of mutation patterns associated with cross-resistance within each antiretroviral drug class. We used five

Results

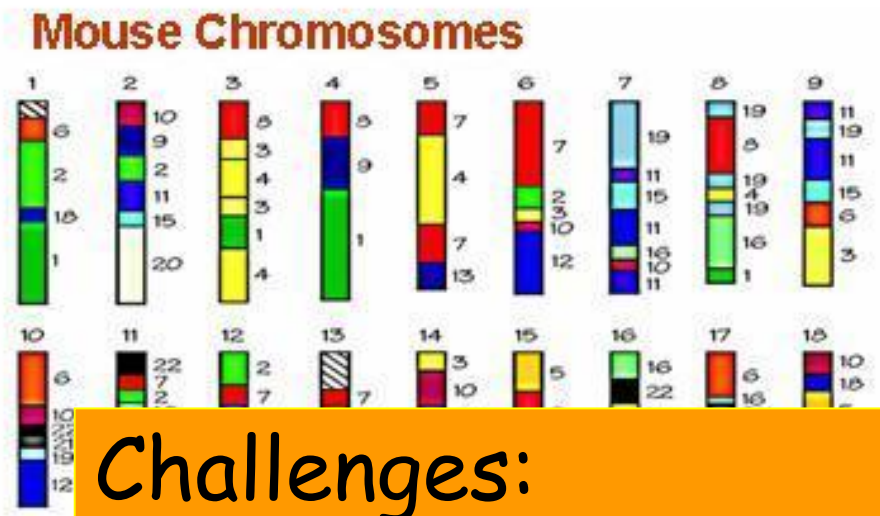
Drug Susceptibility Results, Input Mutations, and Learning Methods. For each of the three drug classes, we created four mutation sets that included (*i*) a complete set of all mutations present in ≥2

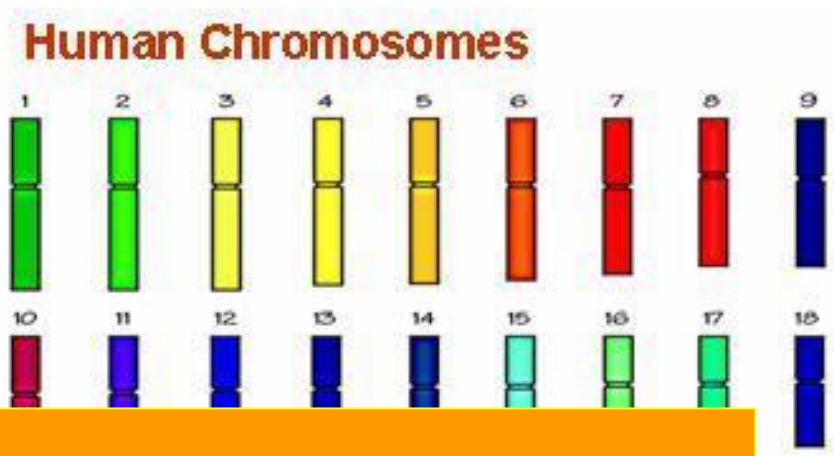
§ Genome Rearrangements

- Rearrangement is a change in the order of complete segments along a chromosome.



Genome Rearrangements





Challenges:

- Reconstruct the evolutionary path of rearrangements
- Shortest sequence of rearrangements between two permutations

More Examples

- Sequencing cancer genomes
- Large scale proteomics studies
- Single-cell genomics

And much more!

The End

Basic Biotechnology

Restriction Enzymes

- Natural role: break foreign DNA entering the cell.
- Ability:
  - Breaks the phosphodiester bonds of a DNA upon appearance of a certain cleavage (cut) sequence.
  - Different sequence for each enzyme
  - Hundreds of different enzymes known.
- Digestion = application of restriction enzymes to a sequence.

Cloning

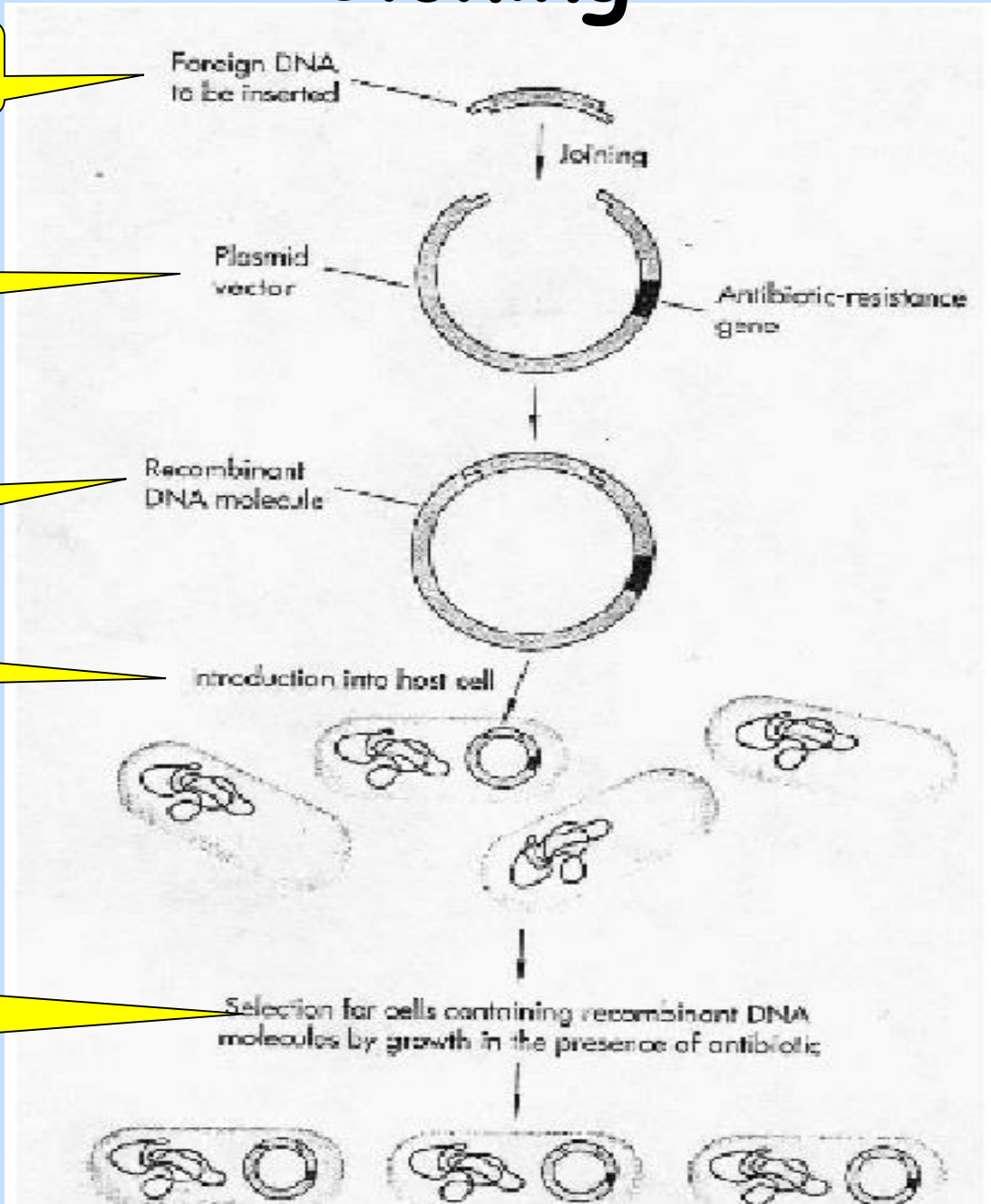
Foreign DNA

Cloning vector (plasmids)

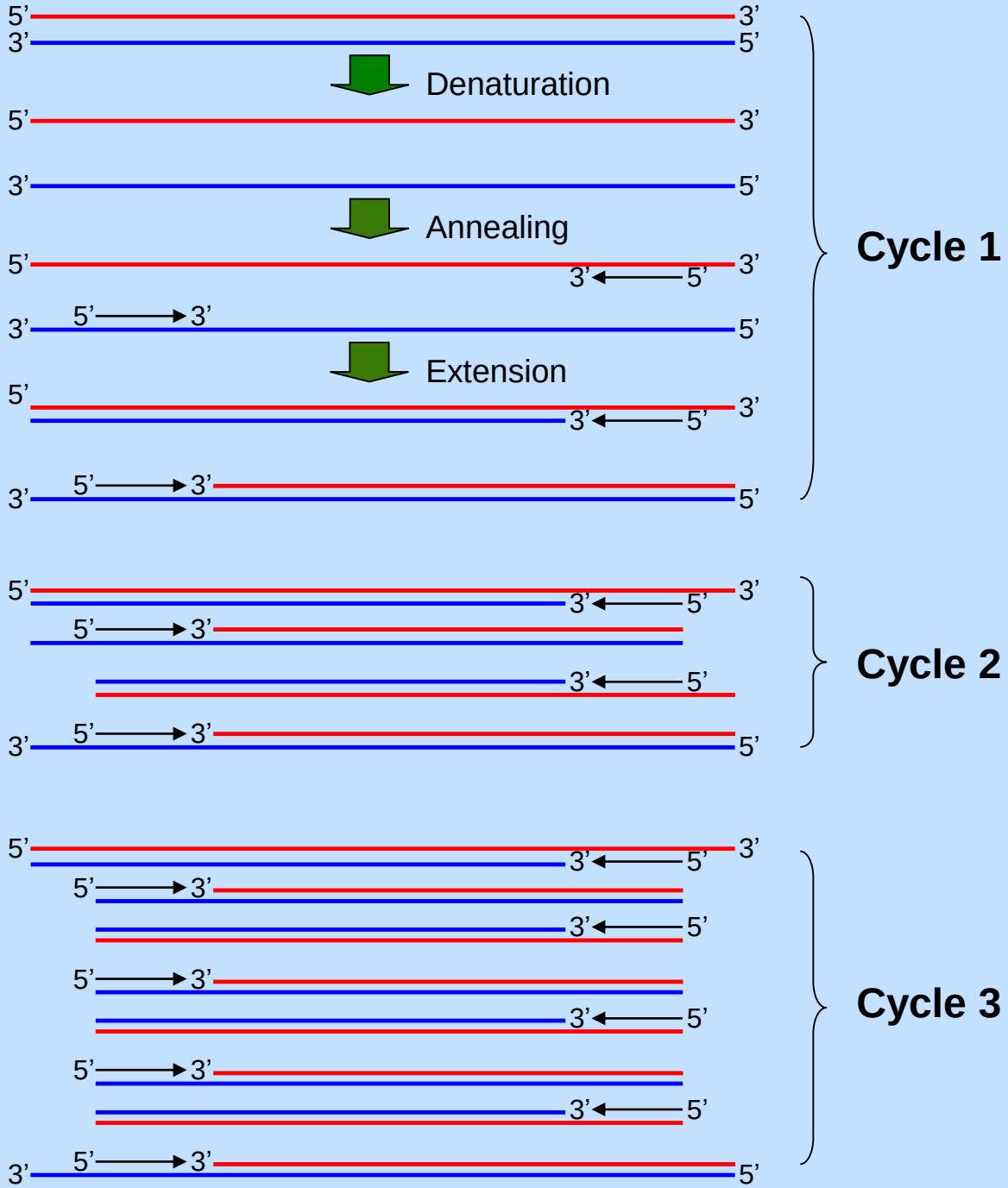
Recombinant DNA

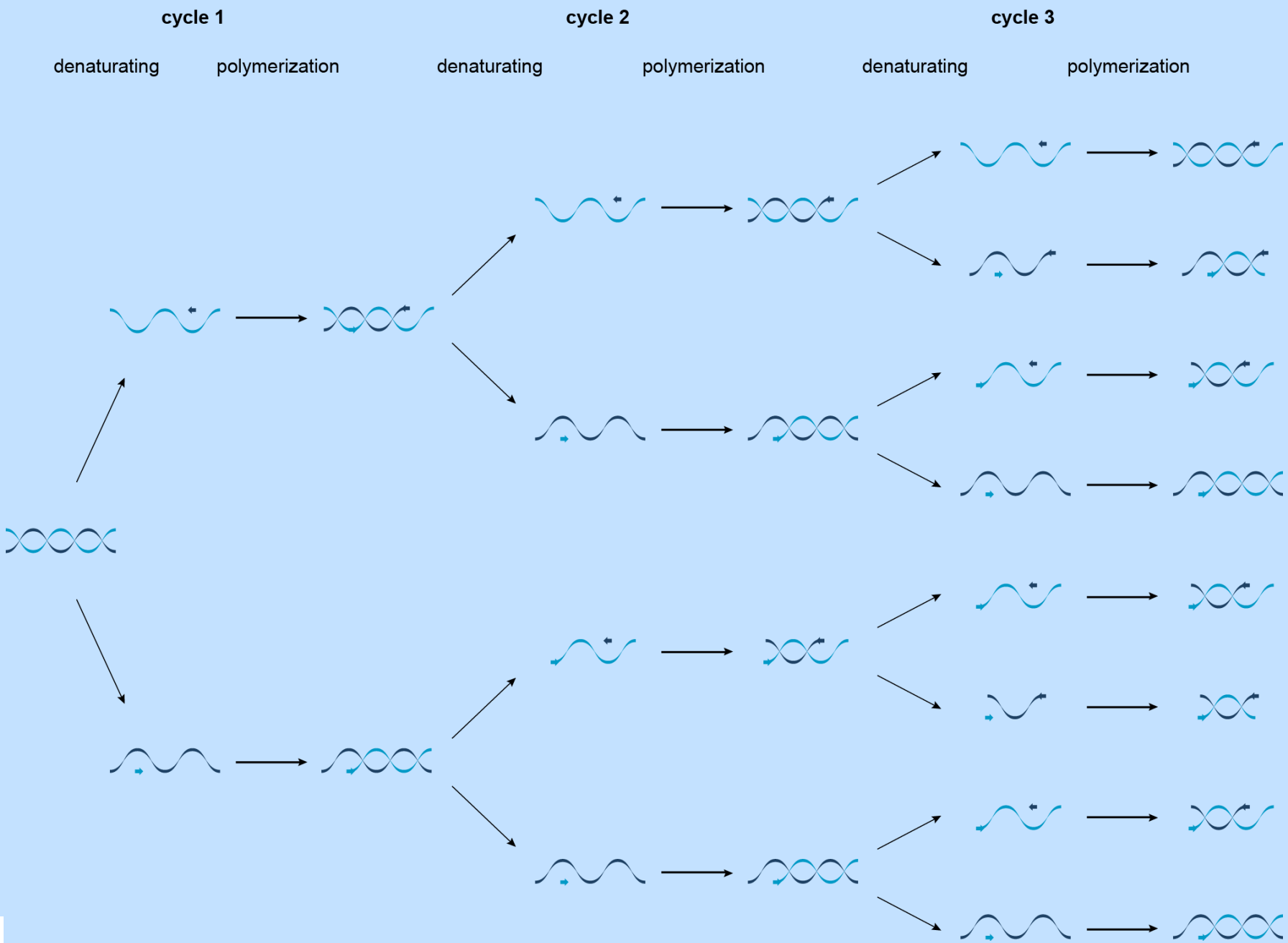
Introduction into host cell

Use of antibiotics to grow recombinant cells



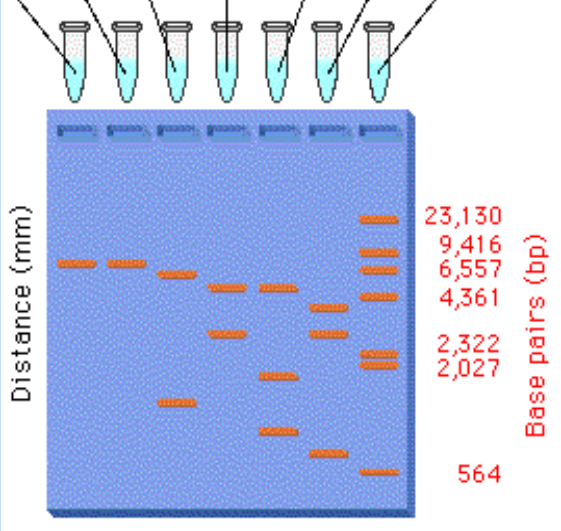
PCR


5'
3'
3'
5'
Denaturation
5'
3'
3'
5'
Annealing
5'
3'
3'
5'
5'
3'
3'
5'
Extension
5'
3'
3'
5'
5'
3'
3'
5'
Cycle 1
5'
3'
3'
5'
5'
3'
3'
5'
3'
5'
5'
3'
3'
5'
Cycle 2
5'
3'
3'
5'
5'
3'
3'
5'
5'
3'
3'
5'
5'
3'
3'
5'
5'
3'
3'
5'
Cycle 3

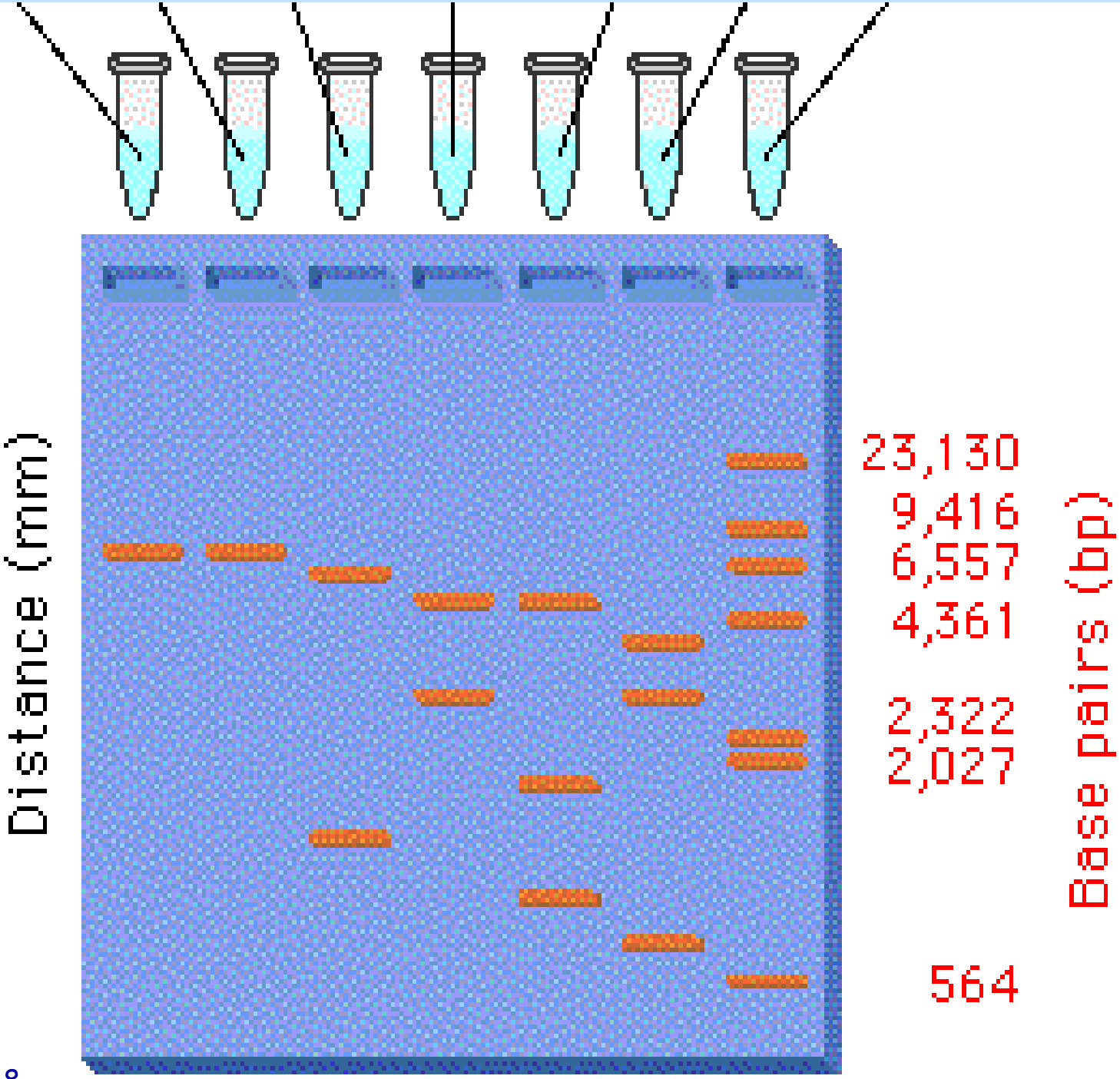


http://www.atdbio.com/content/20/Sequencing-forensic-analysis-and-genetic-analysis

Gel Electrophoresis



- Use: "race" digested DNA fragments through electrically charged gel
- Goals:
  - Separate a mixture of DNA fragments
  - Measure length of DNA fragments
- How does it work:
  - smaller molecule travel faster than larger ones
  - same size and shape ⇒ the same movement speed

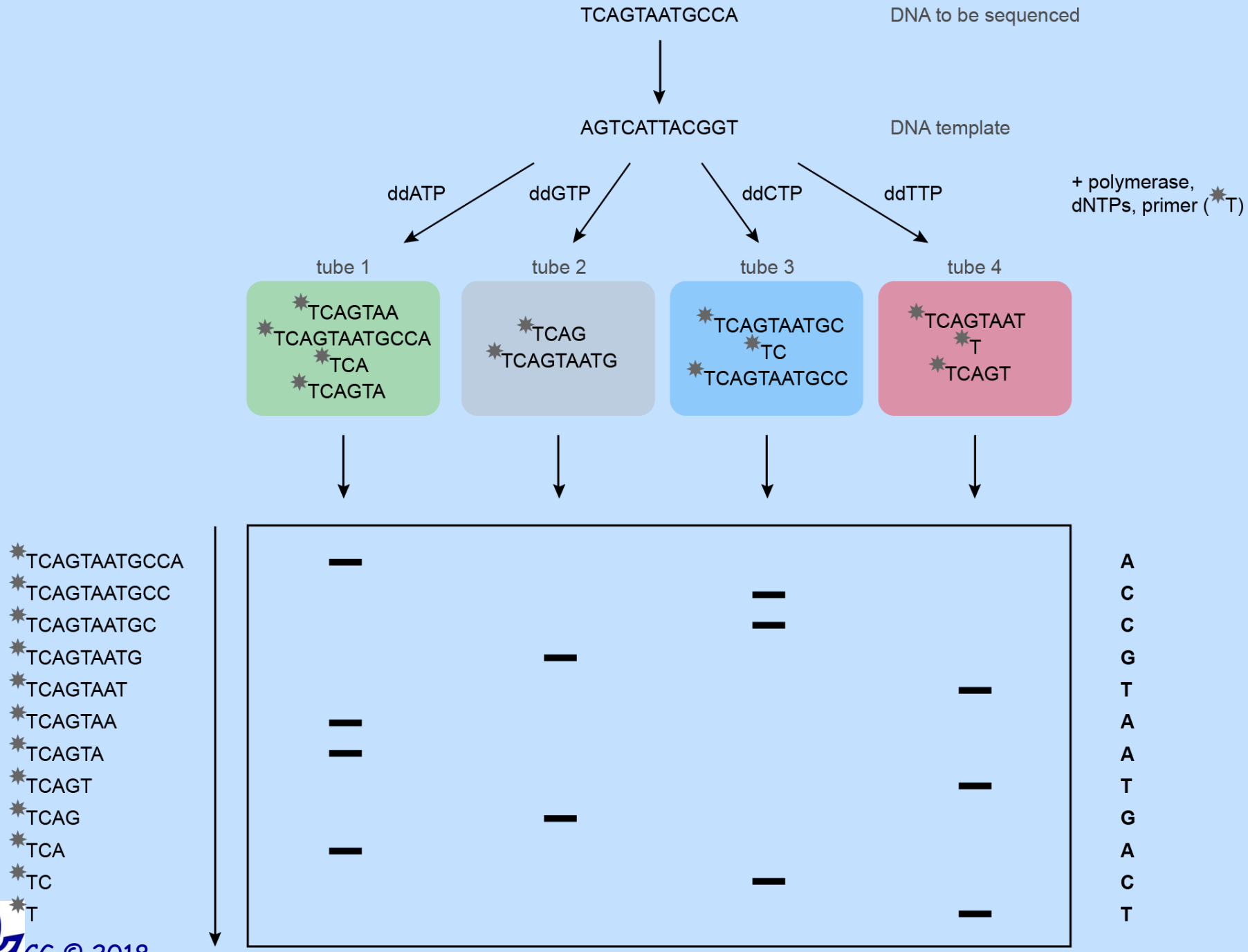


http://dlab.reed.edu/projects/vgm/vgm/VGMProjectFolder/VGM/RED/RED.ISG/mapping.html

```
---A-----A- •
-CC---CC--- •
T---T------ •
-----G----G •
```

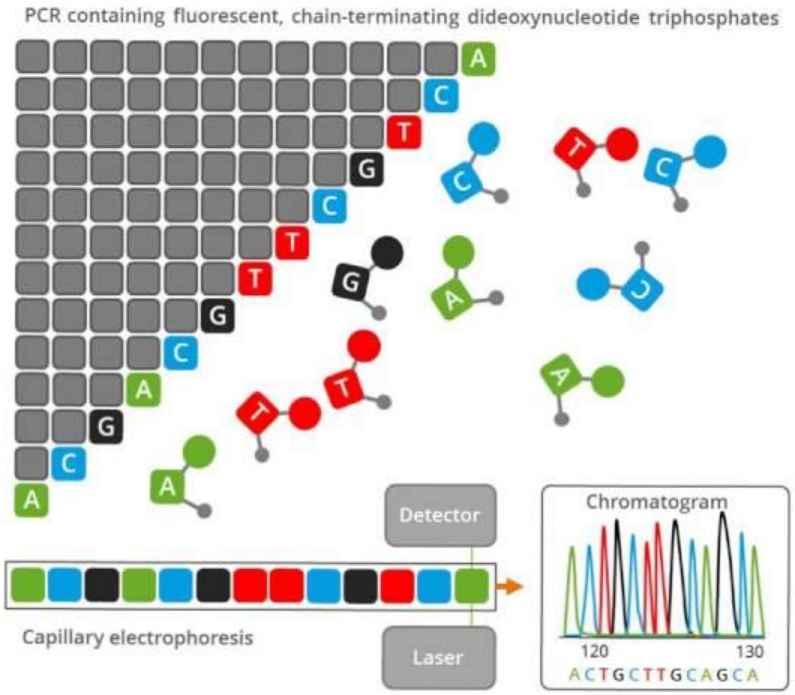
Sequencing

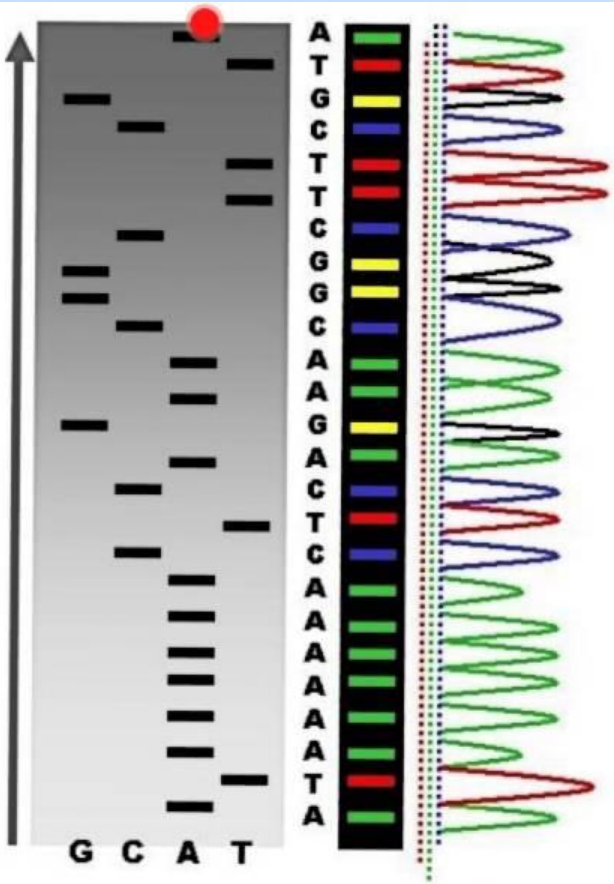
- Sequencing: determining the sequence of bases in a given DNA molecule.
- Classical approach: gel electrophoresis
- Basic idea: knowing the lengths of all prefixes ending with letter X gives a partial seq
- Creating DNA strands of different lengths : catalyzing replication in environment with "terminator" A*.
- Repeat separately with C*, G*, T*
- Abilities: reconstructs sequences of 500-1000 nucleotides.



http://www.atdbio.com/content/20/Sequencing-forensic-analysis-and-genetic-analys

Sanger Sequencing





Sanger sequencing uses ddNTPs (dideoxynucleotide triphosphates) which do not have a free 3' OH mixed in with dNTPs
Whenever the DNA polymerase incorporates a ddNTP it won't be able to add any other nucleotides
Then gel electrophoresis is used to separate the DNAvv

https://www.youtube.com/watch?v=593zWZNwbJI

The End (now for real)