Introduction to Bioinformatics for Computer Scientists

Lecture 2

Preliminaries

- Email: Alexandros.Stamatakis@kit.edu

  → please send an email

  → to be added to the course mailing list

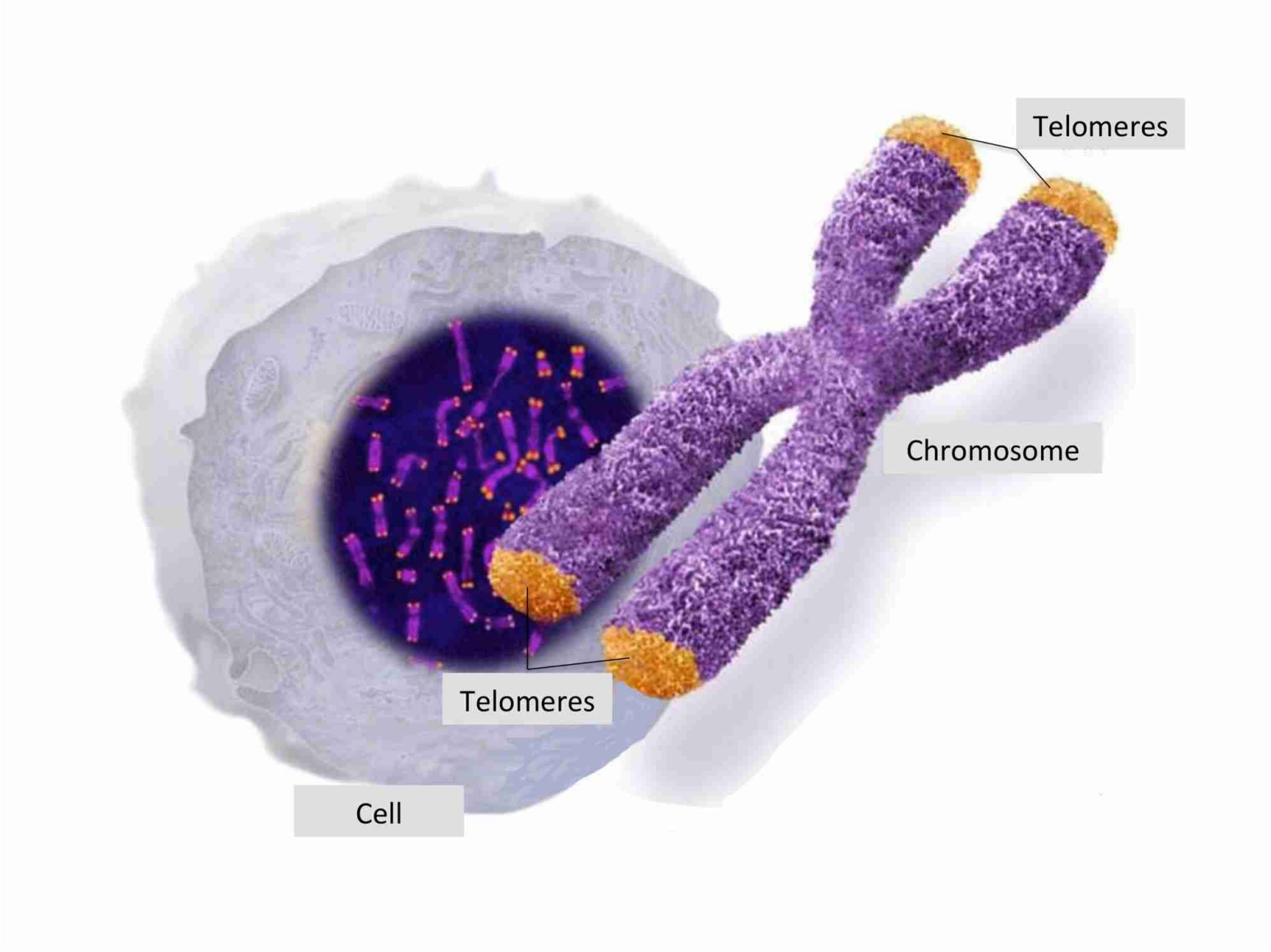
Last lecture

- Sequence data/sequence
- Nucleotide/base-pair
- DNA/RNA
- Ambiguity coding
- Sequencing
  - Sanger Sequencing
  - Next Generation Sequencing
- Genome
- Model Organism
- Double-stranded DNA
- Coding versus non-coding DNA

My Homework

- The human genome project (from Wikipedia)
  - The project ended up costing less than expected at about $2.7 billion (Financial Year 1991). When adjusted for inflation, this costs roughly $5 billion (Financial Year 2018).
  - The project did not sequence all DNA in human cells. It sequenced only *euchromatic* (Euchromatin comprises the most active portion of the genome within the cell nucleus) regions of the genome, which make up 92.1% of the human genome.
  - In May 2020, ... 79 "unresolved" gaps approx. 5% of the human genome
  - Months later new long-range sequencing techniques … led to the first telomere-to-telomere, truly complete sequence of a human chromosome, the X-chromosome.
  - In 2021 it was reported that the Telomere-to-Telomere (T2T) consortium had filled in all of the gaps. Thus there came into existence a complete human genome with almost no gaps, but it still had five gaps in ribosomal DNA.
- For more details see https://en.wikipedia.org/wiki/Human_Genome_Project

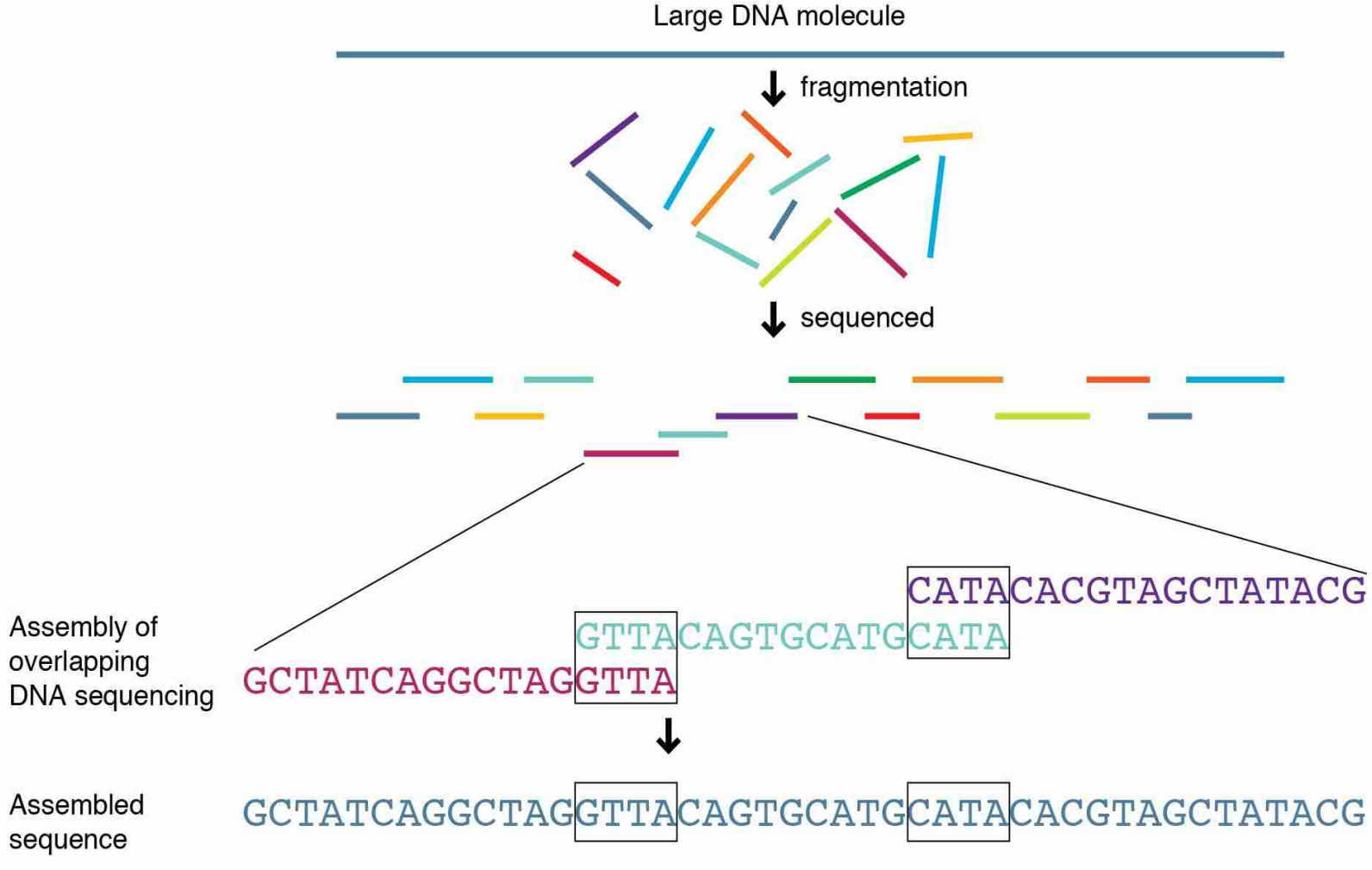
Telomeres – Greek: *end parts*



Today's outline

- More terminology & biological background

Shotgun Sequencing

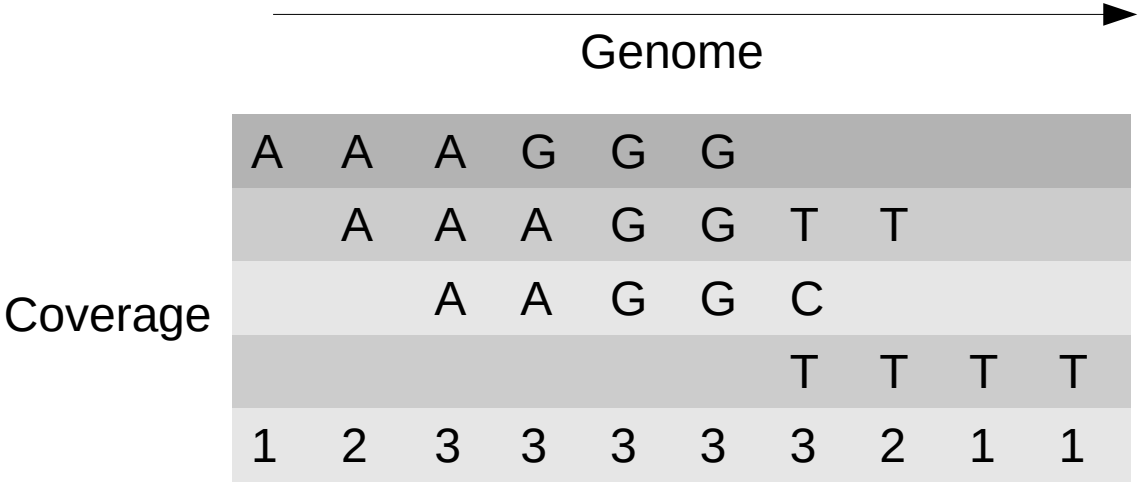


Shotgun Sequencing

- In the last lecture: we can read fragments up to a length of ≈ 1000 bp

  → 1000 bp correspond roughly to the length of an average gene
- What do we do for reading *genomes*?

  1) Break up genome randomly into fragments
  2) Read fragments
  3) Assemble fragments into a genome with computers
- Important characteristics:
  - *Coverage*: how many fragments/reads cover one nucleotide on the genome

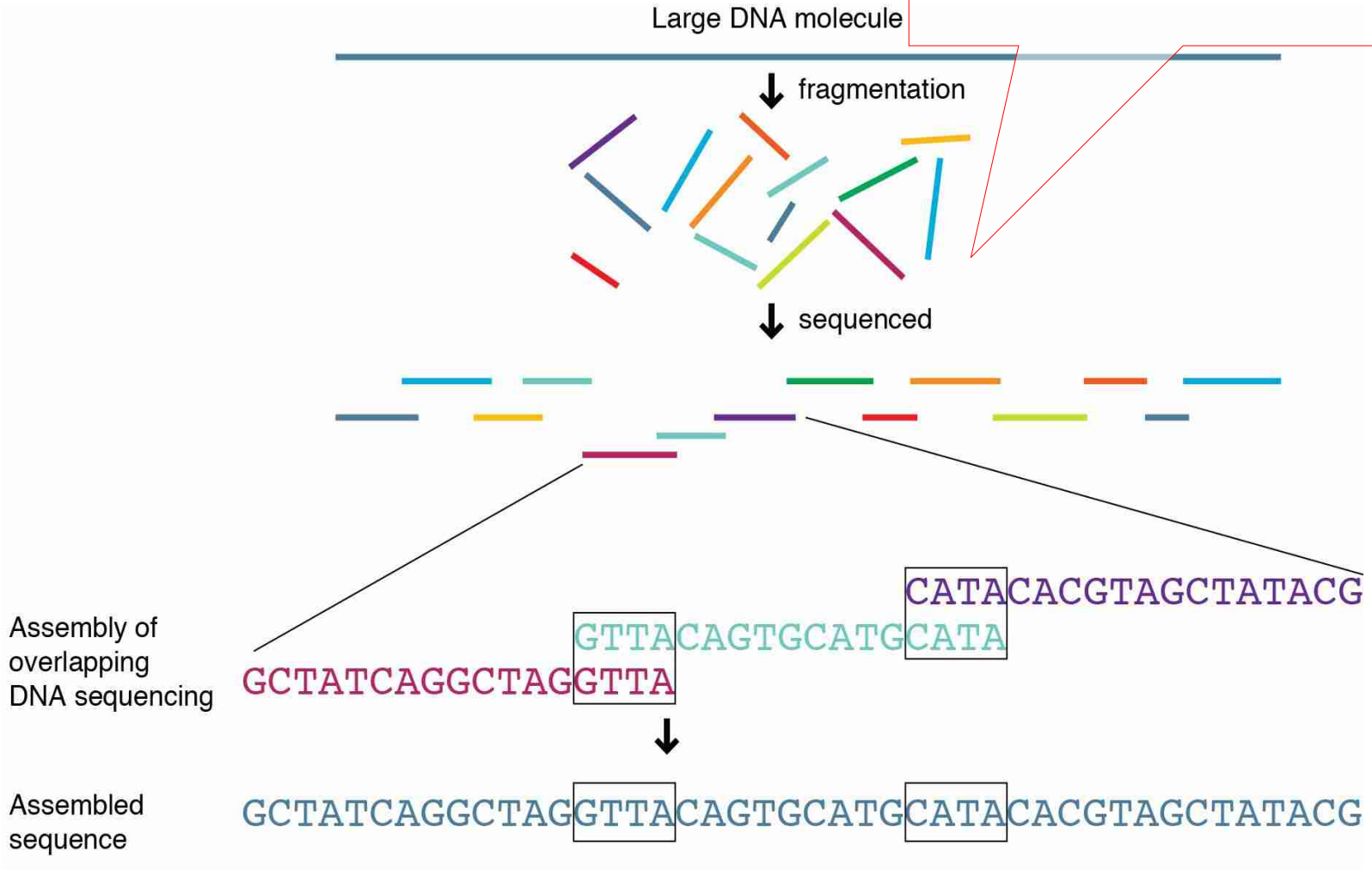


Shotgun Sequencing

- In the last lecture: we can read fragments up to a length of ≈ 1000 bp (Sanger Sequencing)
- What do we do for reading genomes?
  1) Break up genome randomly into fragments
  2) Read fragments
  3) Assemble fragments into a genome with computers
- Important characteristics:
  - *Coverage*
  - *Fragment* length
  - *Paired-end* versus *Single-end* reads
  - *De novo* versus *by reference* assembly

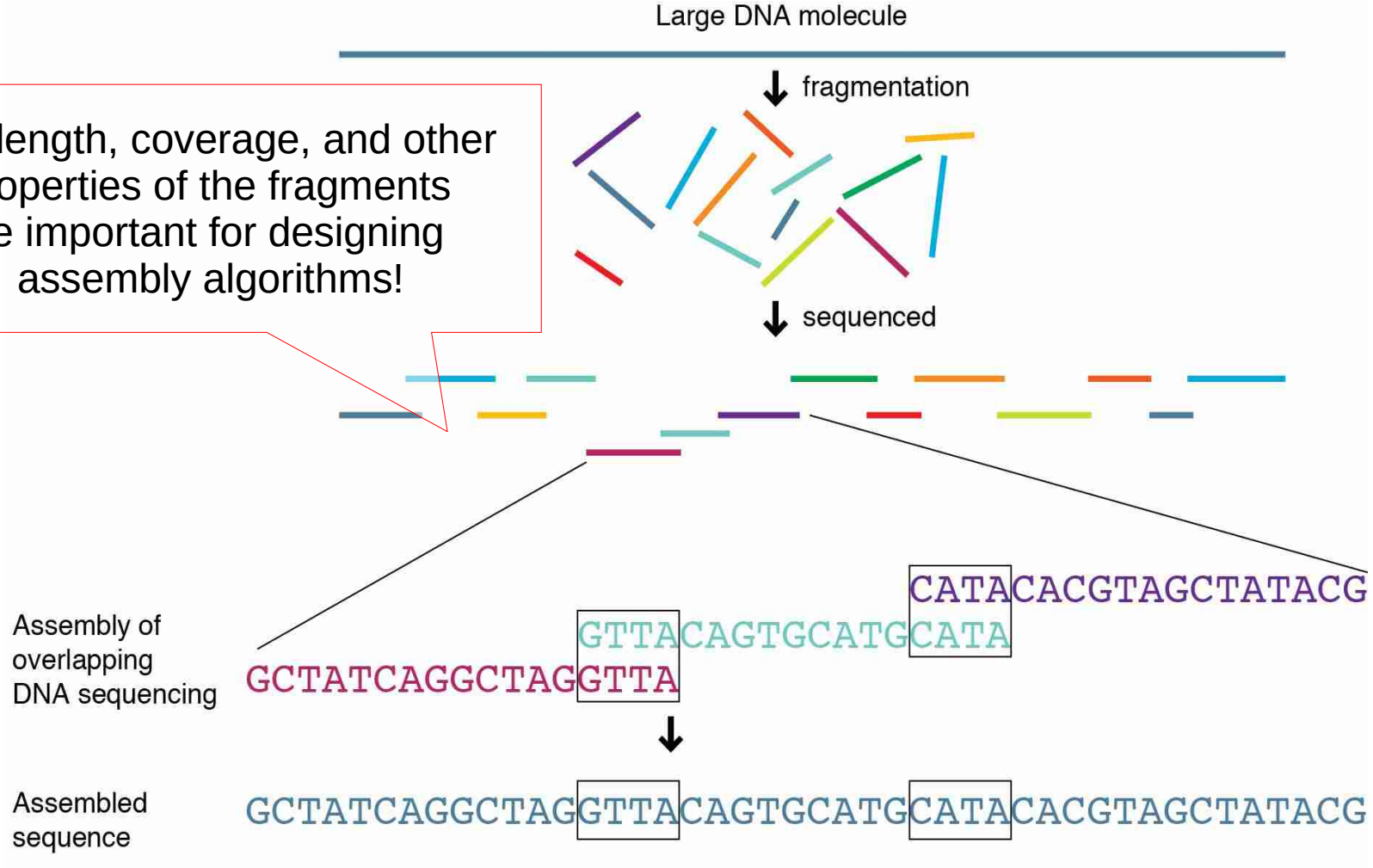
Shotgun Sequencing

This is a simplistic view,
omitting many technical (lab) details



Shotgun Sequencing

The length, coverage, and other properties of the fragments are important for designing assembly algorithms!



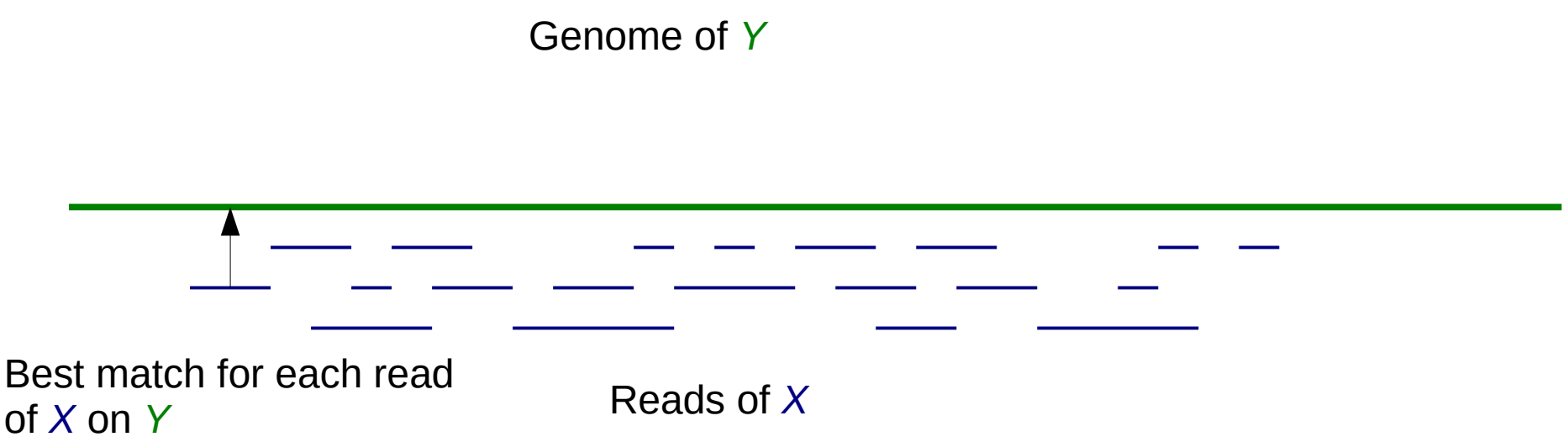
De novo versus *by reference* assembly

- There are two ways to conduct assemblies
- *By reference*: we want to assemble the genome of species *X*

  → there is a closely related species *Y* whose genome is already available

  → map reads of *X* to genome of *Y* to assemble them

  → also known as read mapping



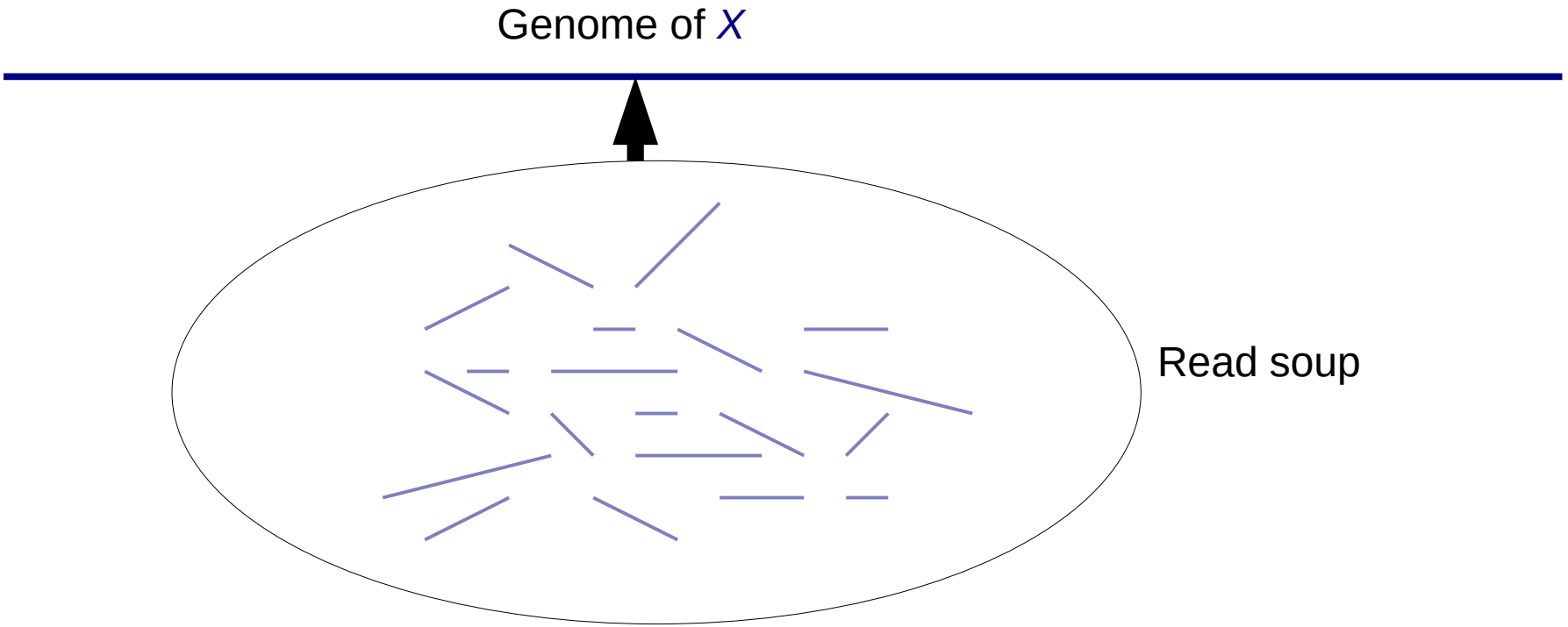
De novo versus *by reference* assembly

- There are two ways to conduct assemblies
- *De novo*: we want to assemble the genome of species *X*

  → there is no closely related species of *X* whose genome is already available

  → assemble genome out of read soup

  → computational problem is much harder, in particular when reads are short



Paired-end Reads

- Two DNA fragments at both ends of the sequence read

- AAAGGGTTT----------TTTTTTAAAGGC
- We know the distance between fragments denoted by - here which is *13*
- This is the same for all paired-end reads

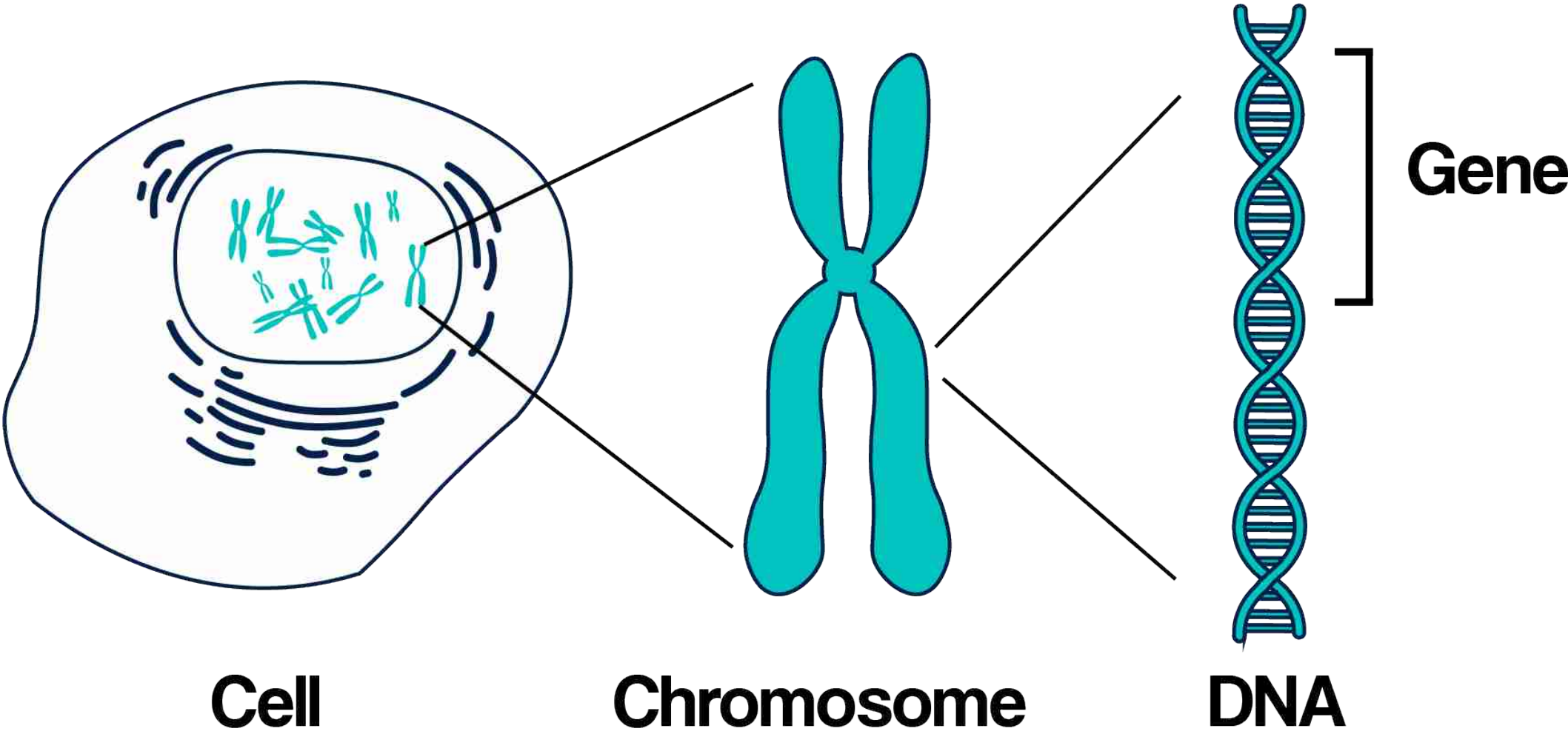
  → contains additional information

  → makes assembly process easier

Back to DNA

- DNA encodes – *coding DNA*
  - Protein information
  - RNA information
- DNA is also know as the *blueprint of life*
- In a cell, the DNA is organized in long molecules called *Chromosomes*

A Chromosome



Back to DNA

- DNA encodes – *coding DNA*
  - Protein information
  - RNA information
- DNA is also know as the *blueprint of life*
- In a cell, the DNA is organized in long molecules called *Chromosomes*
- Keep in mind
  - Some parts of the DNA are *coding*
  - Some parts of the DNA are *non-coding (junk DNA)*

What's a *gene*?

- The coding parts of the DNA
- Each *gene* (a contiguous string of DNA) encodes for
  - Either *RNA*
  - Or a *protein*

RNA & Protein sequences

- In RNA we just replace character T by U
- Protein data has a *20* letter alphabet!
- *3* DNA/RNA characters encode for one protein character!
- We call such a triplet of DNA/RNA characters a *Codon*!
- With *3* DNA/RNA characters we could encode for *4 * 4 * 4 = 64* characters
- … but we only have *20*!
- There are some redundancies and other special cases

Protein Alphabet

Amino acid	Codons	Compressed	Amino acid	Codons	Compressed
Ala/A	GCT, GCC, GCA, GCG	GCN	Leu/L	TTA, TTG, CTT, CTC, CTA, CTG	YTR, CTN
Arg/R	CGT, CGC, CGA, CGG, AGA, AGG	CGN, MGR	Lys/K	AAA, AAG	AAR
Asn/N	AAT, AAC	AAY	Met/M	ATG	
Asp/D	GAT, GAC	GAY	Phe/F	TTT, TTC	TTY
Cys/C	TGT, TGC	TGY	Pro/P	CCT, CCC, CCA, CCG	CCN
Gln/Q	CAA, CAG	CAR	Ser/S	TCT, TCC, TCA, TCG, AGT, AGC	TCN, AGY
Glu/E	GAA, GAG	GAR	Thr/T	ACT, ACC, ACA, ACG	ACN
Gly/G	GGT, GGC, GGA, GGG	GGN	Trp/W	TGG	
His/H	CAT, CAC	CAY	Tyr/Y	TAT, TAC	TAY
Ile/I	ATT, ATC, ATA	ATH	Val/V	GTT, GTC, GTA, GTG	GTN

Protein characters

Codons

Compressed representation,
using the IUPAC ambiguous DNA character encoding we saw last time

Protein Alphabet

Amino acid	Codons	Compressed	Amino acid	Codons	Compressed
Ala/A	GCT, GCC, GCA, GCG	GCN	Leu/L	TTA, TTG, CTT, CTC, CTA, CTG	YTR, CTN
Arg/R	CGT, CGC, CGA, CGG, AGA, AGG	CGN, MGR	Lys/K	AAA, AAG	AAR
Asn/N	AAT, AAC	AAY	Met/M	ATG	
Asp/D	GAT, GAC	GAY	Phe/F	TTT, TTC	TTY
Cys/C	TGT, TGC	TGY	Pro/P	CCT, CCC, CCA, CCG	CCN
Gln/Q	CAA, CAG	CAR	Ser/S	TCT, TCC, TCA, TCG, AGT, AGC	TCN, AGY
Glu/E	GAA, GAG	GAR	Thr/T	ACT, ACC, ACA, ACG	ACN
Gly/G	GGT, GGC, GGA, GGG	GGN	Trp/W	TGG	
His/H	CAT, CAC	CAY	Tyr/Y	TAT, TAC	TAY
Ile/I	ATT, ATC, ATA	ATH	Val/V	GTT, GTC, GTA, GTG	GTN

This list contains only *61* out of *64* triplets.
Where are the remaining three?

Protein Alphabet

Amino acid	Codons	Compressed	Amino acid	Codons	Compressed
Ala/A	GCT, GCC, GCA, GCG	GCN	Leu/L	TTA, TTG, CTT, CTC, CTA, CTG	YTR, CTN
Arg/R	CGT, CGC, CGA, CGG, AGA, AGG	CGN, MGR	Lys/K	AAA, AAG	AAR
Asn/N	AAT, AAC	AAY	Met/M	ATG	
Asp/D	GAT, GAC	GAY	Phe/F	TTT, TTC	TTY
Cys/C	TGT, TGC	TGY	Pro/P	CCT, CCC, CCA, CCG	CCN
Gln/Q	CAA, CAG	CAR	Ser/S	TCT, TCC, TCA, TCG, AGT, AGC	TCN, AGY
Glu/E	GAA, GAG	GAR	Thr/T	ACT, ACC, ACA, ACG	ACN
Gly/G	GGT, GGC, GGA, GGG	GGN	Trp/W	TGG	
His/H	CAT, CAC	CAY	Tyr/Y	TAT, TAC	TAY
Ile/I	ATT, ATC, ATA	ATH	Val/V	GTT, GTC, GTA, GTG	GTN

Note that, mainly the **third** Codon position differs
→ it is less vulnerable to mutations than the *1st* and *2nd* codon positions

Protein Evolution

- This redundancy plays a role in protein evolution
- We distinguish between

  1) *Synonymous* substitutions/mutations
  (*GCC* → *GCT* ≡ Alanine → Alanine)

  versus

  2) *Non-synonymous* substitutions/mutations
  (*GGT* → *GTT* ≡ Glycine → Valine)

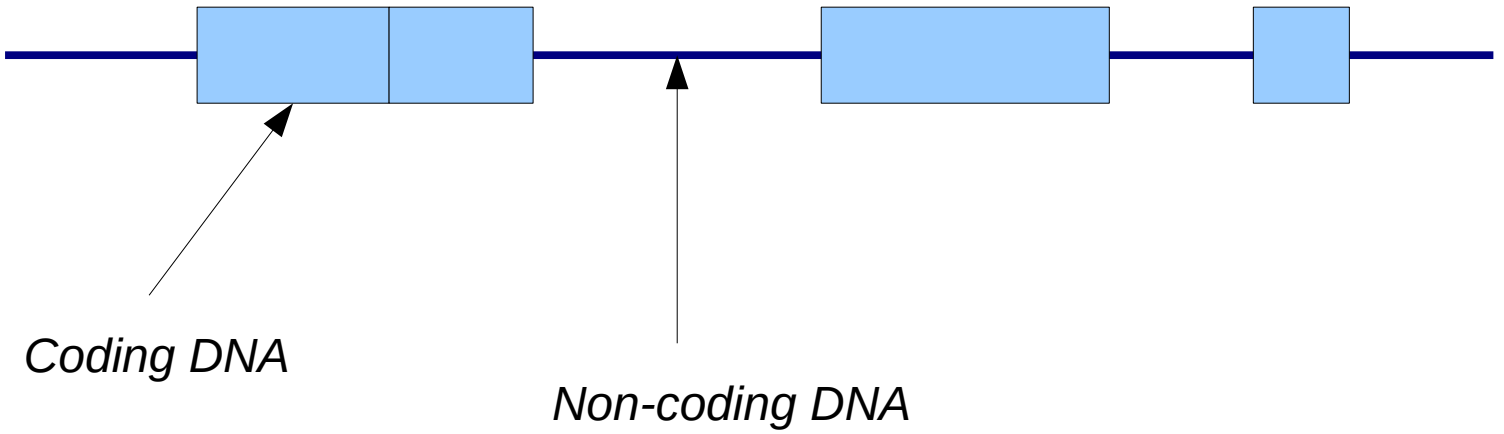
Translating DNA ↔ Protein data

- DNA → Protein: not ambiguous, but redundant
- Protein → DNA: ambiguous, several DNA triplets can encode for the same Amino Acid
- In bioinformatics we sometimes directly use the Codons (triplets) instead of amino acids to utilize all information available!
- See for instance *Codon evolution models*

  → *http://www.inf.ethz.ch/personal/anmaria/papers/Chapter%202.pdf*

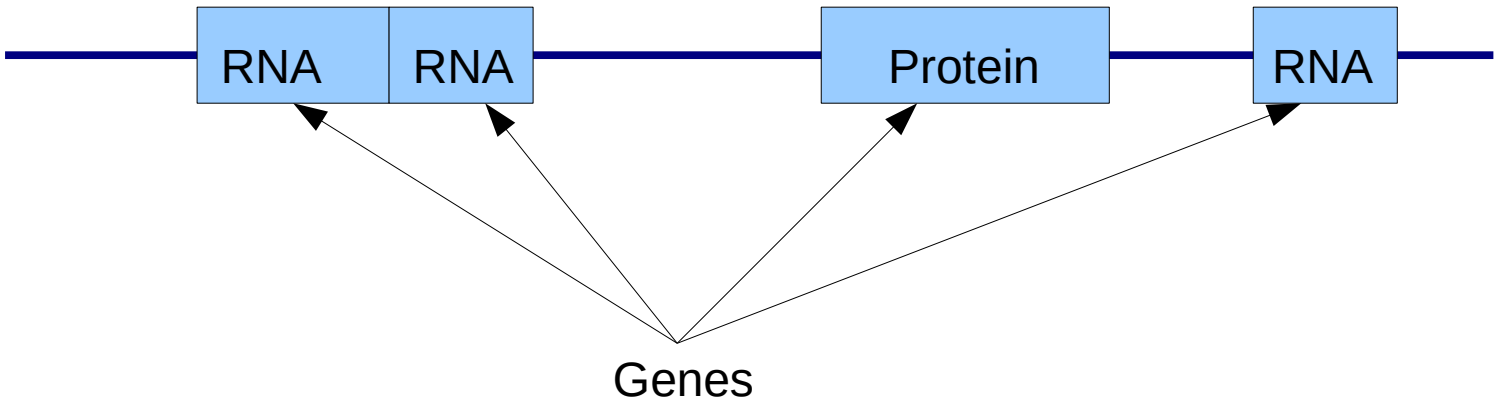
Top-level view

Chromosome: a long DNA molecule



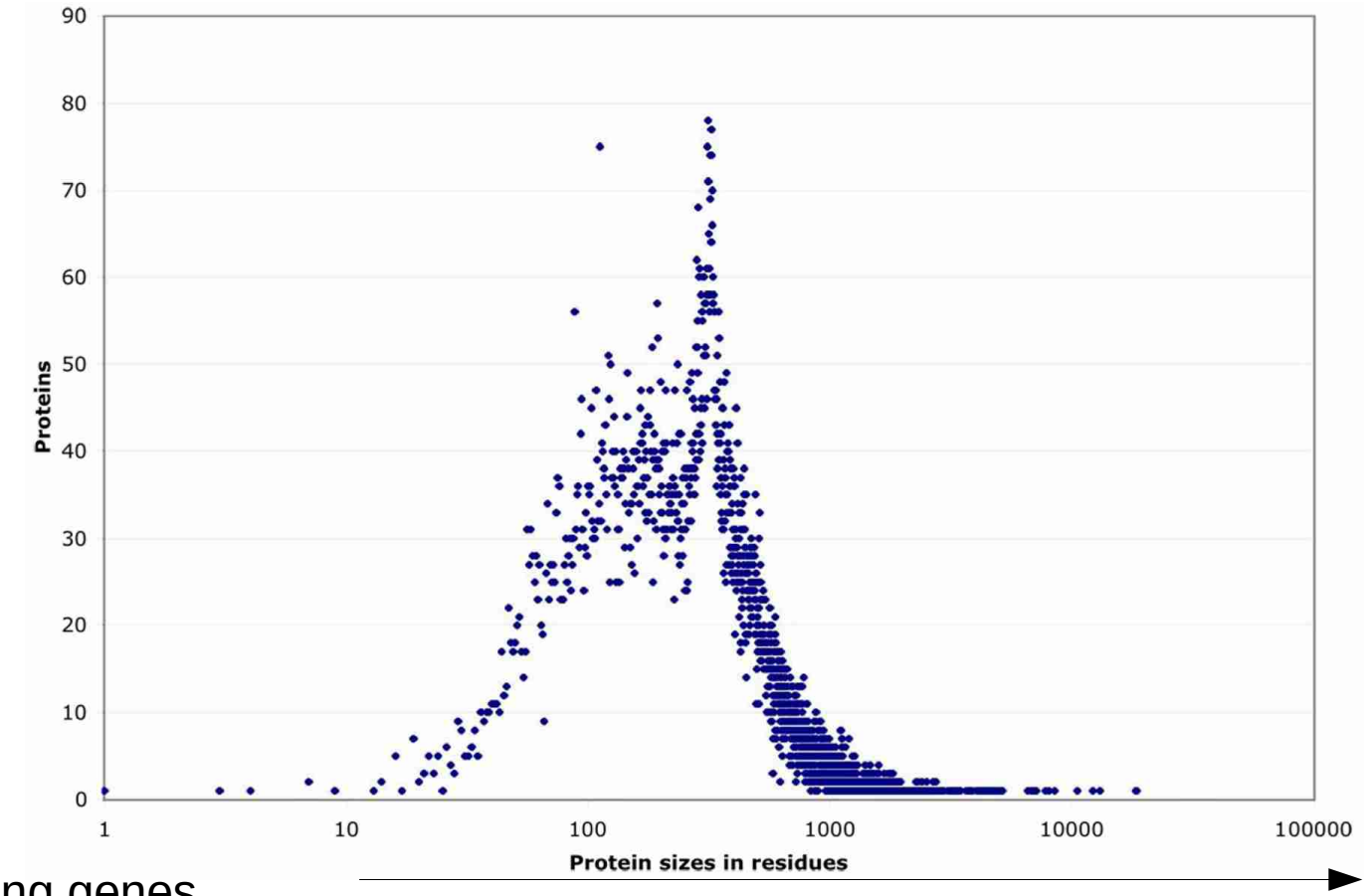
Top-level view

Chromosome: a long DNA molecule



Gene lengths vary: a typical gene is ≈1000 bp long

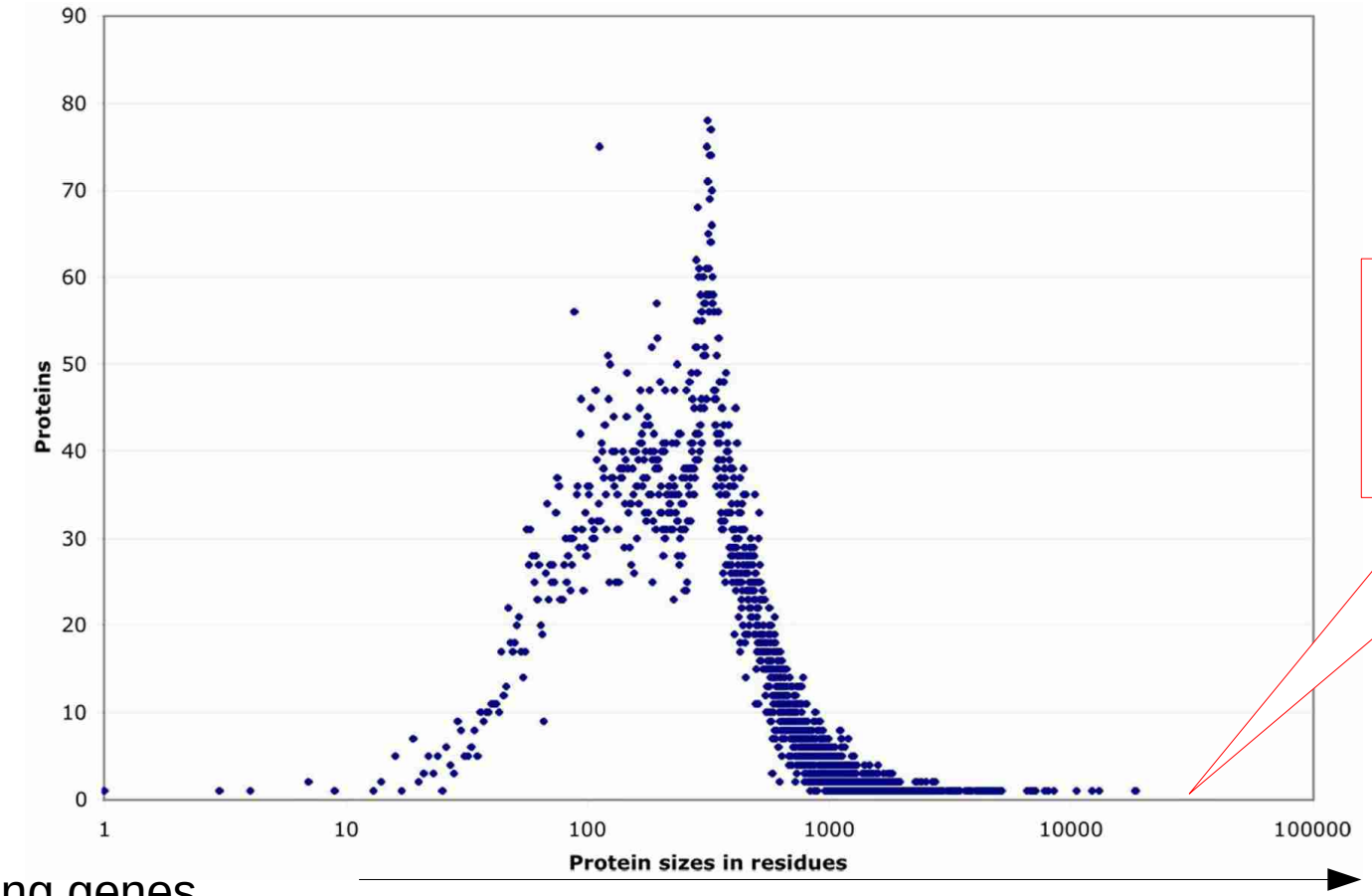
Average Protein gene Lengths



Number of Protein-coding genes

Protein sequence length → this is counted in # amino acid characters, not nucleotides, multiply by three to obtain DNA length!

Average Protein gene Lengths



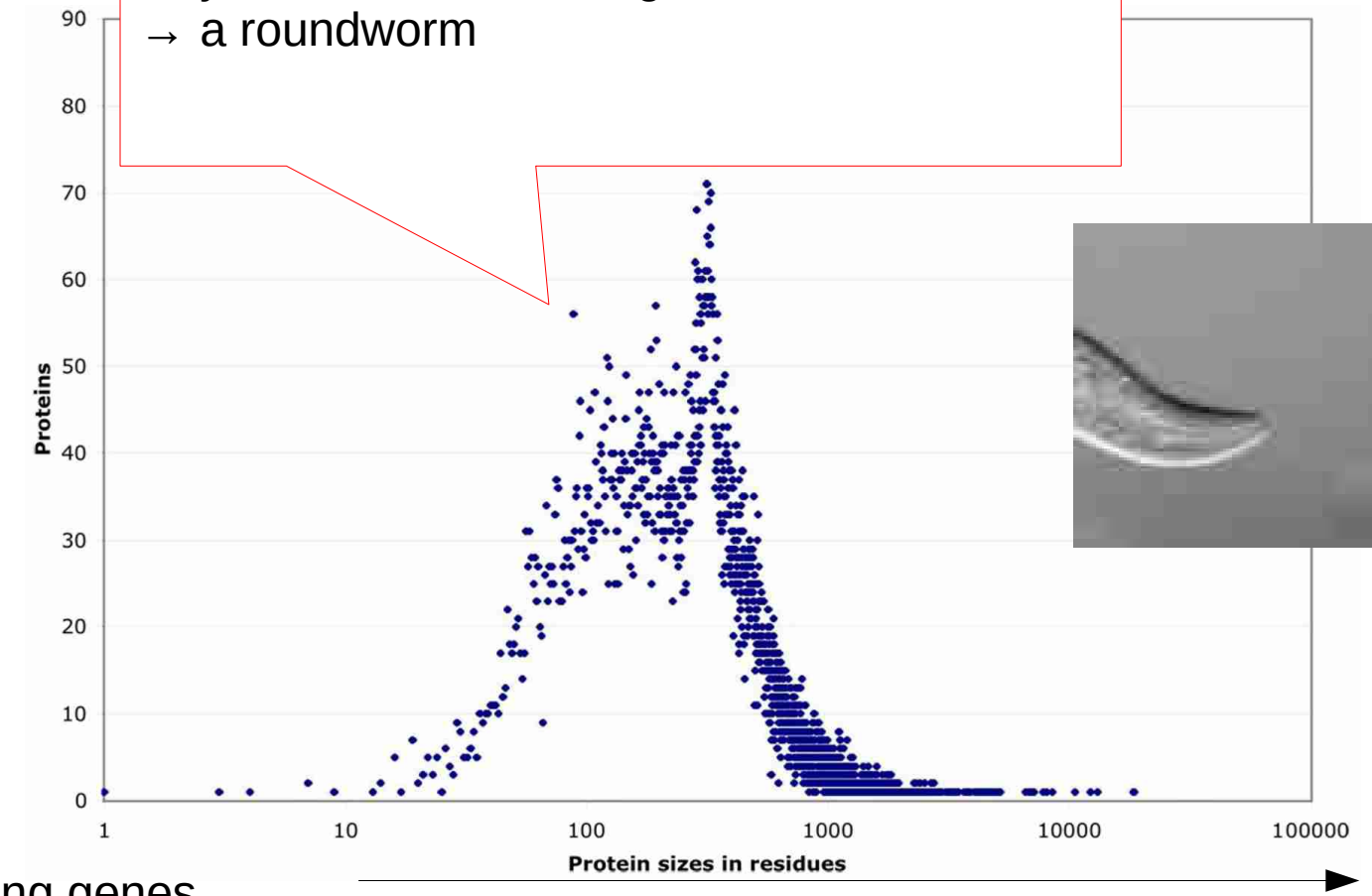
Logarithmic scale!

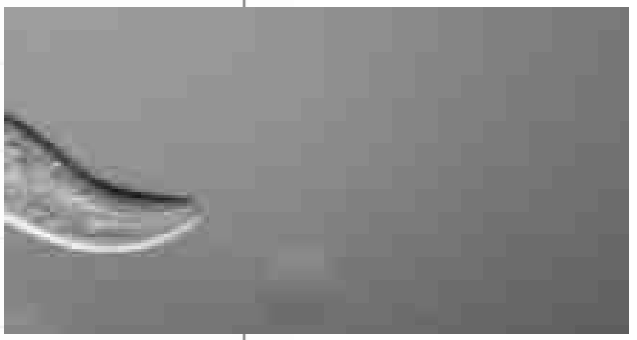
Number of Protein-coding genes

Protein sequence length → this is counted in # amino acid characters, not nucleotides, multiply by three to obtain DNA length!

Average Protein gene Lengths

Data for Caenorhabditis Elegans (C. Elegans)
→ yet another model organism
→ a roundworm

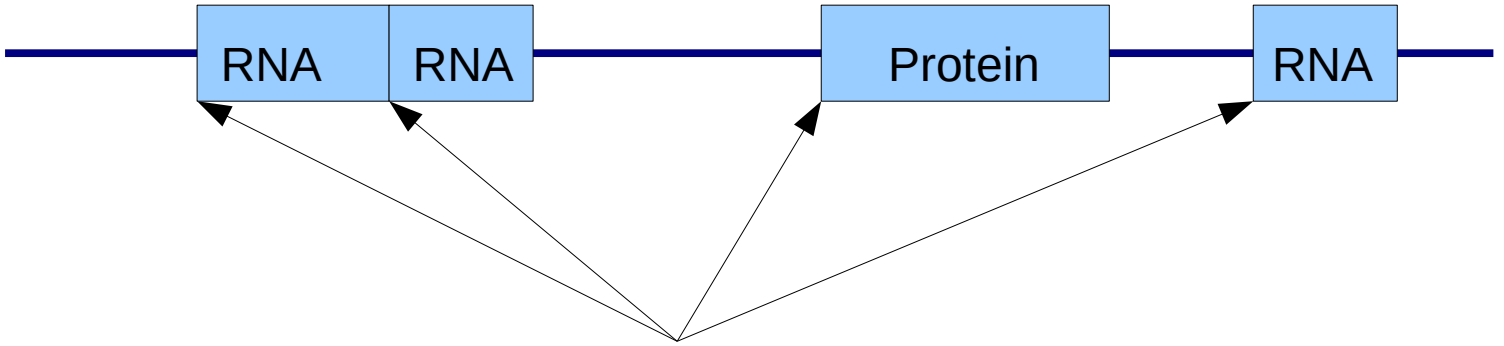




Number of
Protein-coding genes

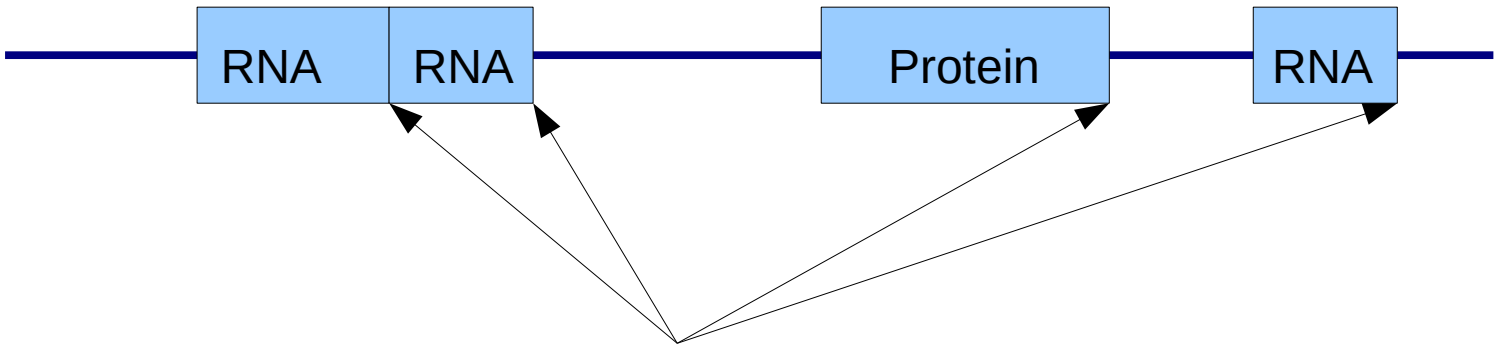
Protein sequence length → this is counted in # amino acid characters, not nucleotides, multiply by three to obtain DNA length!

Top-level view



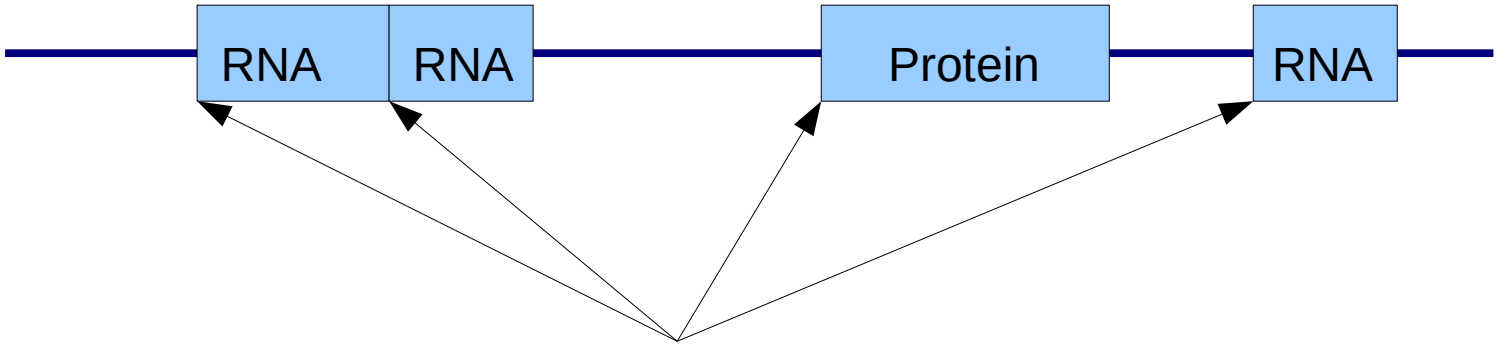
How do we know where genes start?

Top-level view



How do we know where genes end?

Top-level view



Gene boundaries:
→ special *START/STOP Codons* (DNA triplets)

All Codons

Amino acid	Codons	Compressed
Ala/A	GCT, GCC, GCA, GCG	GCN
Arg/R	CGT, CGC, CGA, CGG, AGA, AGG	CGN, MGR
Asn/N	AAT, AAC	AAY
Asp/D	GAT, GAC	GAY
Cys/C	TGT, TGC	TGY
Gln/Q	CAA, CAG	CAR
Glu/E	GAA, GAG	GAR
Gly/G	GGT, GGC, GGA, GGG	GGN
His/H	CAT, CAC	CAY
Ile/I	ATT, ATC, ATA	ATH
START	ATG	

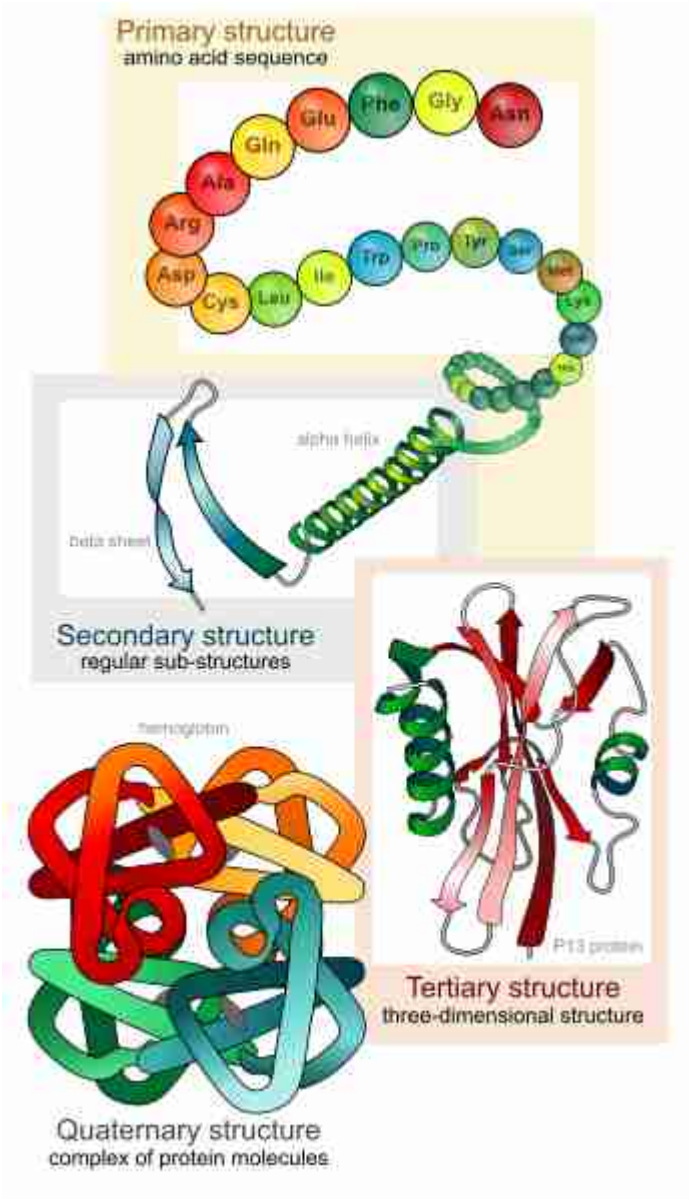
Amino acid	Codons	Compressed
Leu/L	TTA, TTG, CTT, CTC, CTA, CTG	YTR, CTN
Lys/K	AAA, AAG	AAR
Met/M	ATG	
Phe/F	TTT, TTC	TTY
Pro/P	CCT, CCC, CCA, CCG	CCN
Ser/S	TCT, TCC, TCA, TCG, AGT, AGC	TCN, AGY
Thr/T	ACT, ACC, ACA, ACG	ACN
Trp/W	TGG	
Tyr/Y	TAT, TAC	TAY
Val/V	GTT, GTC, GTA, GTG	GTN
STOP	TAA, TGA, TAG	TAR, TRA

Now we have all *64* combinations 

Proteins

- What do they do?
- Structural proteins → tissue building blocks
- Enzymatic proteins → catalysts (steering/accelerating) of specific biochemical reactions in the body
- Examples:
  - oxygen transport
  - immune defense
  - provide & store energy
- Because there are many such processes we need many proteins
- Homo sapiens ≈ *20,000* proteins → number disputed
- Again: a protein is a sequence/string of amino acid characters
- Terminology: Instead of counting nucleotides/base pairs we count protein letters as *residues*
- Example: the protein string AEFFQQP has *7* residues

Protein Structure



Role of Structure

- A protein does not only consist of a string of residues (called *primary structure*)
- A protein sequence also has:

  1) Secondary
  2) Tertiary
  3) Quaternary

  structure!
- The structure determines the function/effect of a protein
- One would like to predict the structure from the protein sequence (primary structure)
- Still a challenging problem
- We will not deal with this in our course though!

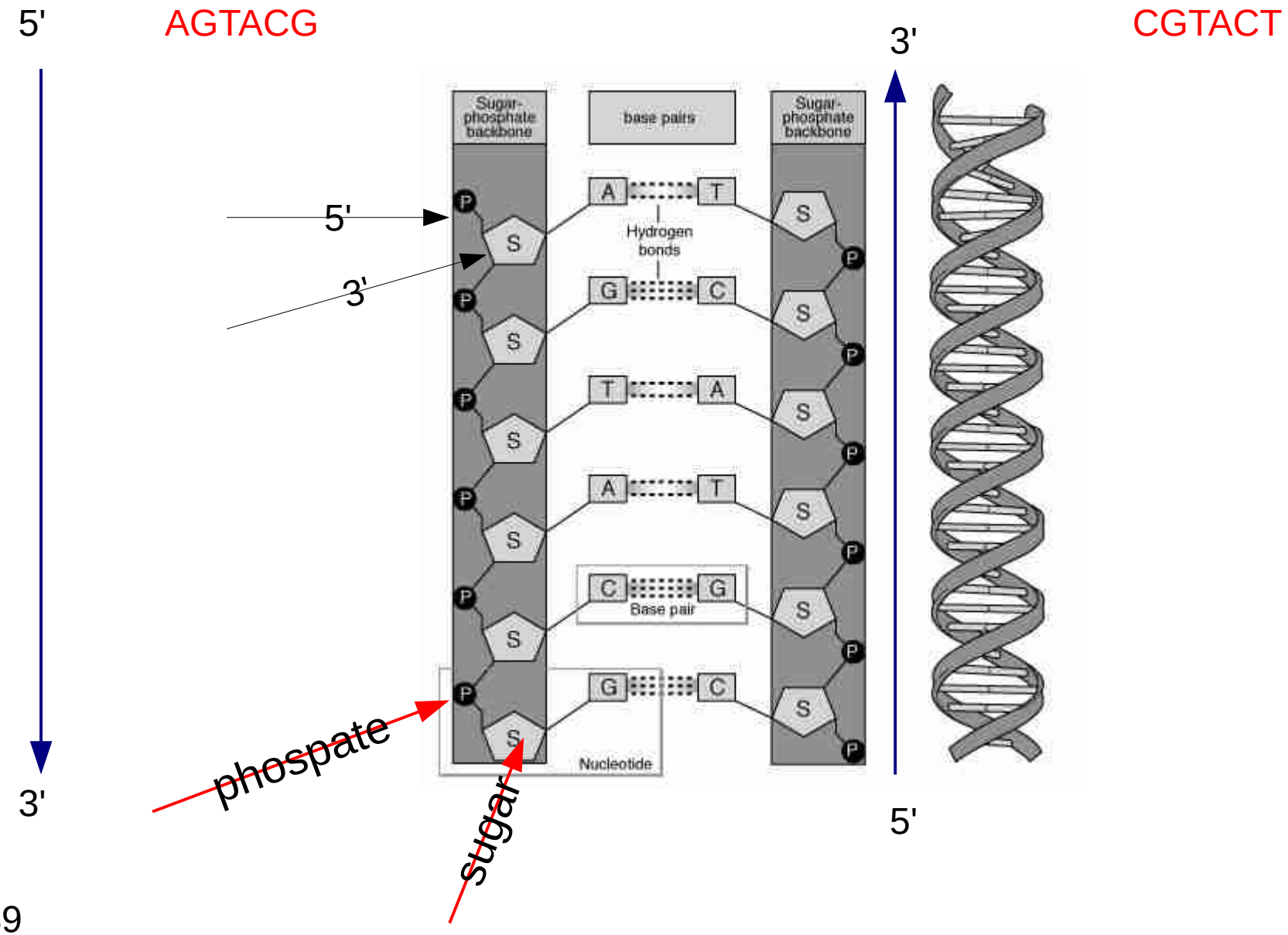
Protein Structure Prediction

- Some protein structures are known → *Crystallography*
- Test prediction programs on these
- Contest: The Critical Assessment of protein Structure Prediction www.predictioncenter.org
- Blind testing and benchmarking of programs

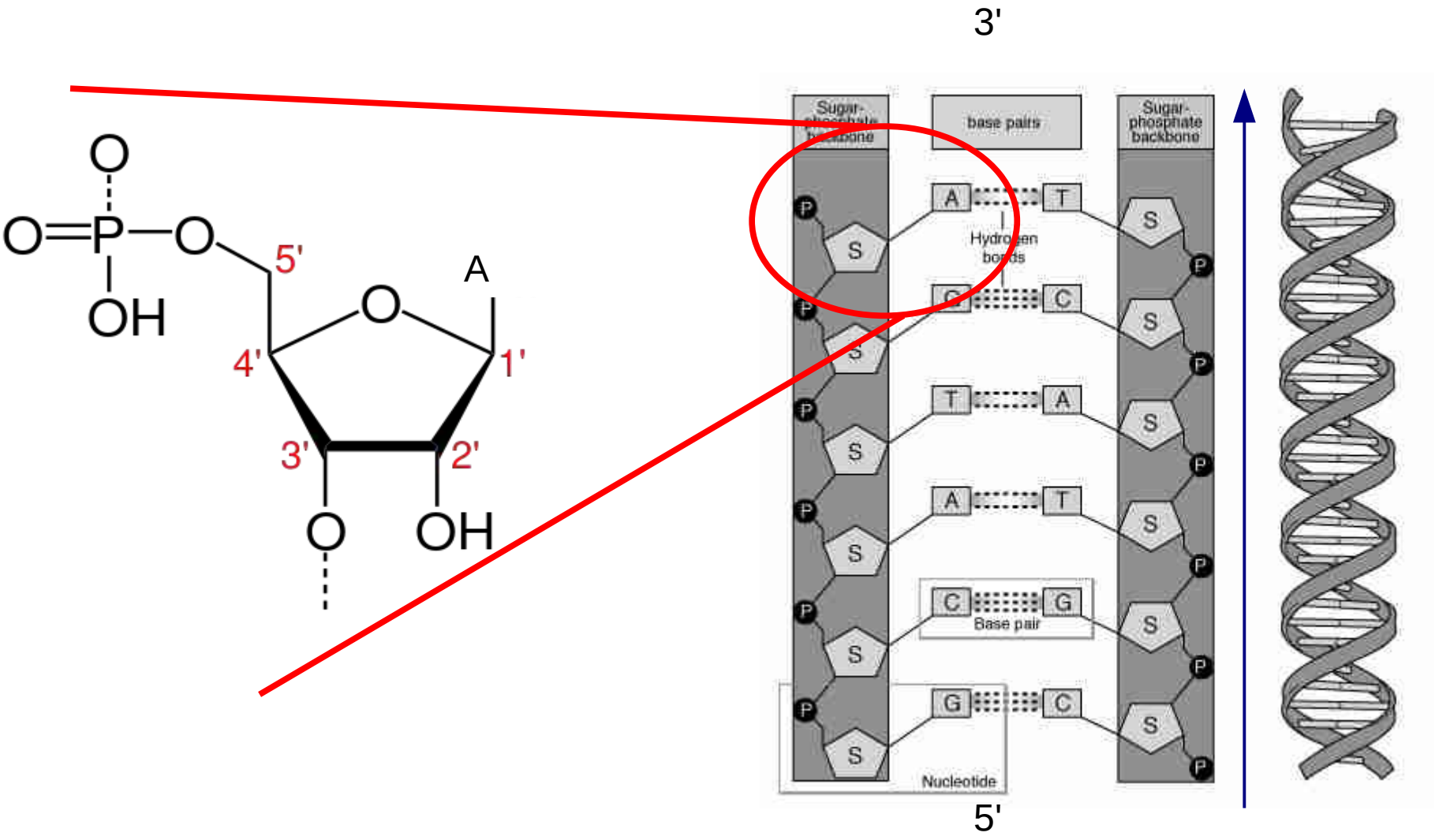
Another challenging problem

- Can we predict the function of a gene and/or protein, based on its sequence?
- Generally known as *gene function prediction*
- We will also omit this topic though

3' and 5'


5'
AGTACG
3'
CGTACT
Sugar-phosphate backbone
base pairs
Sugar-phosphate backbone
5'
3'
Hydrogen bonds
Base pair
Nucleotide
phospate
sugar
3'
5'

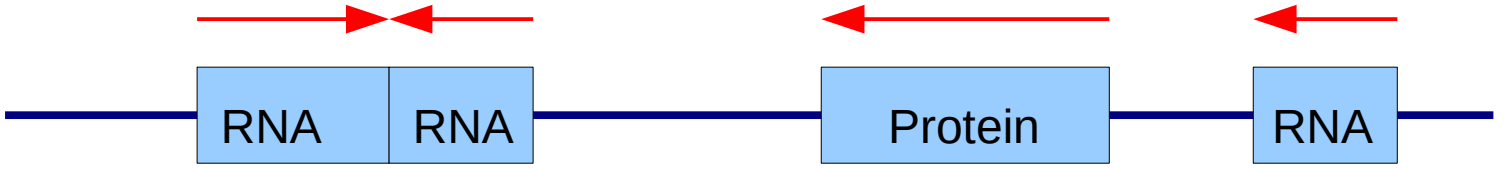
3' and 5'


3'
O
O=P—O
OH
5'
A
O
4'
1'
3'
2'
O
OH
Sugar-phosphate backbone
base pairs
Sugar-phosphate backbone
A
T
Hydrogen bonds
G
C
T
A
A
T
C
G
Base pair
G
C
Nucleotide
5'

Back to DNA again

- DNA comes in a double helix
- A single string of DNA without the complement is also called DNA strand
- The bases *A, C, G, T* are connected via a backbone molecule consisting of *5* carbon atoms labelled *1', 2',...,5'*
- Backbone connections via the *3'* and *5'* units
- Every DNA strand has a direction
- By convention we write DNA sequences in the direction from *5'* → *3'*

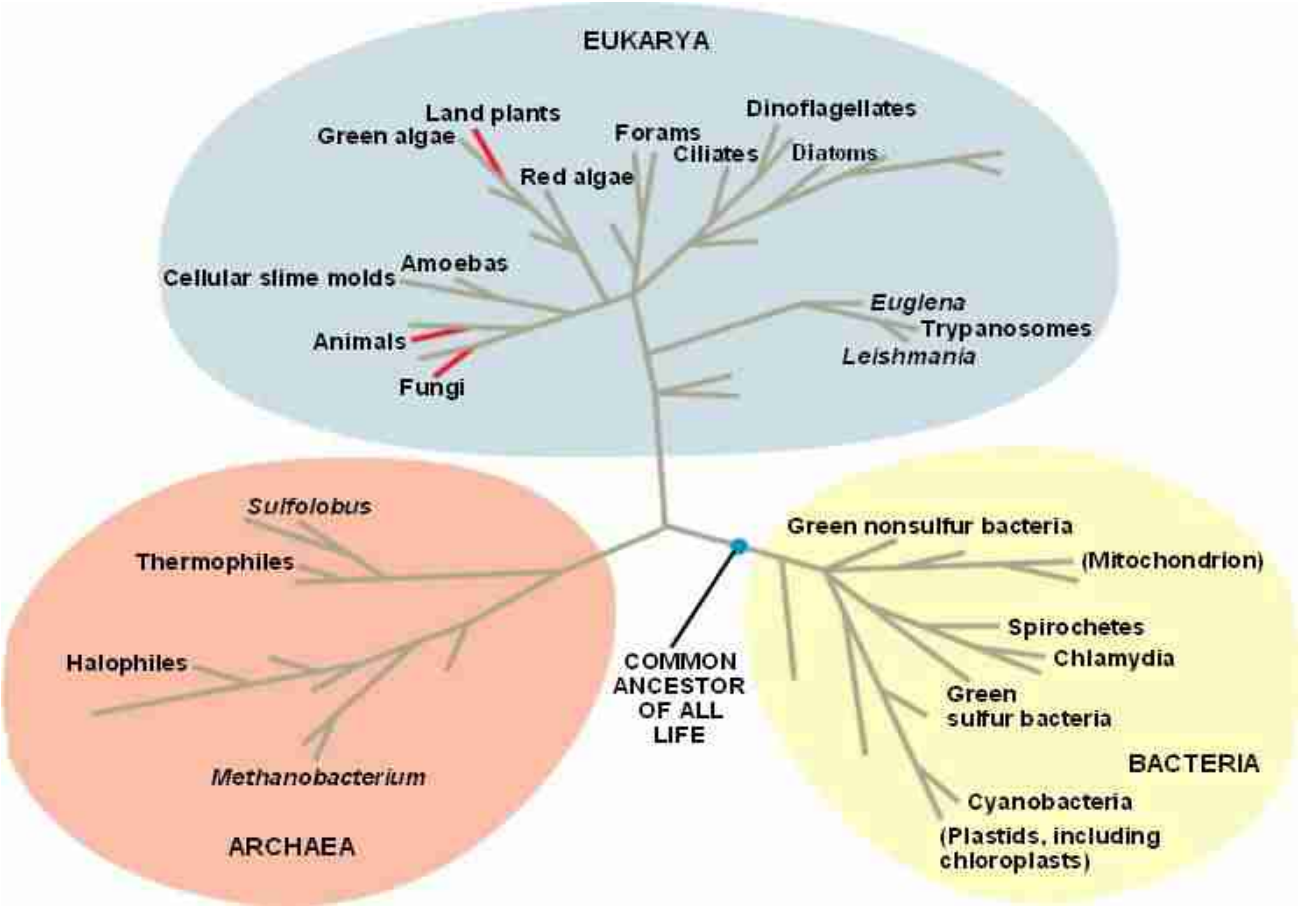
Top-level view



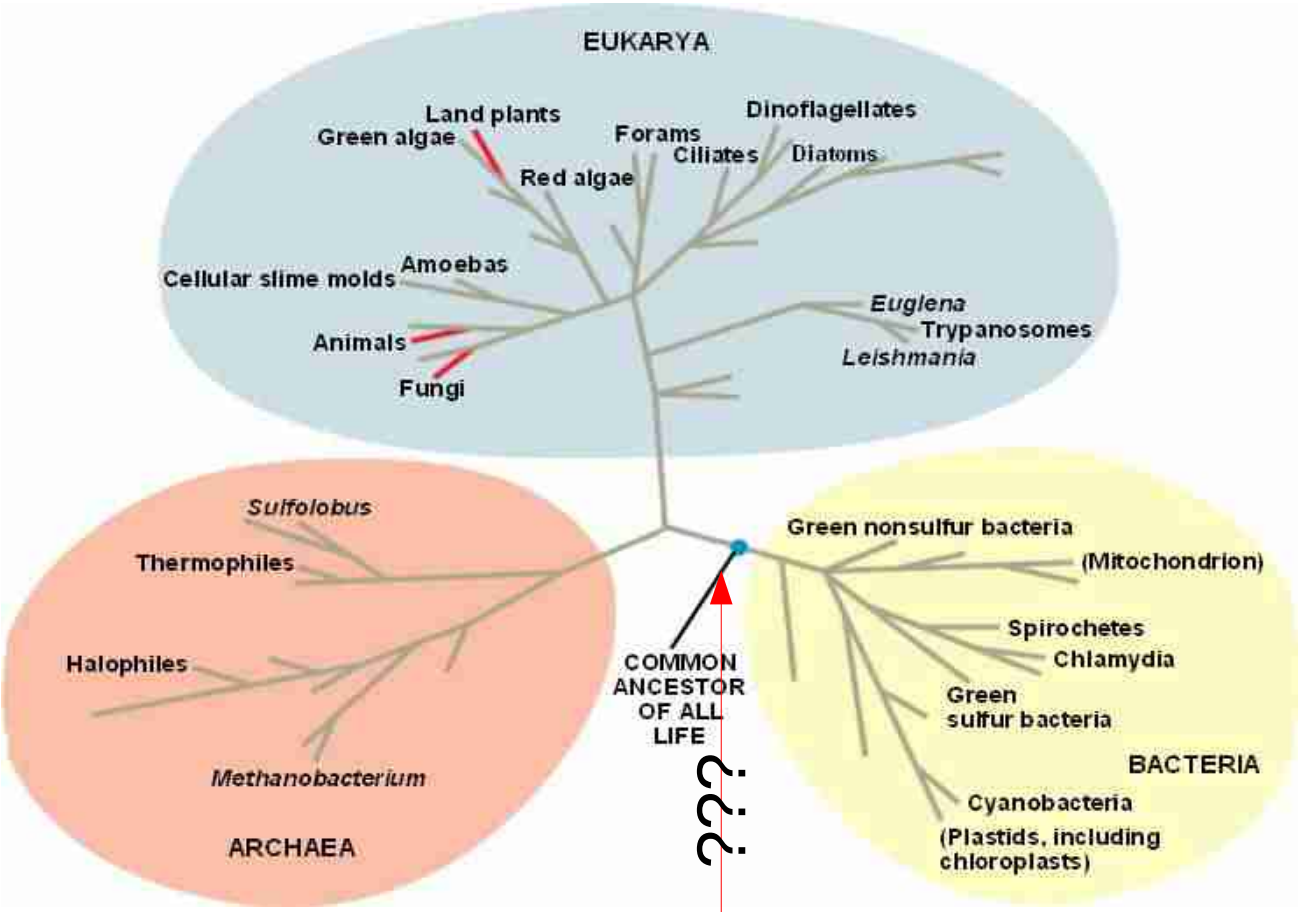
→ Genes have a direction!
→ depending on which strand of the double helix encodes the gene
They must be read from the correct side to be recognized!

The domains of life

Classic paper: Woese C, Kandler O, Wheelis M (1990).
"Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.".
Proc Natl Acad Sci USA 87(12): 4576–9



The domains of life



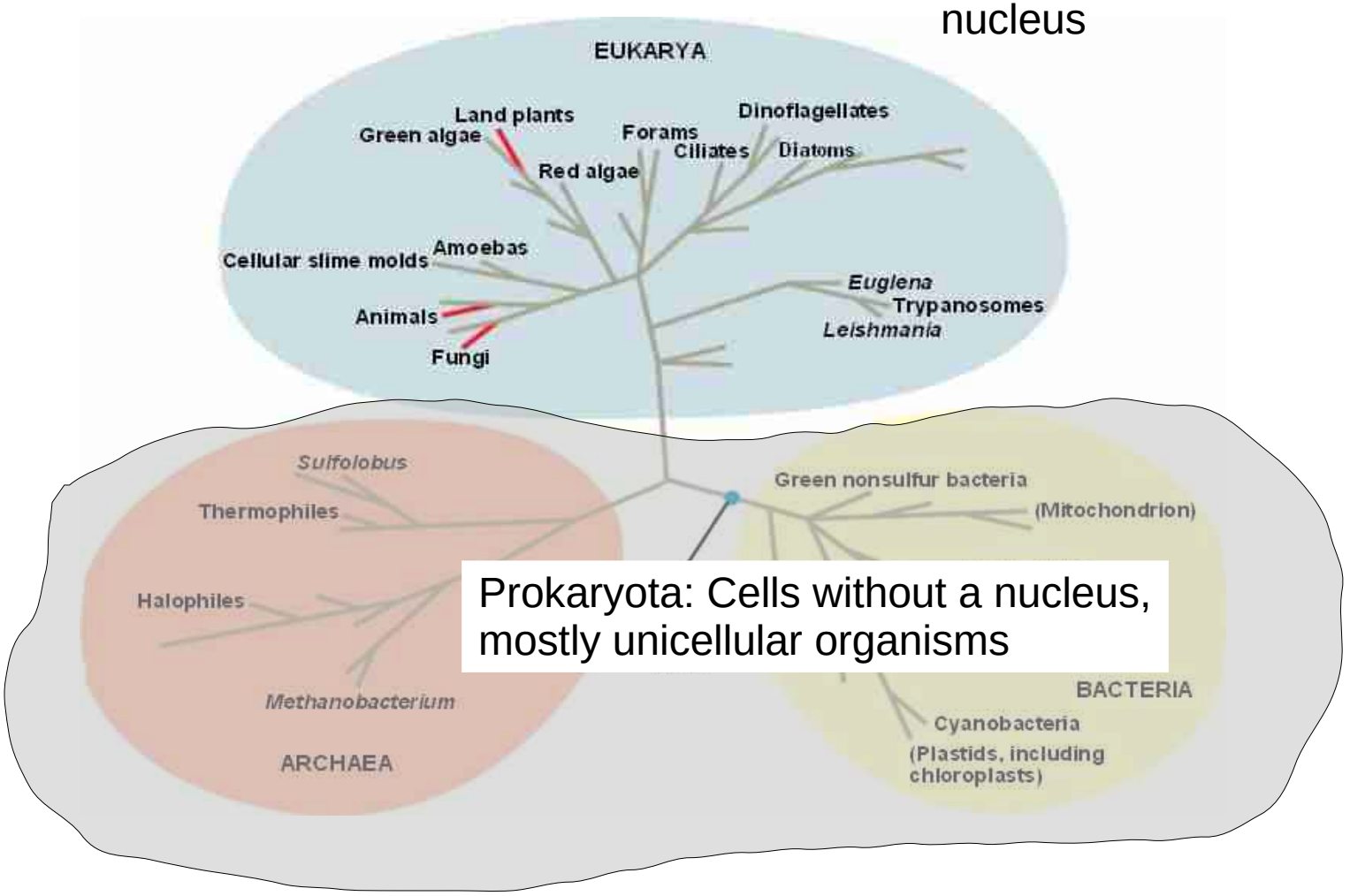
Salty environments

Hot environments

Where is the common ancestor?

The domains of life

Eukaryota: organisms with a cell nucleus



Prokaryota: Cells without a nucleus, mostly unicellular organisms

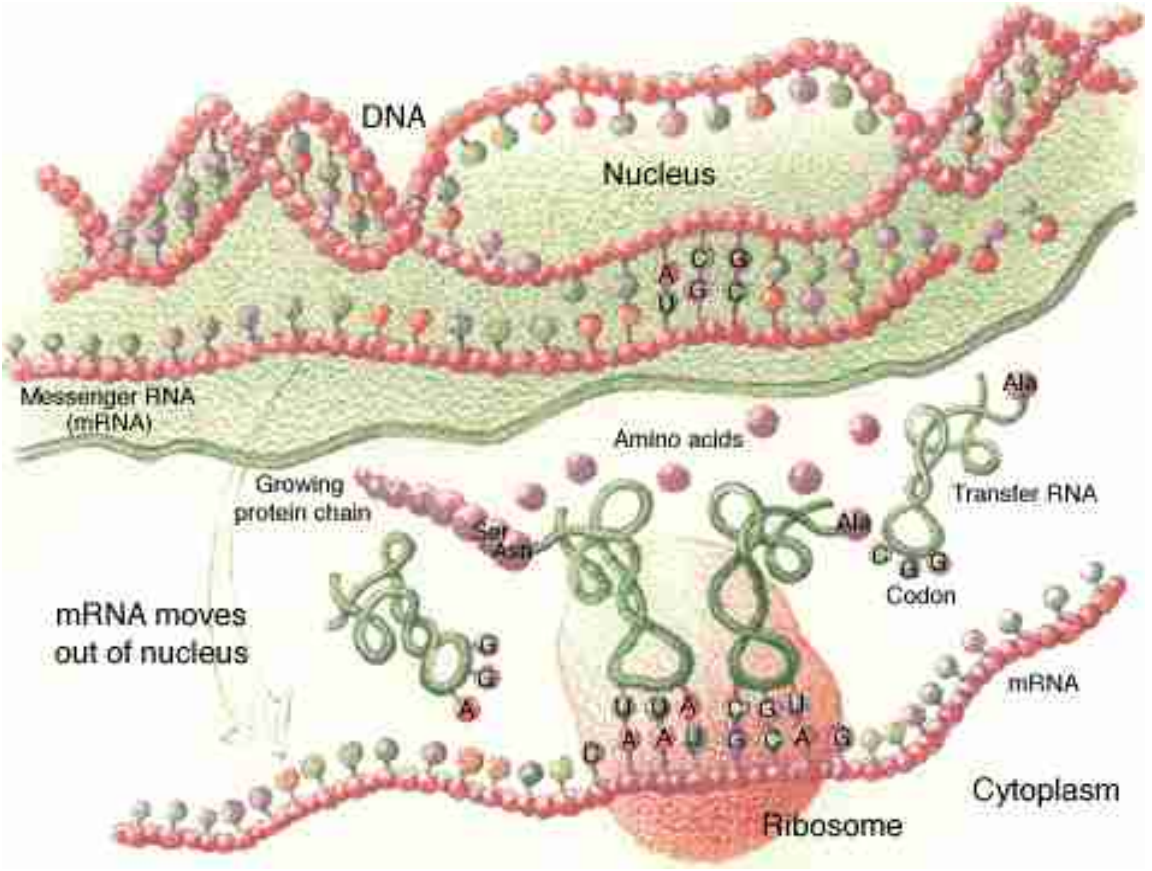
More about genes

- *Prokaryot{es|a}*: A gene encodes a protein or an RNA
- *Eukaryot{es|a}*: it's more complicated
  - Not the entire gene sequence may encode for a protein, just parts of it
  - Within an eukaryotic gene we distinguish between
    - *Introns* → not used in protein synthesis
    - *Exons* → parts of the gene used for protein synthesis

What does RNA do?

- As we already know RNA is similar to DNA
- There are some chemical differences
- RNA does not form a double-stranded helix
- DNA stores information
- Like proteins, RNA performs different functions in the cell
- An analogy:
  - DNA is something like the hard disk
  - RNA and proteins are processing elements

An overview


DNA
Nucleus
Messenger RNA
(mRNA)
Amino acids
Growing
protein chain
Ser
Asn
Ala
Transfer RNA
Codon
mRNA moves
out of nucleus
mRNA
Cytoplasm
Ribosome

RNA

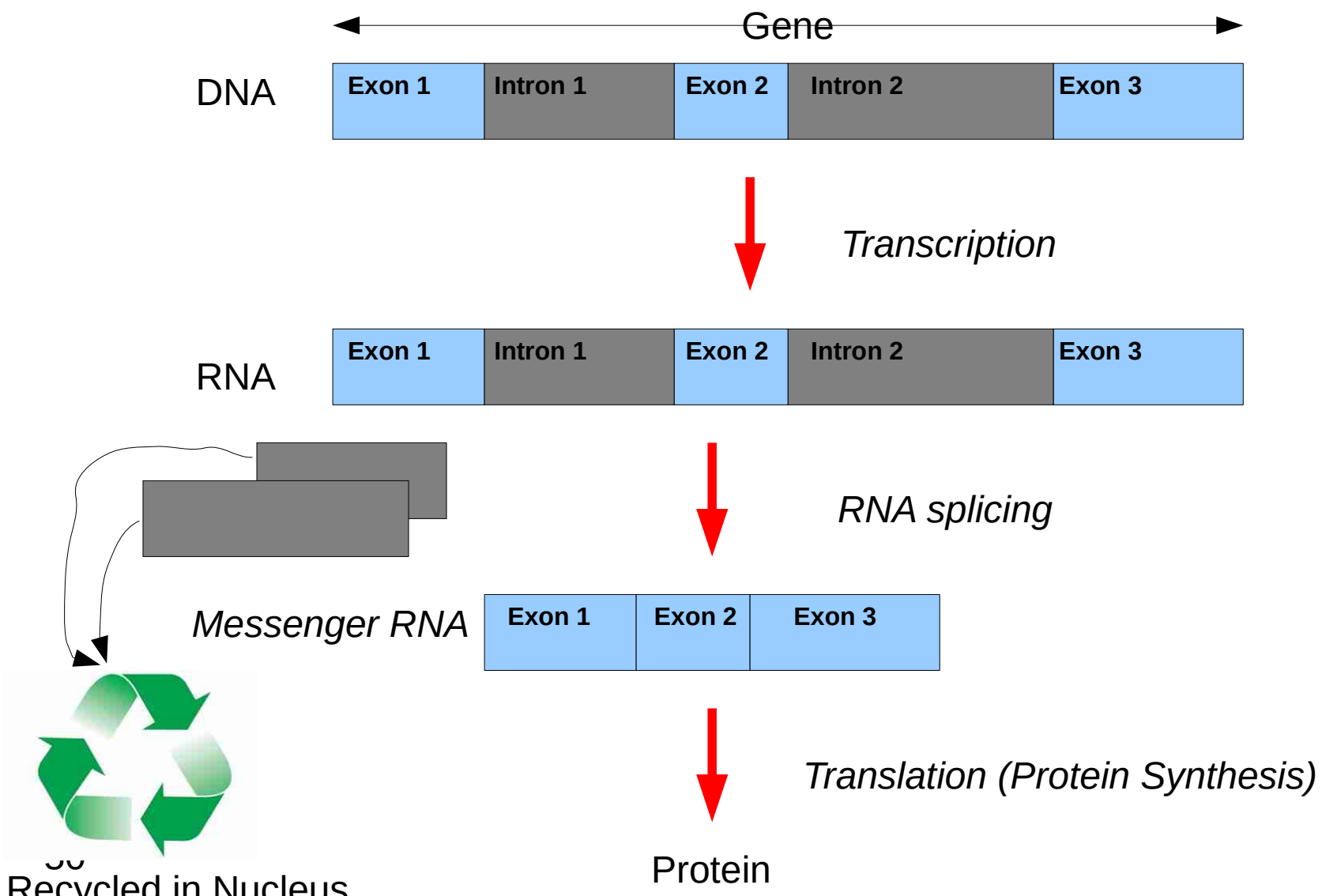
- RNA is involved in the process of DNA *Transcription*
- RNA is a copy of a coding DNA strand (a gene)
- And involved in the process of Transcription to construct either:

  1) A protein: DNA → RNA → Protein

     This is called translation (coding RNA)

  2) A non-coding RNA: DNA → RNA that has some other direct function in the cell

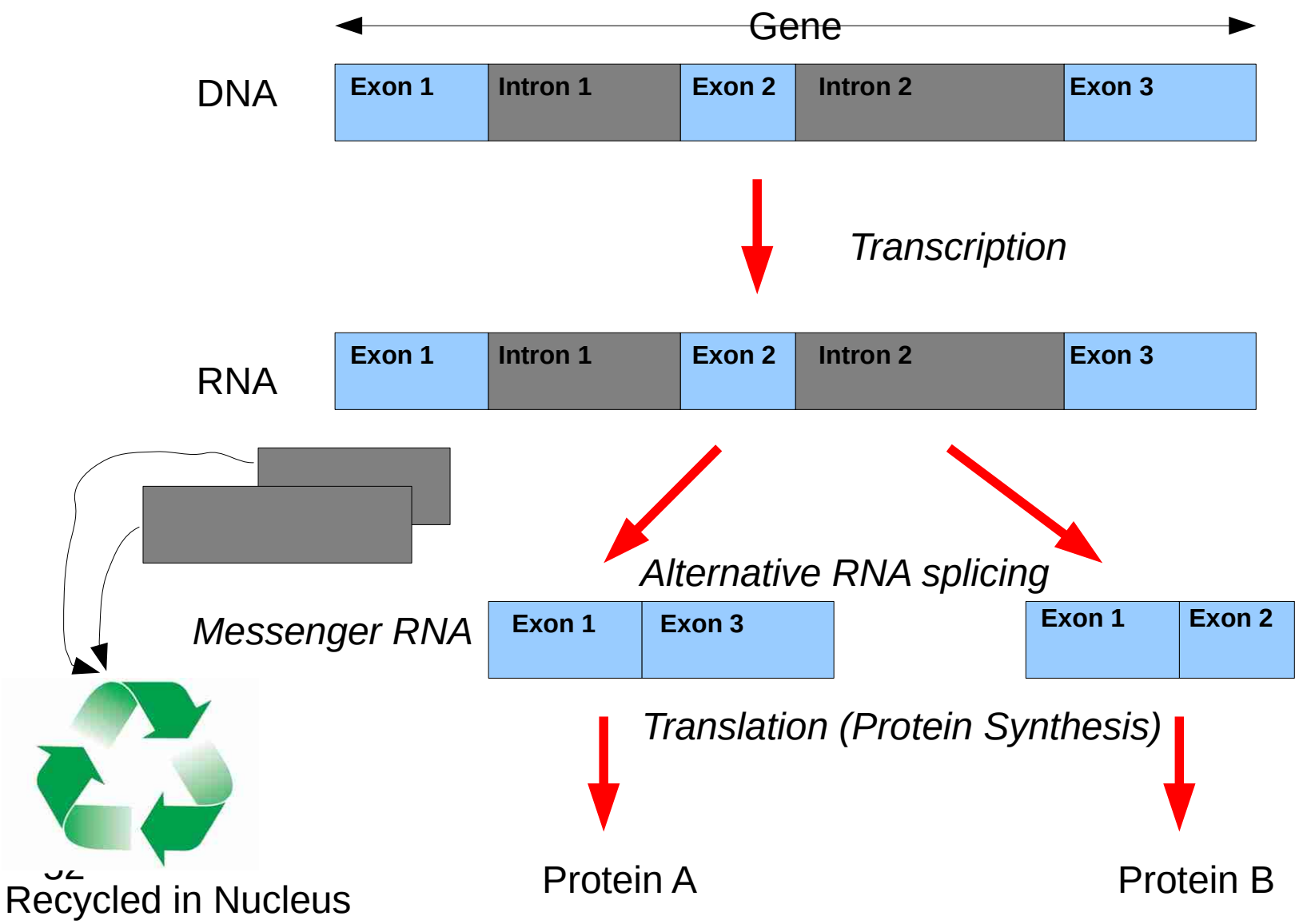
RNA Splicing *Eukaryota*



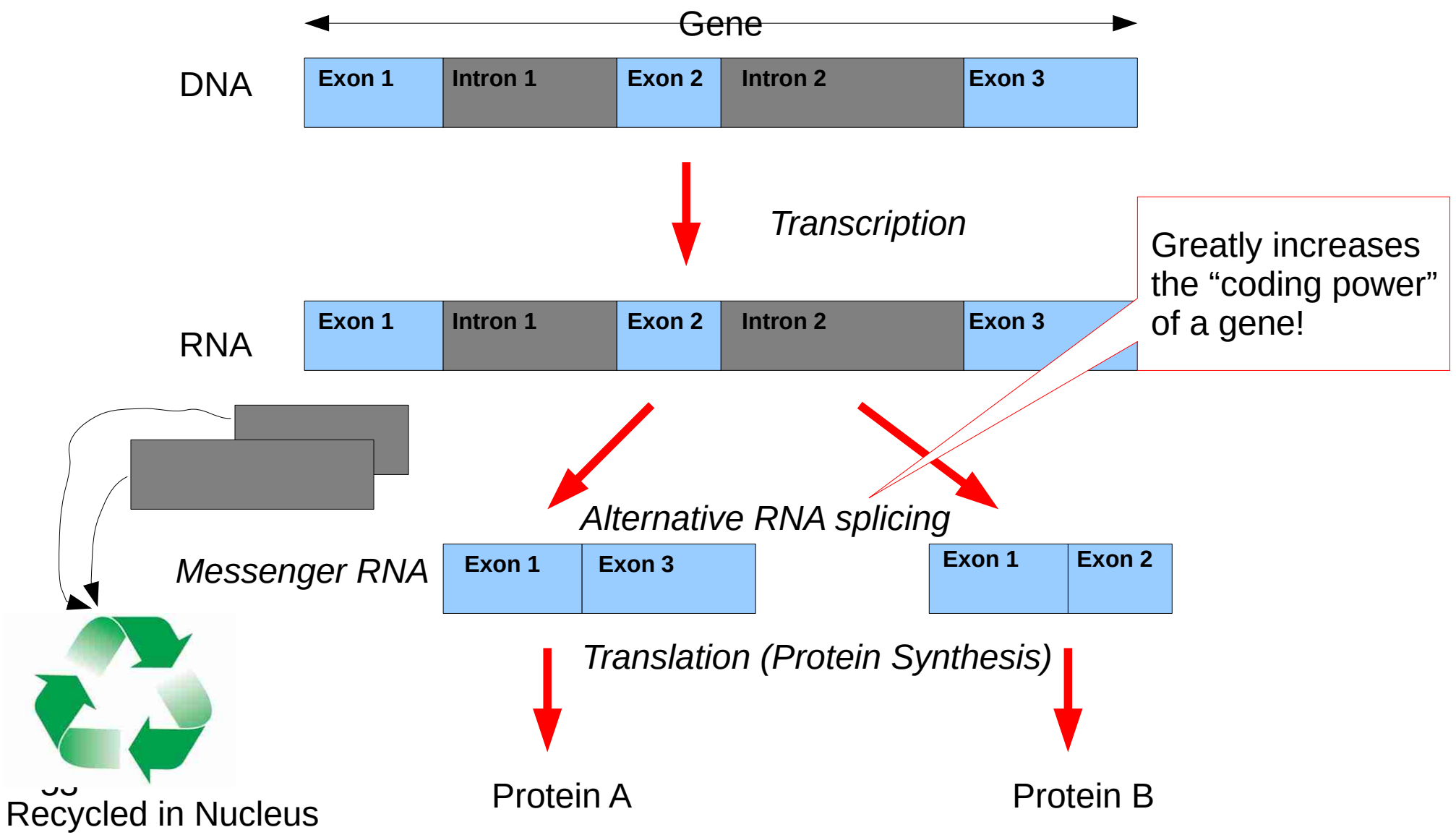
Eukaryotic RNA

- Remember: Not the entire gene sequence may be transcribed/used
- *Introns* → not used
- *Exons* → used
- Introns are spliced out (“ausgestossen”) from the RNA strand (corresponding to the full gene), after transcription

Alternative Splicing



Alternative Splicing



Types of RNA

- *mRNA*: messenger RNA

  → transports RNA data to the ribosome for protein synthesis

- *rRNA*: ribosomal RNA

  → carries out the translation in the ribosome via catalysis

- *tRNA*: transfer RNA

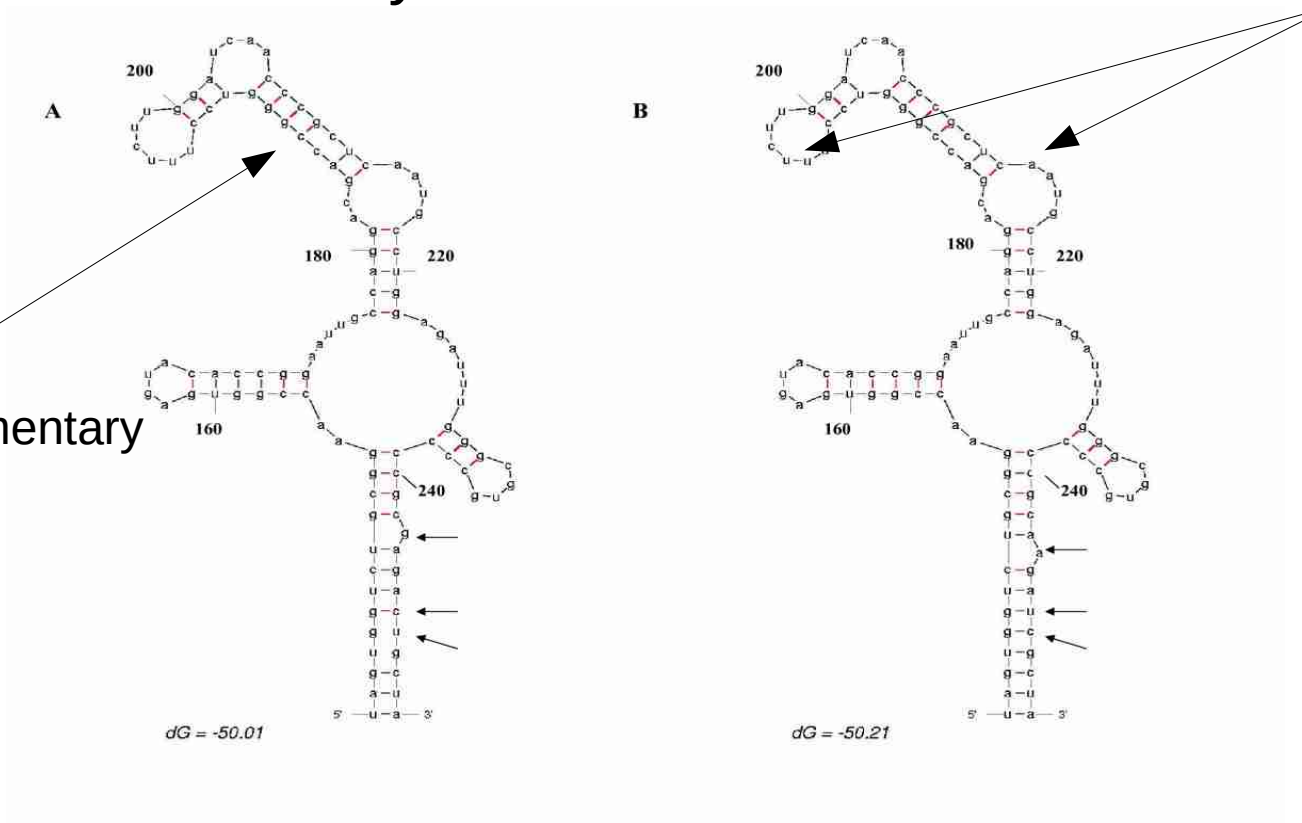
  → brings in the amino acids

The importance of ribosomal RNA

- Different species do not have the same set of genes
- Only few genes are common to *all* species
- The *rRNA* is such a gene
- The most well-known gene is the *16S* gene
- Therefore, it can be used to infer evolutionary relationships among **all** species

RNA Secondary Structure

- RNA is a single-stranded sequence!
- Secondary structure has an influence on the function of the molecule
- There is also a tertiary structure!



Loops: no Matching bases

Stem:complimentary Bases bind

A ↔ U

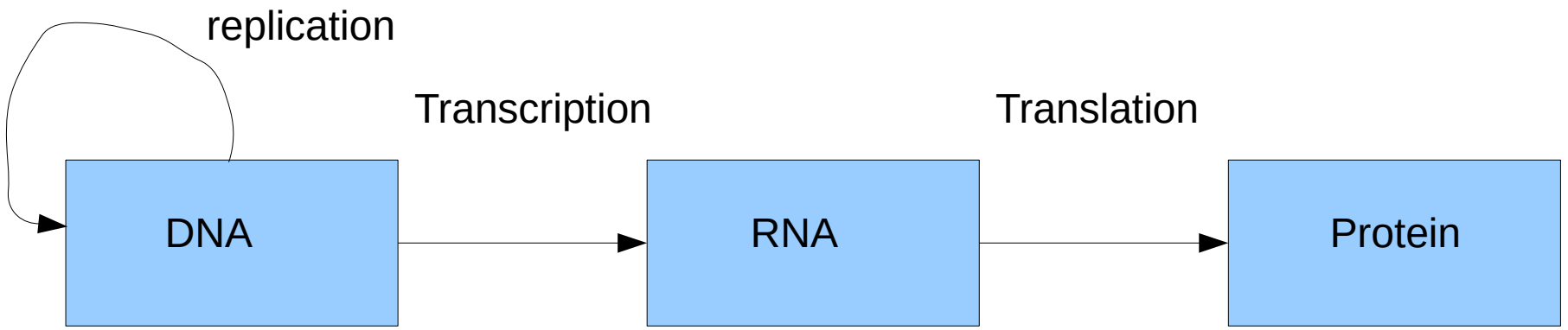
C ↔ G

RNA Secondary Structure

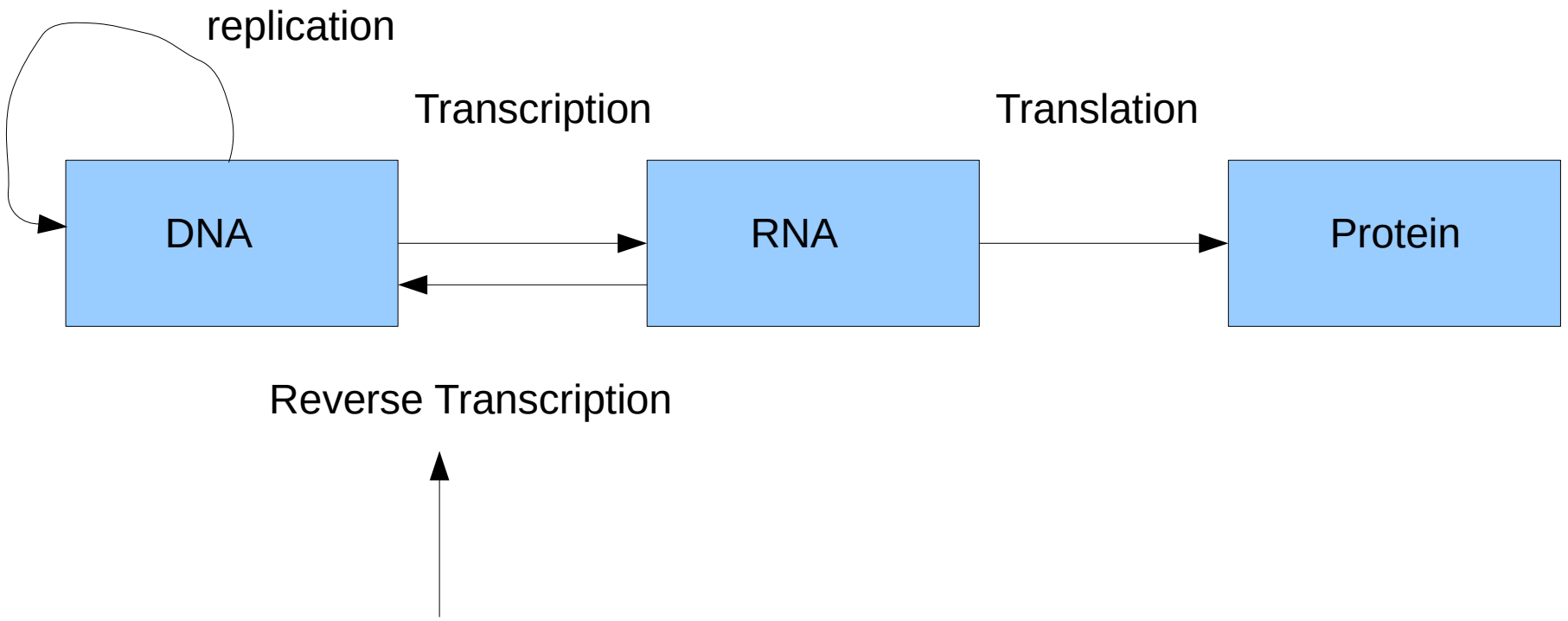
- Importance for RNA evolution

  → matching bases in a stem can not mutate independently from each other
- Research on predicting secondary structure from a plain RNA sequence

Central Dogma of Molecular Biology



Central Dogma of Molecular Biology



Serves some functions mainly in Viruses
1975 Nobel prize

What is a *Transcriptome*?

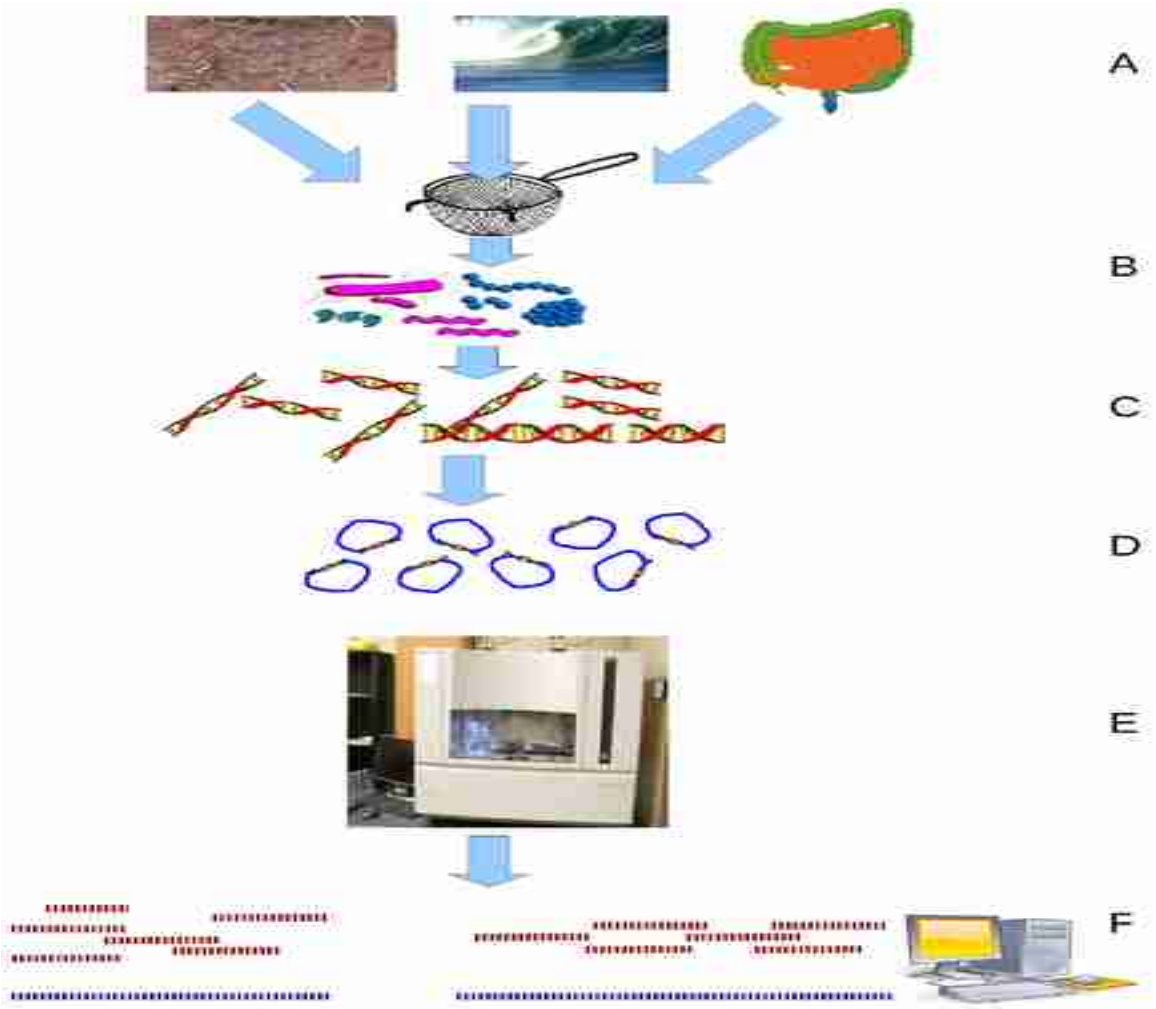
- The set/entirety of all RNA (mRNA, tRNA, rRNA) molecules in a cell
- In contrast to a genome, the transcriptome reflects the activity in a cell!

  → the interesting stuff is going on in there!
- Note the temporal and spatial component
  - Depending on the point of time and specialization/location of the cell, the transcriptome may be different

    → different genes are active in those specialized cells

    → sample from different cells
- *1600* insect transcriptome project 1KITE www.1kite.org

What is a *Meta-Genome*?



The Meta-Genome

- Example: Blind sequencing of all genetic material of a bacterial community → many species
- Figure out what the microbial diversity is
- Current hot topic!
- Can be done at:
  - Whole-genome level → metagenomics
  - Gene level, target specific gene → metagenetics
    - e.g., 16S RNA for Bacteria

Chromosome

- All *Chromosomes,* put together, form the *genome*
- # of chromosomes varies across species!
  - Human: 46
  - Mouse: 40
  - Donkey: 62
- Prokaryotes (simple organisms)

  → one chromosome
- Eukaryotes

  → many chromosomes

  → they are organized in pairs (paternal/maternal)

Eukaryotic Chromosomes

- Paired chromosomes are called homologous
- Some genes in homologous (parental/maternal) chromosomes are exactly identical
- … some are not → they have different genotypes!
- The genes that appear in different forms are called *Alleles*
- Cells containing pairs of chromosomes are called *diploid*
- Cells containing only one chromosome of each pair are called *haploid* → sexual reproduction

What's a species?

- Tricky question
- Different definitions
  - → generally debated
  - → more than *30* definitions exist
- By reproduction
  - → two species that can reproduce
  - → what about bacteria/viruses ????
- Evolutionary species concept
  - → via ancestral descent in an evolutionary tree
- General lineage (Abstammung/Verzweigung) concept
  - → an independently evolving lineage
- Phylogenetic Species Concept
  - → “an irreducible (basal) cluster of organisms, diagnosably distinct from other such clusters, and within which there is a parental pattern of ancestry and descent”
- By sequence similarity & statistical methods → *species delimitation*

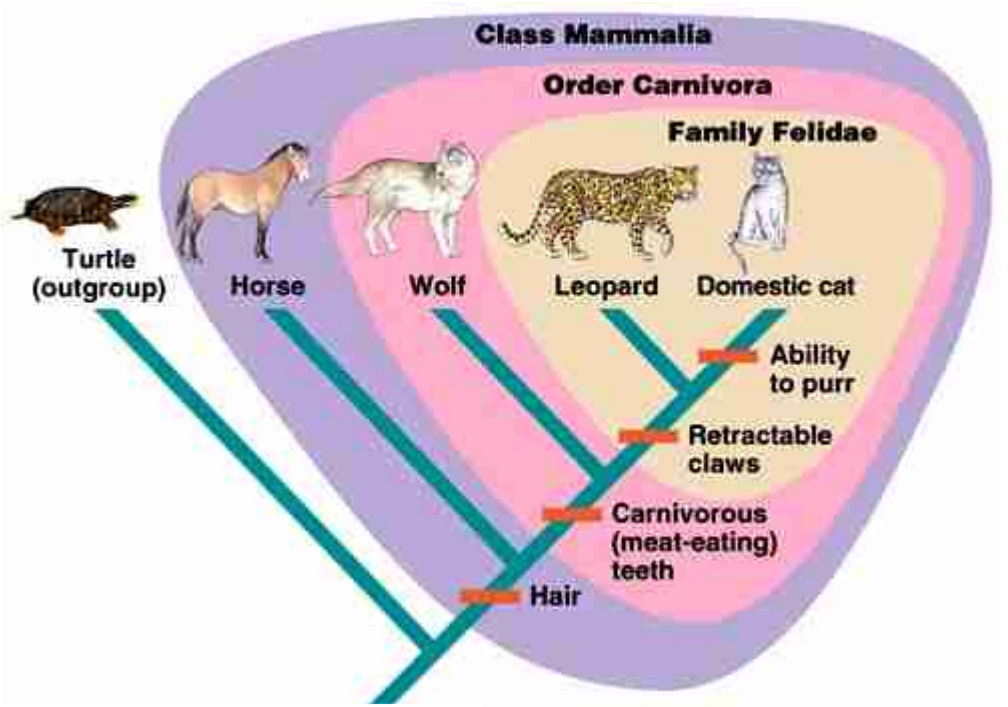
What's a species?

Interesting paper on this:
http://www.sciencedirect.com/science/article/pii/S016953471200100 0

"Coalescent-based species delimitation in an integrative taxonomy"

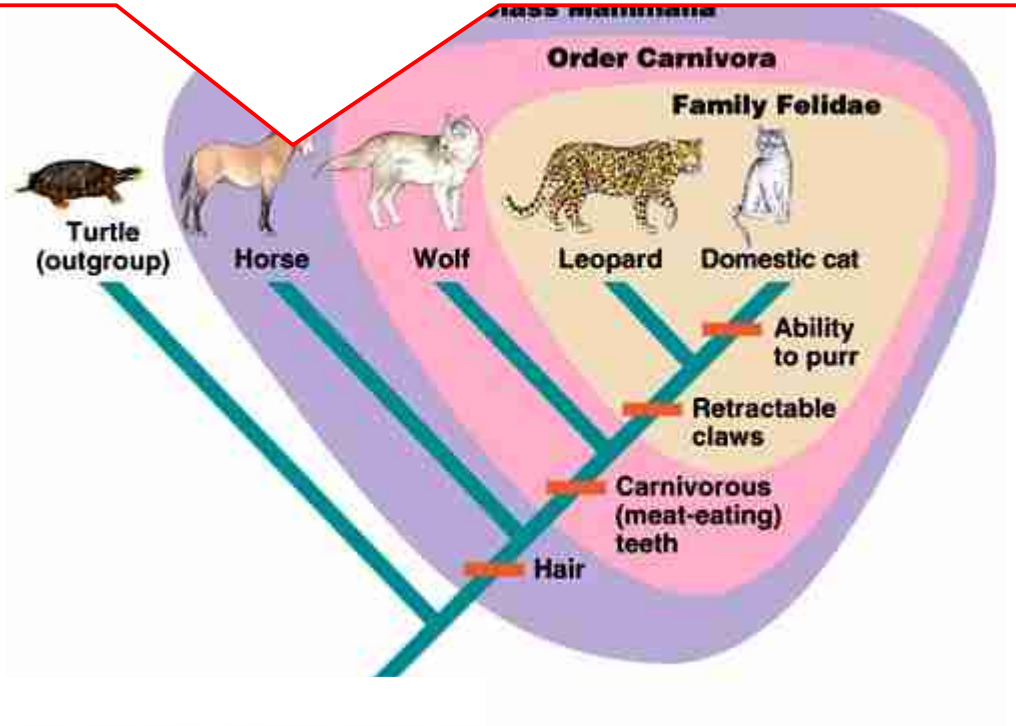
- Tricky que
- Different c
  - → gener
  - → more than *30* definitio
- By reproduction
  - → two species that can reproduce
  - → what about bacteria/viruses ????
- Evolutionary species concept
  - → via ancestral descent in an evolutionary tree
- General lineage (Abstammung/Verzweigung) concept
  - → an independently evolving lineage
- Phylogenetic Species Concept
  - → "an irreducible (basal) cluster of organisms, diagnosably distinct from other such clusters, and within which there is a parental pattern of ancestry and descent"
- By sequence similarity & statistical methods → *species delimitation*

A Taxonomy


Class Mammalia
Order Carnivora
Family Felidae
Turtle (outgroup)
Horse
Wolf
Leopard
Domestic cat
Ability to purr
Retractable claws
Carnivorous (meat-eating) teeth
Hair

A Taxonomy

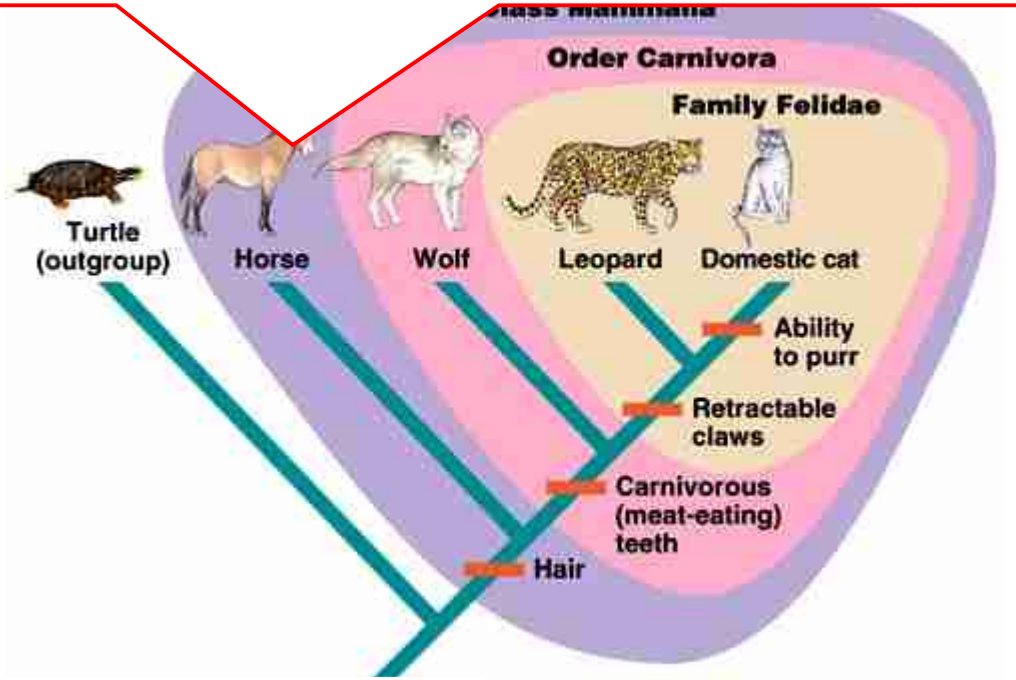
First systematic classification of living beings by Aristotele *384 -382* BC
Some terms still in use today, e.g., classification of animals into *Vertebrates* versus *Invertebrates*



A Taxonomy

First systematic classification of living beings by Aristotele *384 -382* BC
Some terms still in use today, e.g., classification of animals into *Vertebrates* versus *Invertebrates*

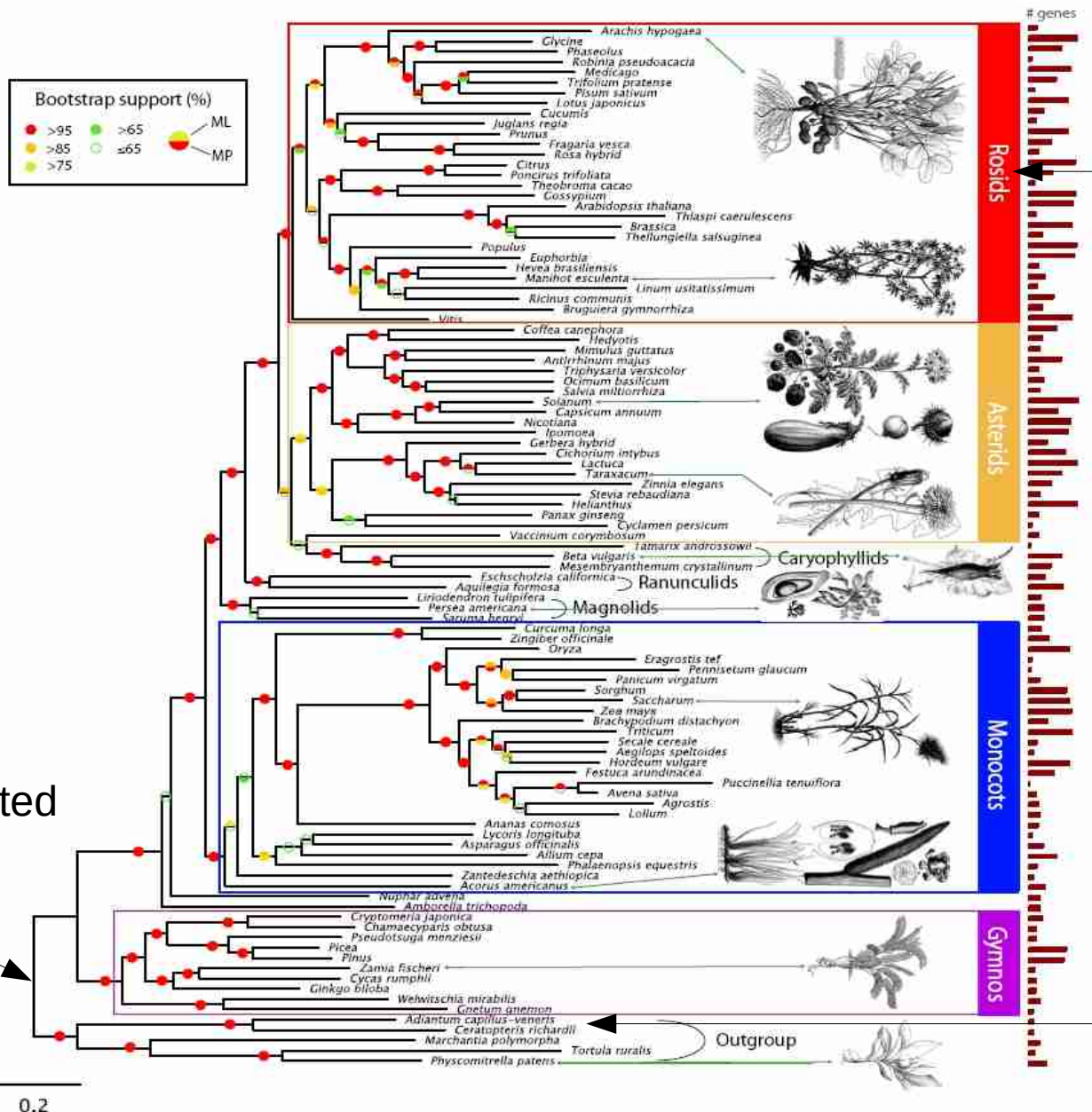
Wirbeltiere



Taxonomy

- Group biological organisms (species) into groups with similar characteristics
- Define characteristics of groups at different hierarchy levels, e.g., animals > mammals > great apes
- Taxonomic ranks
  - Domain → three domains of life
  - Kingdom
  - Phylum
  - Class
  - Order
  - Family
  - Genus
  - Species

A Phylogeny or Phylogenetic Tree



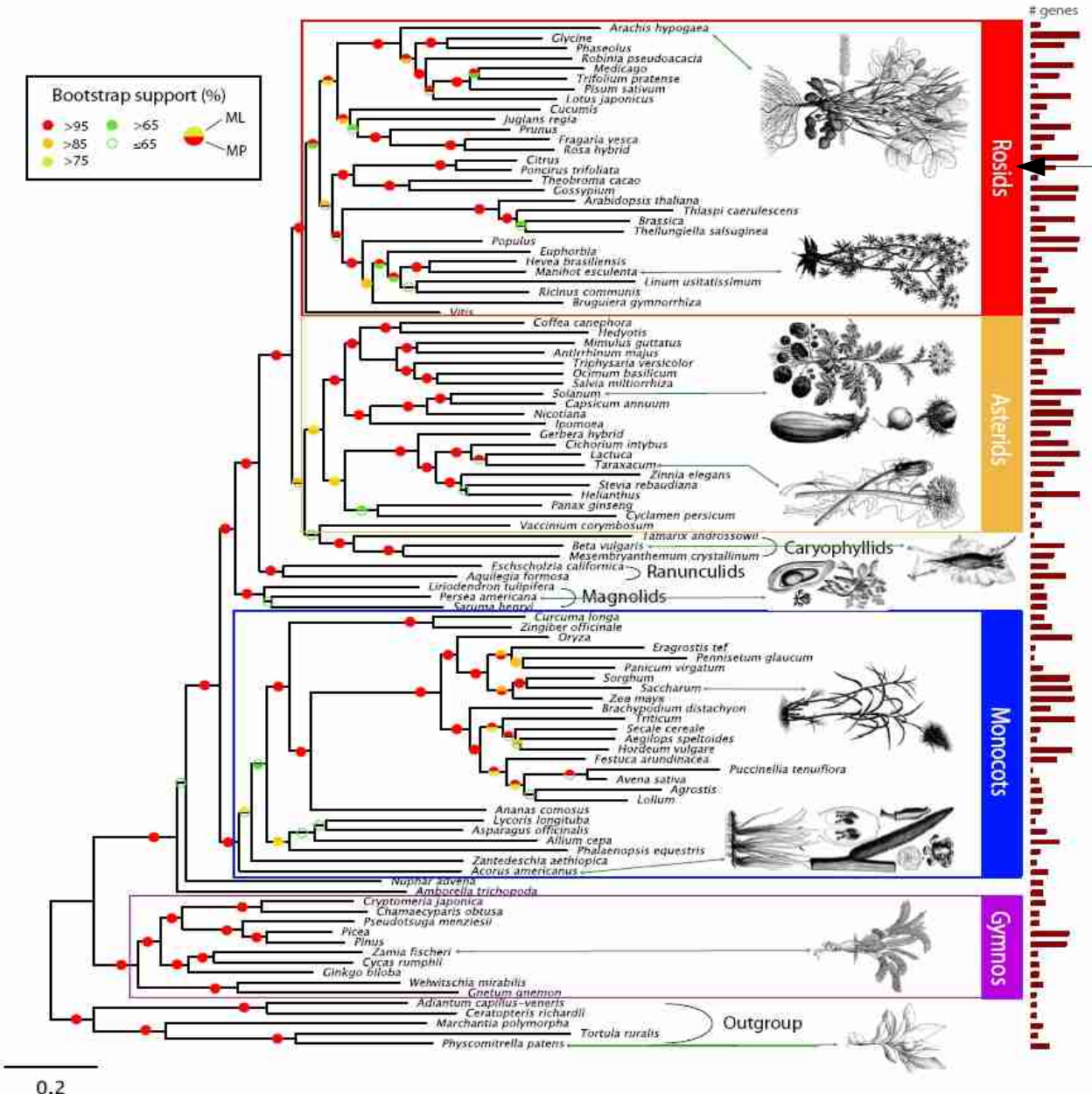
A taxonomic subclass

The ingroup

This tree is unrooted

The outgroup

A Phylogeny or Phylogenetic Tree



In Phylogenetics such a subtree is often also called *Lineage*!

Phylogeny

- An unrooted strictly binary tree
- Leafs are labeled by extant “übrig geblieben” (currently living) organisms represented by their DNA/Protein sequences
- Inner nodes represent hypothetical common ancestors
- *Outgroup*: one or more closely related, but different species → allows to root the tree

Taxon

- Used to denote clades/subtrees in phylogenies or taxonomies
- A group of one or more species that form a biological unit
- As defined by taxonomists
  - → subject of controversial debates
  - → part of the culture/fuzziness of Biology
- In phylogenetics we often refer to a single leaf as taxon
  - → the plural of taxon is *taxa*

A final quote

“Nothing in Biology makes sense except in the light of evolution” – Russian evolutionary biologist Theodosius Dobzhansky

Terminology introduced today

- Shotgun sequencing
  - Coverage
  - Paired-end reads
  - De novo versus by reference assembly
- Gene
  - Protein coding
  - RNA
  - Direction
  - Introns versus Exons
  - Splicing & alternative splicing
  - Function prediction
- RNA
  - tRNA
  - mRNA
  - rRNA
    - present in all organisms
    - important for inferring/calculating evolutionary relationships
    - 16S gene
  - Secondary RNA structure

Terminology introduced today

- Three domains of life
  - Eukaryota (with cell nucleus → splicing mechanisms)
  - Prokaryota (no cell nucleus)
  - Tree of life
- Codons
  - Redundancy
  - Start/stop Codons
  - Synonymous versus non-synonymous substitutions
- DNA
  - 3' versus 5' end
  - Default convention 5' → 3'
- Protein synthesis
- Transcription & translation
- The central dogma of molecular biology
- Transcriptome
- Meta-Genome
- Chromosome
  - Allele
- Species
- Taxonomy
- Phylogeny

Next Lecture

- Benoit Morel
  - Comparing sequences computationally
  - Algorithms on strings of DNA