Introduction to Bioinformatics for Computer Scientists

Lecture 1

Why is this a special course?

- As of 01.01.2023 and until 31.12.2027 I am working at the Foundation for Research and Technology in Crete, Greece
- … in the context of a large EU grant to set up a second research group at FORTH
- At the same time I am maintaining my professorship (including all rights and obligations) at KIT
- … and my research group at the Heidelberg Institute for Theoretical Studies
- I live and work in Crete most of the time

FORTH

FOUNDATION FOR RESEARCH AND TECHNOLOGY - HELLAS

Learning, Research, Innovation

So what about my teaching at KIT?

- **Idea:** set up and teach a joint, simultaneous CS Master level course at the computer science departments of the University of Crete (UoC) and KIT
- This is a totally new teaching experiment – I am looking forward to this
- However, the semesters only partially overlap

Teaching Schedule

- 4 live lectures at KIT → streamed to UoC
- 5 live lectures at UoC → streamed to KIT
- To fit the lecture content into the semester overlap we unfortunately have to do 3 hour lectures :-(
- This is not only tiresome for you ;-)
- However, we will be done before Christmas :-)
- Zoom links will be communicated via course mailing list

Live Lecture Schedule

When?	Where?	Who?
October 23	**KIT**	Alexis
October 30	**KIT**	Alexey & Lukas
November 6	**UoC**	Alexis
November 13	**KIT**	Alexis
November 20	**UoC**	Alexis
November 27	**UoC**	Alexis
December 4	**KIT**	Alexis
December 11	**UoC**	Alexis
December 18	**UoC**	Alexis

Preliminaries

- Lectures will be in English, evidently :-)
- Please send me an email to be included in the course mailing list
- Emails
  - Alexandros.Stamatakis@kit.edu
  - stamatak@ics.forth.gr
  - Alexandros.Stamatakis@h-its.org
- I usually reply within a day

Preliminaries

- Lab web-sites:
  - www.exelixis-lab.org (Heidelberg lab)
  - www.biocomp.gr (Crete lab)
- Course web-site:

  http://www.exelixis-lab.org/web/teaching/BioinformaticsModule.html
- Exelixis is the Greek word for evolution
- Slides & Videos
- Slides and videos from previous semesters
  - https://cme.h-its.org/exelixis/web/teaching/slides.html
  - Live lectures will deviate from pre-recorded videos
- Help us improve the course :-)

Etiquette

- Address me as Alexis in English, German, Greek if you like
- Please address me by name when writing me an email, don't start emails with *“Hi,“* or *“Hello,”*
- Office hours

  → send me an email to arrange for a virtual meeting
- Laptop, smartphones, tablets **CLOSED** policy
- Feel free to interrupt and ask as many questions as you like!
- Science needs controversial discussions!

Exam

- **KIT**: 20 minute oral exam → we can discuss the dates for this
- **UoC**: Also 20 minute oral exam planned, but I still need to figure this out

Instructors

- Mostly me
- However, the second lecture block will be taught by a staff scientist (Alexey) and a PhD student (Lukas) from the Heidelberg lab

The Heidelberg Lab

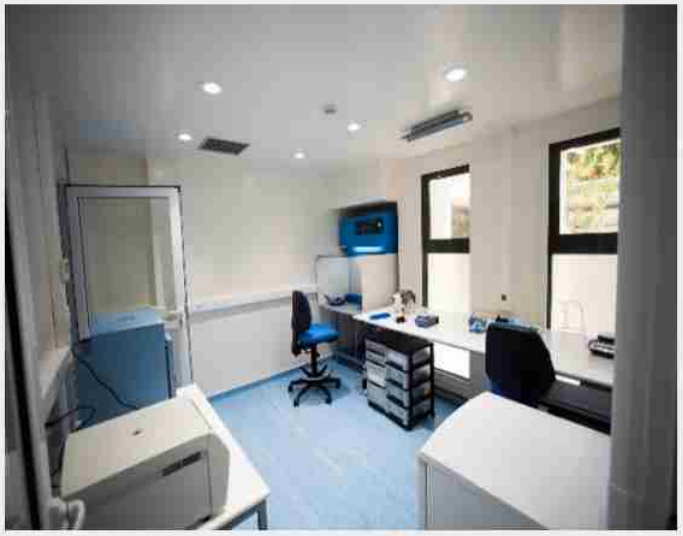
- *Computational Molecular Evolution Group*
  - 5 Phd students: *Julia*, *Dimitri*, *Luise*, *Anastasis*, *Lukas*
  - 1 PostDoc: *Benoit*
  - 1 staff scientist: *Alexey*
  - Several Master/Bachelor students & HiWis

The Crete Lab

- *Biodiversity Computing Group*
  - 3 PhD students will join in early 2024
  - 3 PostDocs: *Ben, Giorgos, Panos*

Another Lab in Crete

- I am also involved (a bit) in the ancient DNA lab
- Lab web-site: https://ancient-dna.gr/index.php/en/



Your Instructors in chronological order

- Alexis

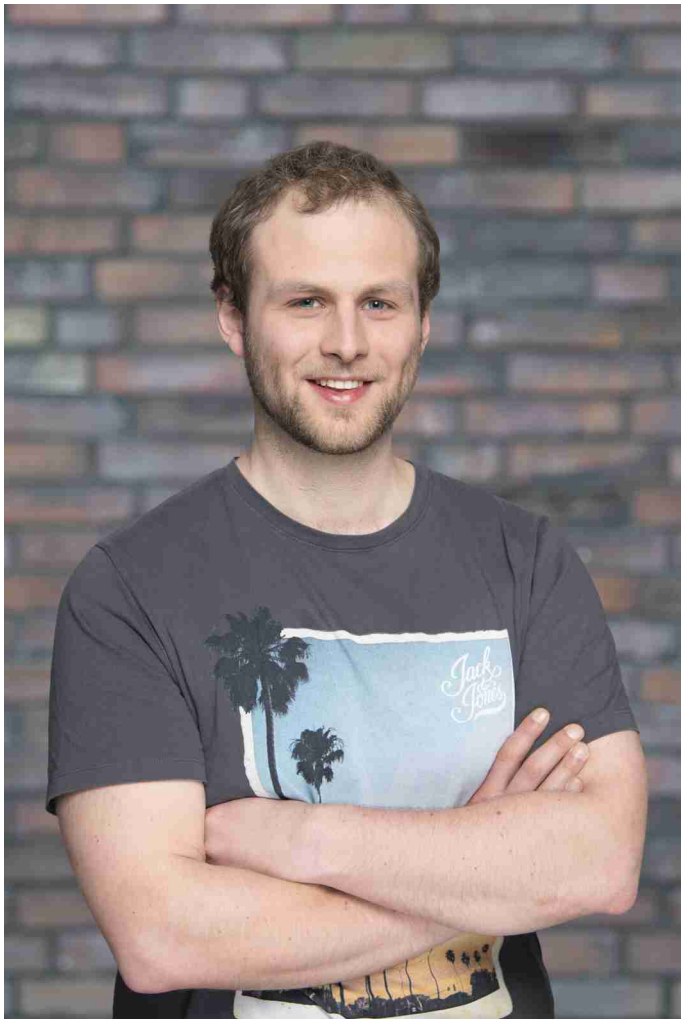
ERA (European Research Area) Chair at FORTH
Full Prof. at KIT
Associated Research Group Leader at Heidelberg Institute for Theoretical Studies

Some Biographical Bullets

- until 1995: grown up in Athens, Greece
- 1995-2004: Diploma & PhD in CS at TU Munich
- 2005-2006: PostDoc in Crete
- 2006-2008: PostDoc at ETH Lausanne
- 2008-2010: Emmy-Noether group leader at LMU and then TU Munich
- Since 2010: Research group leader at HITS Heidelberg
- Since 2012: Full professor at KIT
- Since 2020: Stuck in Crete due to the pandemic
- Since 2023: ERA chair at Institute of Computer Science at FORTH

Your Instructors in chronological order

- Lukas Hübner



Shared PhD student with Peter Sanders & former Master's student at KIT

Your Instructors in chronological order

- Alexey Kozlov

Former PhD student & former Master's student at KIT, staff scientist at HITS

Goals of this Course

- introduce *some* biological terminology
- present *some* areas of Bioinformatics
- provide an overview
- show that there are interesting algorithmic & computational problems
- provide you the knowledge you need to work with us on research projects (Master's thesis etc.)

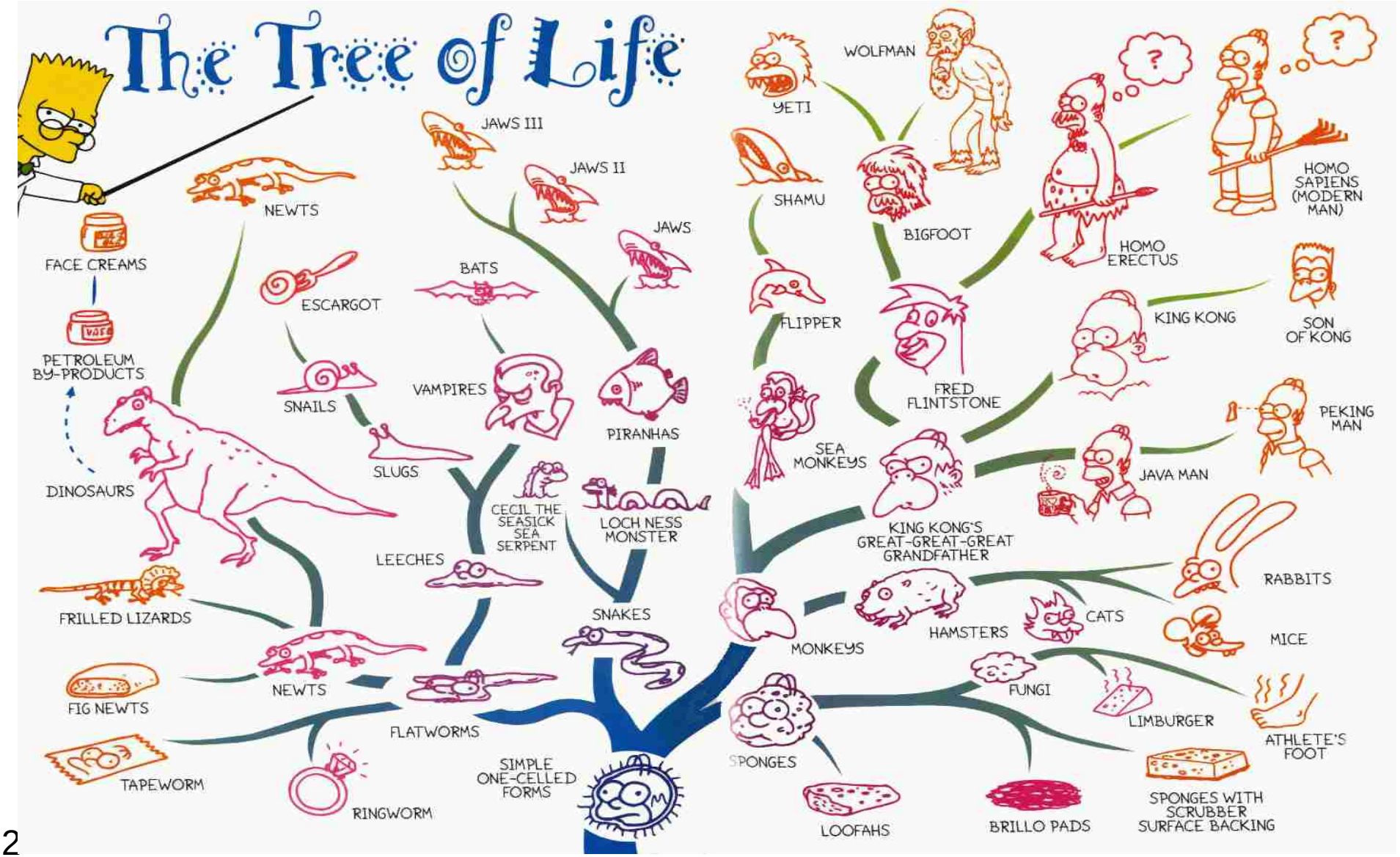
Course Structure

- **October 23**
  - **Introduction & Basic Molecular Biology**
- **October 30**
  - **Pair-wise Sequence Alignment**
  - **BLAST & Genome Assembly**
- **November 6**
  - **Multiple Sequence Alignment**
  - **Introduction to Phylogenetics**
- **November 13**
  - **Introduction to Phylogenetics (continued)**
  - **Phylogenetic Search Algorithms**
- **November 20**
  - **A brief introduction to Markov Chains**
  - **Maximum Likelihood Lecture**

Course Structure

- **November 27**
  - **Maximum Likelihood Lecture (continued)**
  - **Advanced Maximum Likelihood Stuff**
- **December 4**
  - **Bayesian inference & MCMC**
  - **Advanced Bayesian inference & MCMC**
- **December 11**
  - **Advanced Bayesian inference & MCMC (continued)**
  - **Introduction to Population Genetics**
- **December 18**
  - **Introduction to Population Genetics (continued)**
  - **Wrap up & exam preparation**

Main Research Focus of my Lab



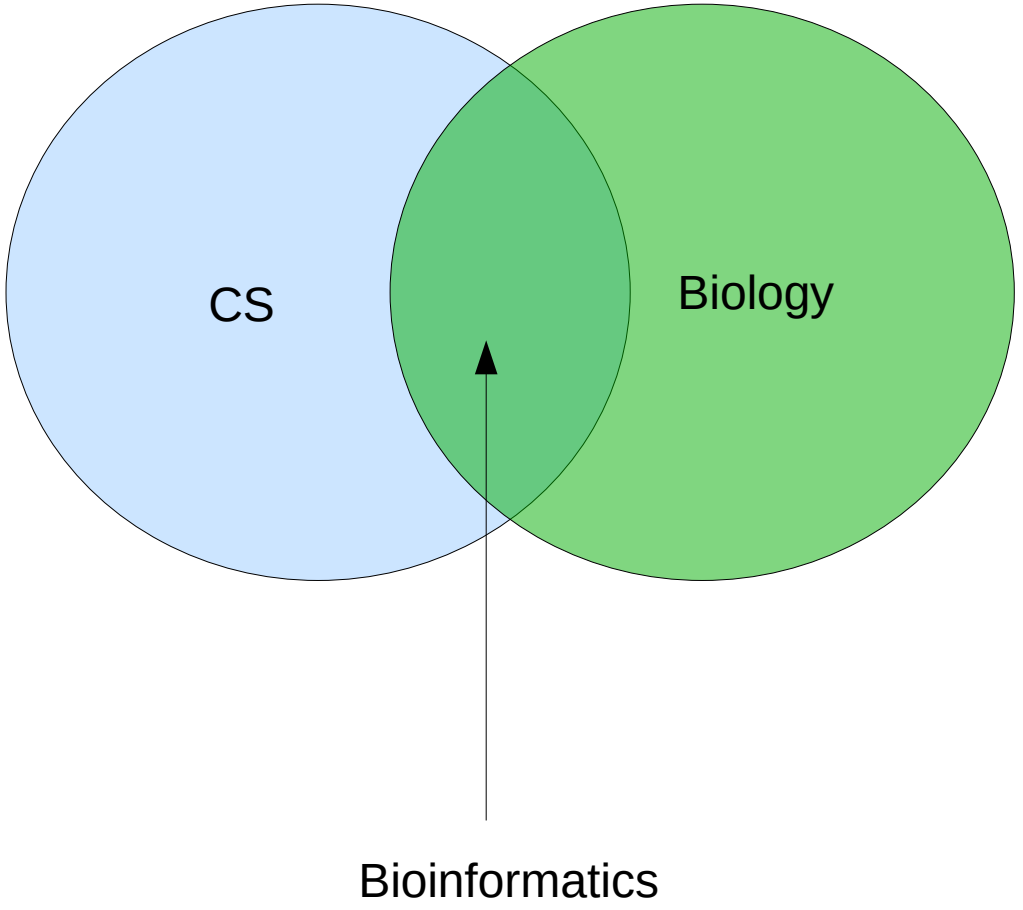
What is Bioinformatics?

- Term introduced by P. Hogeweg & B. Hesper in 1970 http://en.wikipedia.org/wiki/Paulien_Hogeweg
- There are many definitions
- I will provide my own:
  - In bioinformatics we intend to develop, optimize, and parallelize algorithms, models, and production-level software for analyzing, storing, and extracting knowledge from, biological raw data.
  - Key differences to CS
    - proof-of-concept implementations are not sufficient
    - we need to produce code that can be used by biologists
    - we need to provide support for the code
    - have a look at http://groups.google.com/group/raxml
    - Most famous Bioinformaticians are known for one or more widely-used and highly cited algorithms & tools they have developed
- “Biology easily has 500 years of exciting problems to work on” – Donald Knuth

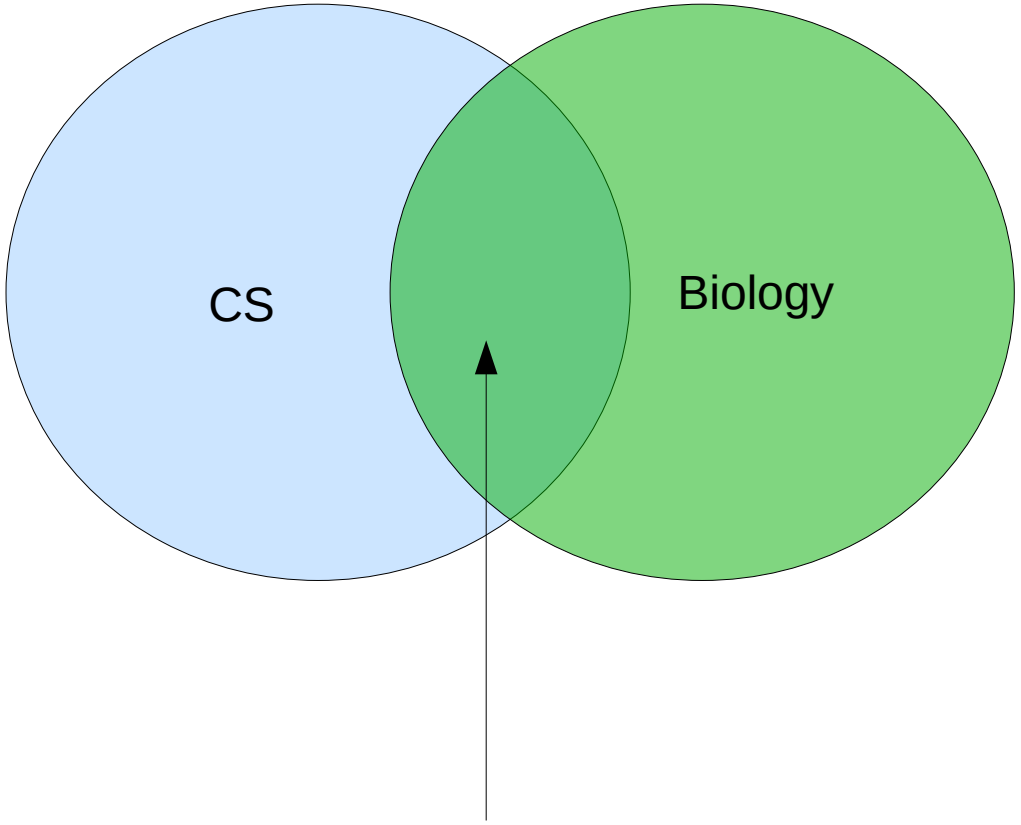
The ideal Bioinformatics tool

What is my hypothesis?

What is Bioinformatics?

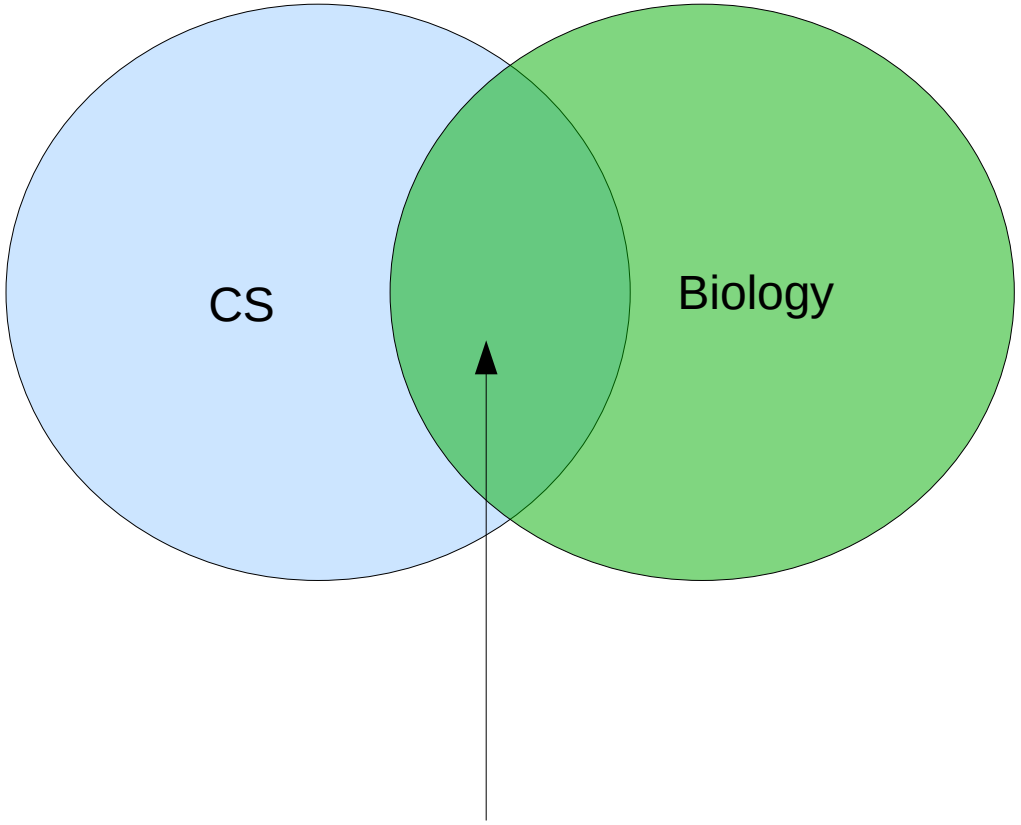


Why is this exciting?



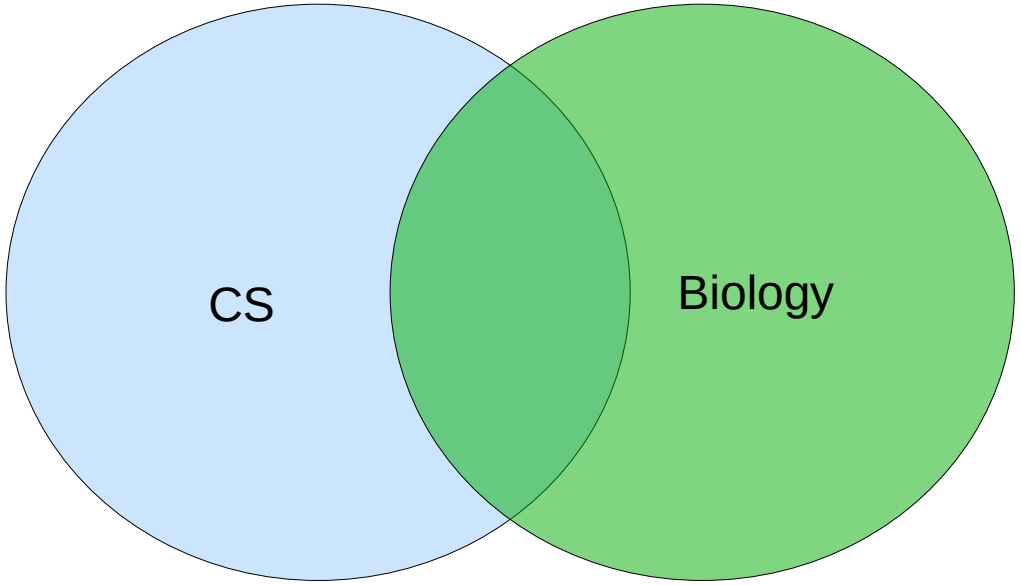
Important problems → medical applications,
Infectious diseases, genetic defects etc.
Masses of data → storage and analysis challenges
HPC → increased need for parallel codes

What are the challenges?



We can't be experts in everything → interdisciplinary collaboration
We need a culture of asking questions when we don't understand a term/concept!

Disciplines involved



Numerics
Statistics
Discrete Algorithms
Algorithm Engineering
Parallel Computing
Supercomputing
Software Engineering (in practice)

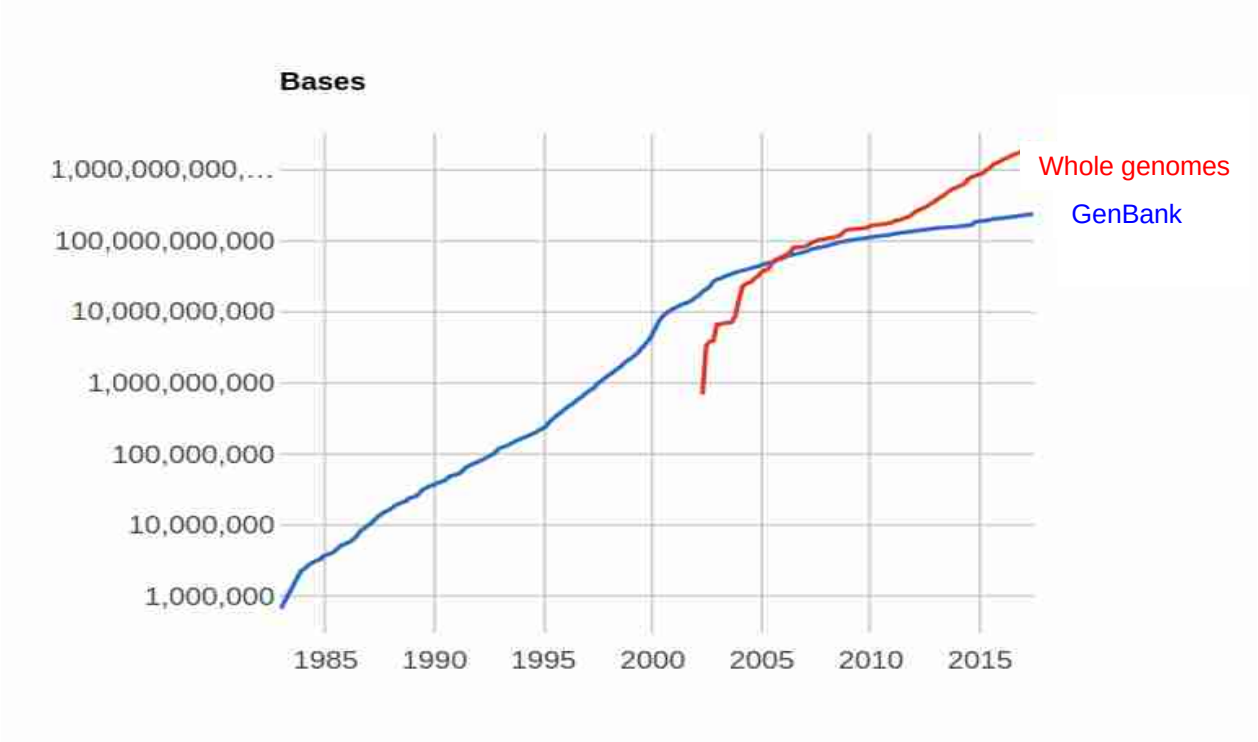
What is Biological Raw Data?

- There are many types of biological raw data
  - Images from microscopes
  - Microarray data
  - Protein structure data
  - Morphological data
  - Ecological data
  - Biogeographical data
  - …
- In this course we will mainly focus on *classic* Bioinformatics, that is, the analysis of molecular sequence data (DNA, protein data)

DNA data

- DNA data is available in public databases
- The most well-known one is GenBank
- Maintained by NCBI: National Center for Biotechnology Information, US
- Other databases for DNA data: EMBL (EU), DDBJ (Japan)

# of nucleotides/
base pairs
log-scale!



DNA data

- Genetic sequence
- Alphabet of 4 basic characters (nucleotides):
  - **A**denine
  - **C**ytosine
  - **G**uanine
  - **T**hymine
- A DNA sequence: *AACGTTTGA*
  - This sequence has 9 base pairs/nucleotides
- In RNA data: **T** is replaced by **U**racil
- A RNA sequence: *AACGUUUGA*
- We will see what RNA is later
- If we use **T** or **U** does usually not matter, computationally

Extended DNA alphabet

- DNA sequencing techniques are not exact
- Need to extend character set to denote:
  - could be an A *or* C
  - could be an A *or* C *or* G
  - ...
- International Union for Pure and Applied Chemistry (IUPAC) encoding

Ambiguity Code

Code	Represents	Complement
A	Adenine	T
G	Guanine	C
C	Cytosine	G
T	Thymine	A
Y	Pyrimidine (C or T)	R
R	Purine (A or G)	Y
W	weak (A or T)	W
S	strong (G or C)	S
K	keto (T or G)	M
M	amino (C or A)	K
D	A, G, T (not C)	H
V	A, C, G (not T)	B
H	A, C, T (not G)	D
B	C, G, T (not A)	V
X/N	any base	X/N

Ambiguity Code

We will talk about this later!

Code	Represents	Complement
A	Adenine	T
G	Guanine	C
C	Cytosine	G
T	Thymine	A
Y	Pyrimidine (C or T)	R
R	Purine (A or G)	Y
W	weak (A or T)	W
S	strong (G or C)	S
K	keto (T or G)	M
M	amino (C or A)	K
D	A, G, T (not C)	H
V	A, C, G (not T)	B
H	A, C, T (not G)	D
B	C, G, T (not A)	V
X/N	any base	X/N
-	Gap	-

DNA Sequencing

- The process of reading the nucleotide bases in a DNA molecule
- There exist various sequencing technologies
- Properties
  - Cost
  - Speed
  - Amount of data/Number of Sequences
  - Sequence length
  - Error rate

DNA Sequencing

- Sanger sequencing (*since 1977)*
  - High accuracy: 99.9%
  - Long sequences: 300-900 nucleotides
  - Expensive: $2400 per 1,000,000 nucleotides
  - Few sequences: up to ≈ 100
- Next-generation sequencing (since *2007*)
  - Lower accuracy 98-99.9%
  - Short sequences (100-400 nucleotides)
  - Inexpensive $1 - $10 per 1,000,000 nucleotides
  - Many sequences: 500 – 3,000,000,000 per sequencer run

A next-Generation Sequencer

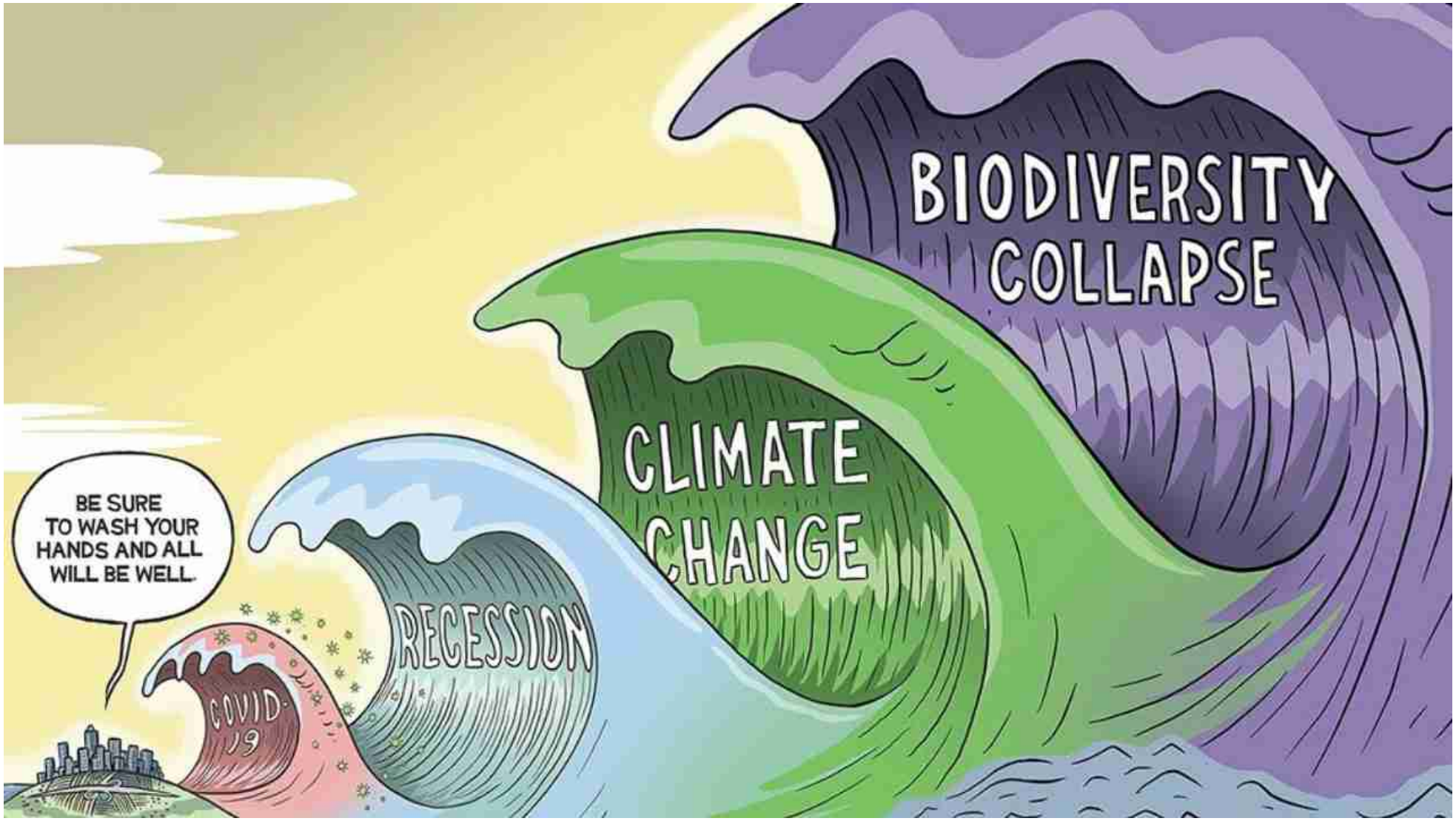
A Next^2 Generation Sequencer

DNA Sequencing

- Sanger sequencing
  - High accuracy: 99
  - Long sequences:
  - Expensive: $2400 per 1 00 nucleotides
  - Few sequences: up to ≈ 00
- Next-generation sequencing (since *2007*)
  - Lower accuracy: 98-99.9%
  - Short sequences (100-400 nucleotides)
  - Inexpensive $1 - $10 per 1,000,000 nucleotides
  - Many sequences: 500 – 3,000,000,000 per sequencer run

This is a revolution!
We will see how this data can be used and analyzed in this course!

Why care about Biodiversity?



The Biodiversity Crisis

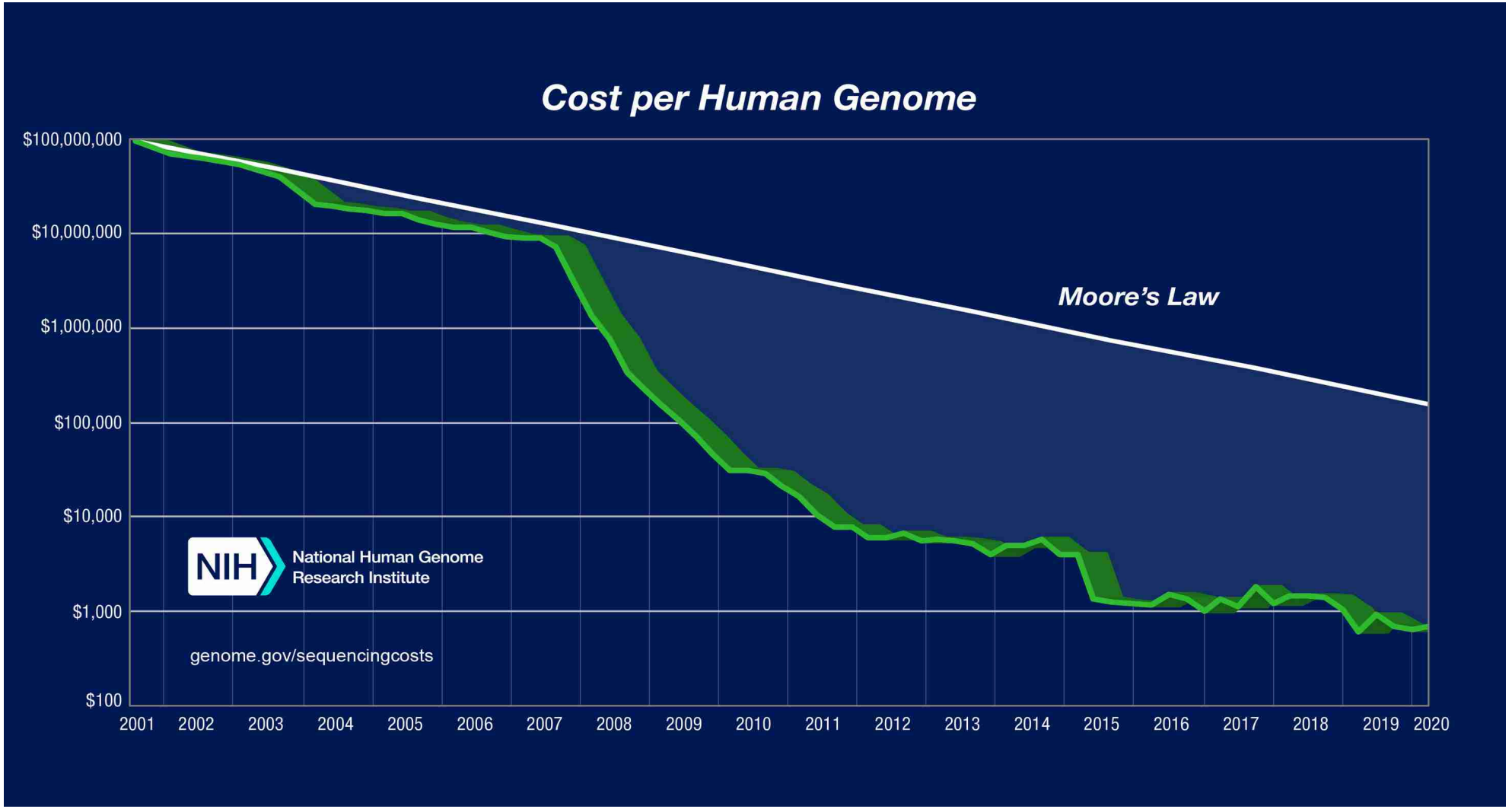
- Suggested Reading: “How genomics can help Biodiversity conservation”

  https://www.cell.com/trends/genetics/fulltext/S0168-9525(23)00020-3
- Maybe present as seminar paper in summer term?

Table 1. List of genomic approaches with application and comparison of raw sequencing costs (i.e., costs of sample collection, researcher time, and analysis are not included)

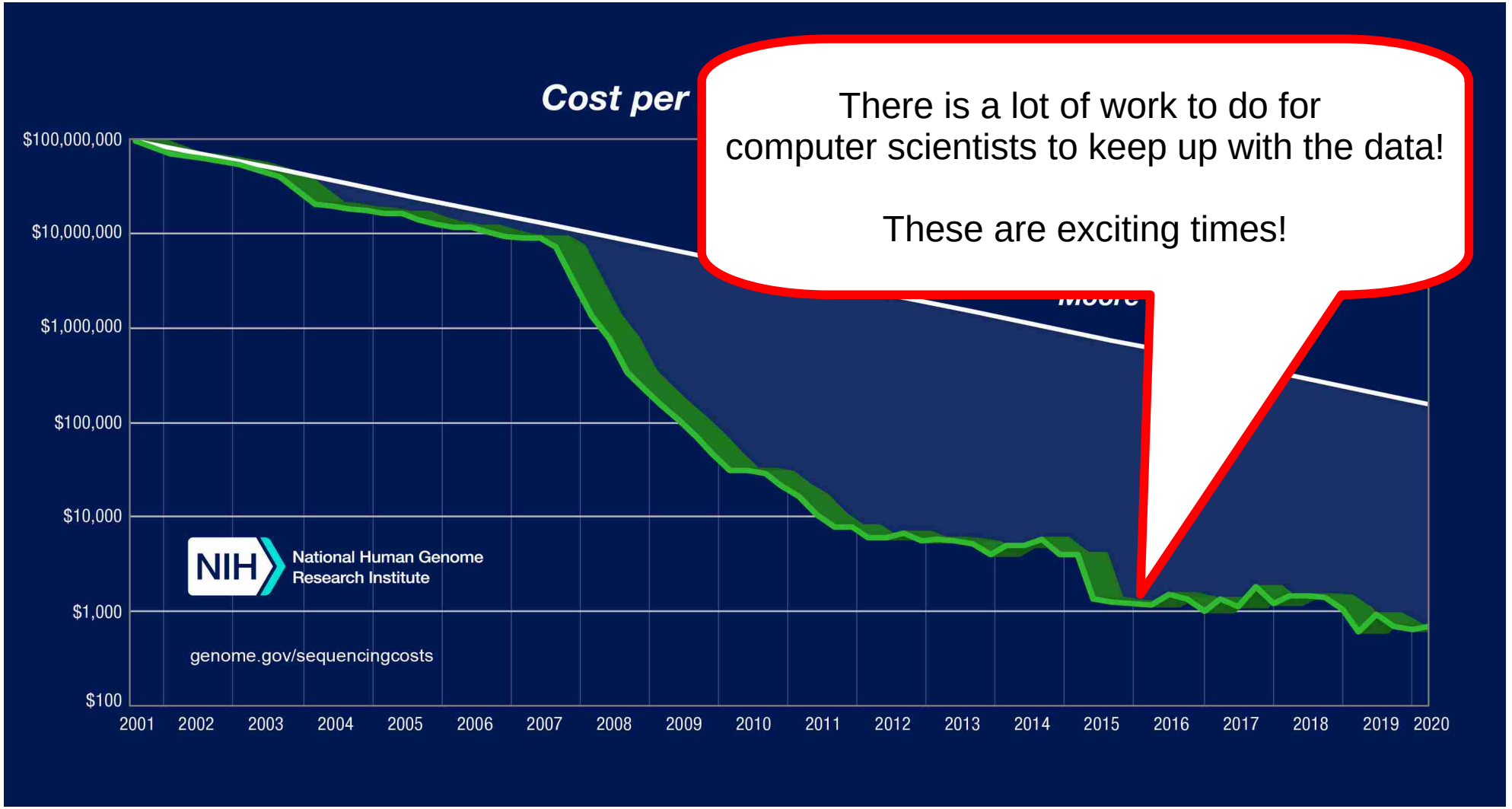
	DNA barcoding/ metabarcoding	Genome skimming	Reduced representation DNA sequencing	Transcriptome sequencing	Whole-genome resequencing
What genome do you get?	None	Organelle, *k*-mer representation of nuclear	None	Coding regions only, variable fragmentation	Nonrepetitive genome, depends on coverage
Cost in dollars (as of date)*	$5 per sample (Sanger/NGS)	$50	$50	$100–$400	$100–$800
What type of samples are needed	Fresh tissue samples, museum specimens, noninvasive samples	Fresh tissue samples, museum specimens	Fresh tissue samples, museum specimens, noninvasive samples	Tissue-specific, live/fresh, flash frozen/in RNA buffer	Fresh tissue samples, museum specimens
Genetic diversity*	Yes, but limited	Yes, but limited	Yes	Yes	Yes
Population structure*	Yes, but weak to detect shallow/cryptic genetic structure; economical for detailed spatial sampling	Yes, typically organelle based	Yes	Yes	Yes
Phylogenetic information	Yes, but barcode based	Yes, typically organelle based	Yes	Yes	Yes
Introgression event	No	No	Yes, but no individual genes	Yes, but limited detection power	Yes
QTL mapping	No	No	Yes, but low resolution	Yes, expression QTL (eQTL)	Yes
Natural selection signal detection	No	Yes, on organelle genes	Yes	Yes	Yes
Gene structure study	No	Yes, on organelle genes	Potentially, if reference genome available	No	Yes, if reference genome available
Gene family analyses	No	Yes, on organelle genes	Potentially, if reference genome available	Yes	Yes, if reference genome available
Genome rearrangement study	No	Yes, on organelle	Potentially, if reference genome available	No	Yes, if reference genome available
Functional genomic study*	No	Yes, on organelle genes	No	Yes	No
Genome size estimation	No	Yes, typically organelle genome	No	No	Depending on coverage
Linkage disequilibrium	No	No	Yes, usually, not always	Limited	Yes
Demographic reconstructions (MSMC)*	No	No	No	No	Yes
Demographic reconstructions from SFS	No	No	Yes	No	Yes
GWAS	No	No	Yes, but low resolution	Yes, T(ranscriptome) WAS	Yes

The revolution



Sequencing cost versus processing cost!

The revolution



Sequencing cost versus processing cost!

Remember

- Back in 2001 the complete sequencing of the human genome made the news!
- Papers appeared in *Science & Nature*
- Now it's almost boring: aha, somebody sequenced yet another genome
- Our lab in 2014
  - Evolutionary analysis of 50 bird *genomes*
  - Evolutionary analysis of 140 insect transcriptomes → we will see what a *transcriptome* is later

Bird & Insect Papers

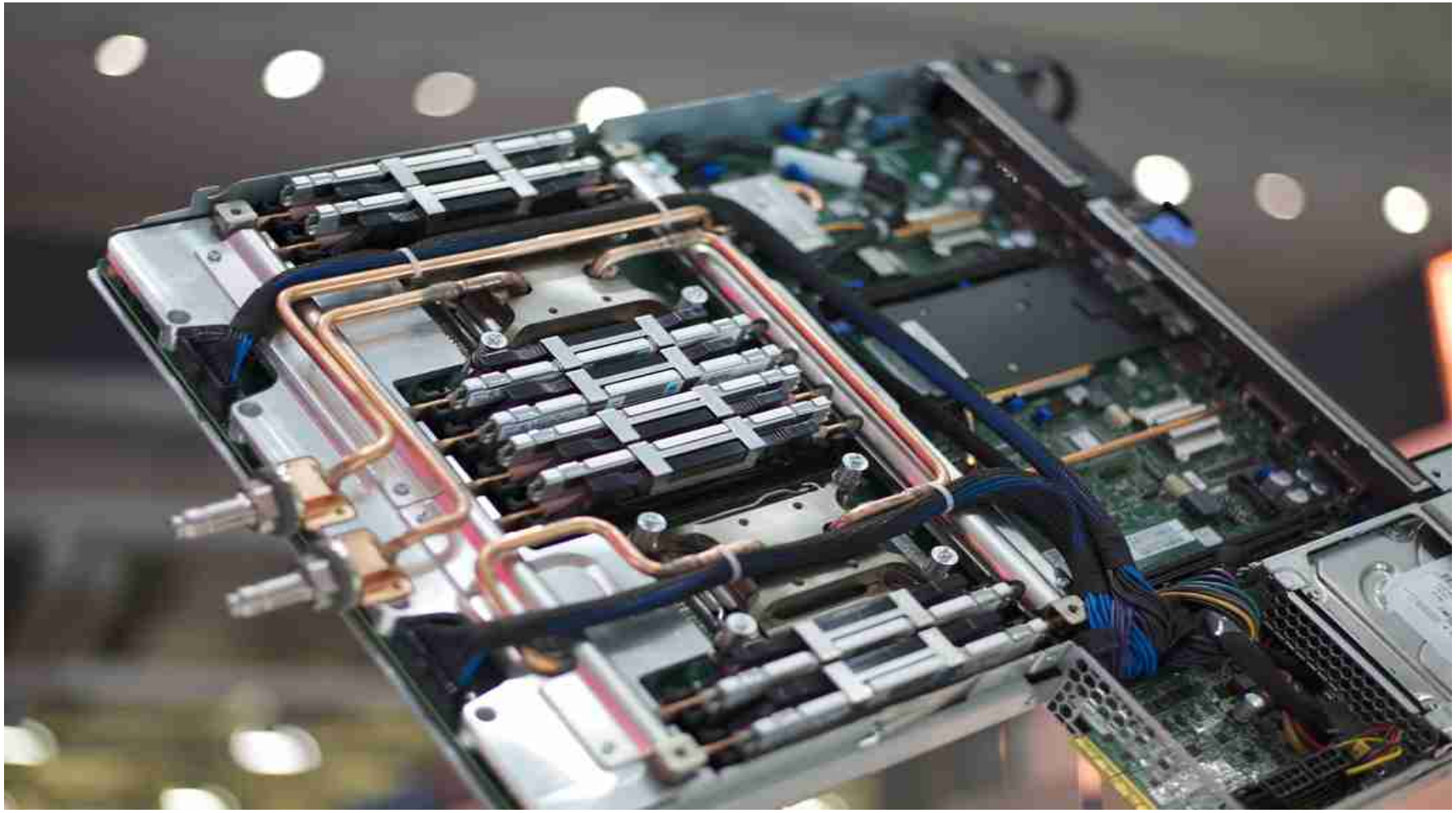
Bird & Insect Papers

In 2018/19 we were analyzing datasets of 1500 insects and 350 birds

Supercomputing

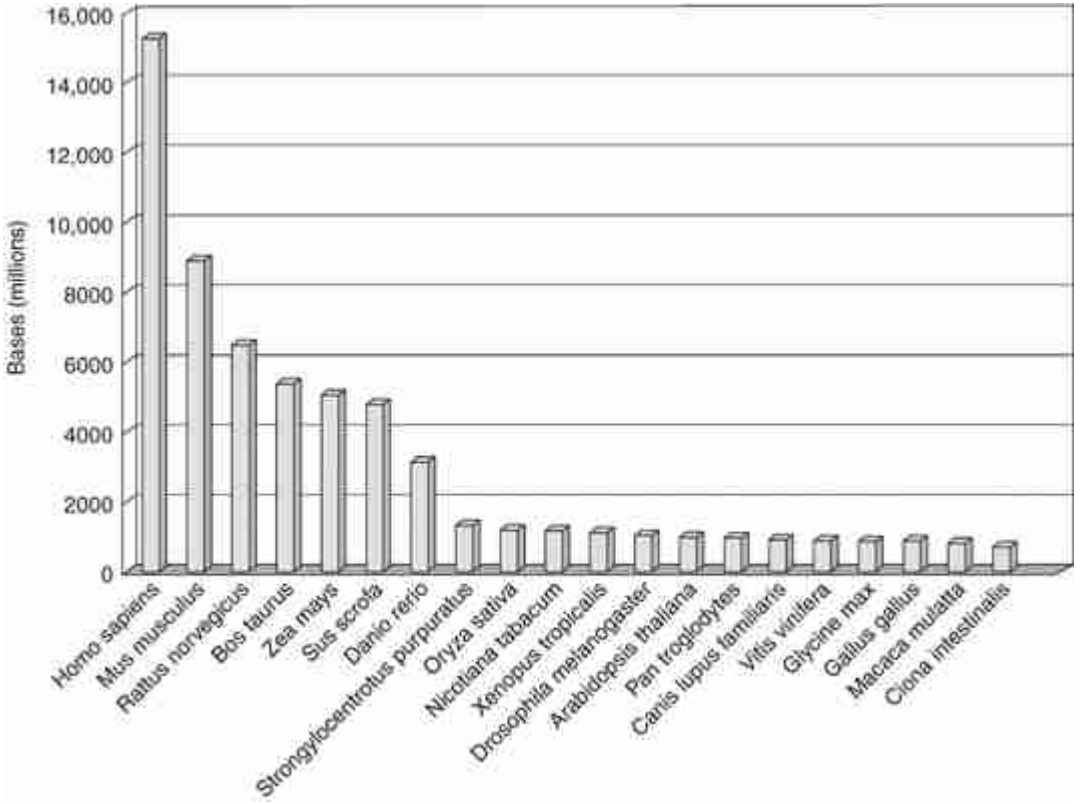
Munich supercomputer: SuperMUC

SuperMUC Cooling



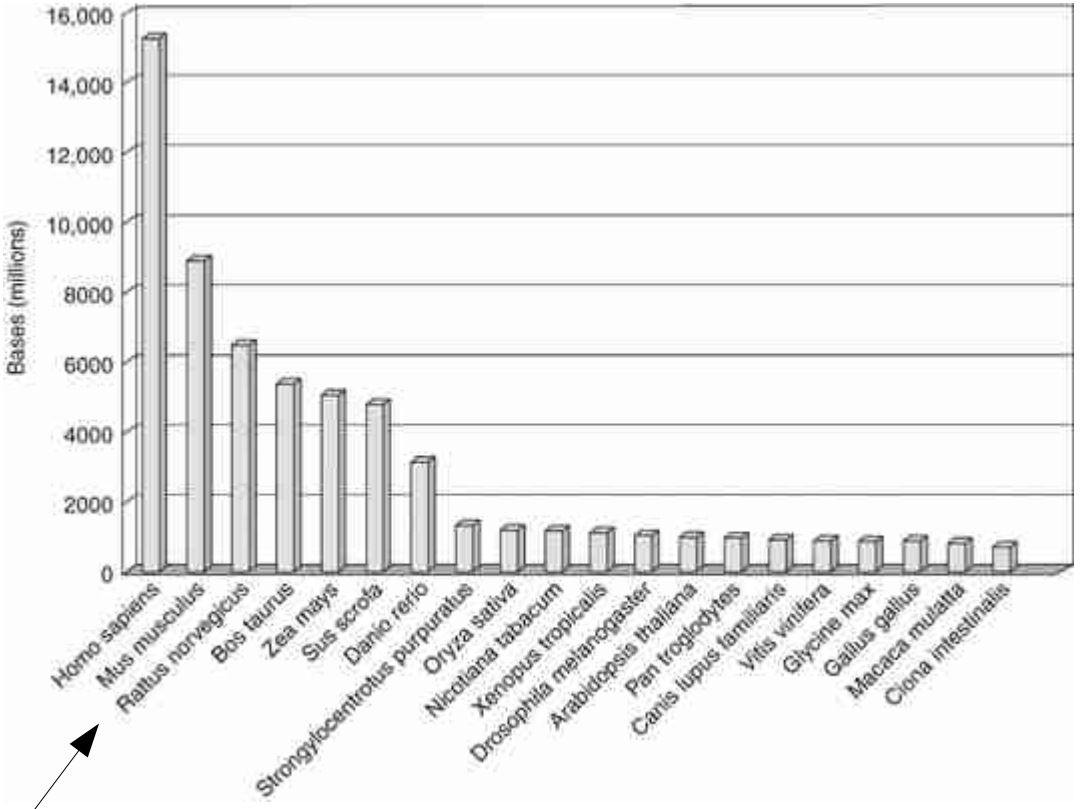
DNA data

- GenBank: most-sequenced species



DNA data

- GenBank: most-sequenced species



 Some of these species are so-called *model organisms*

Model Organism

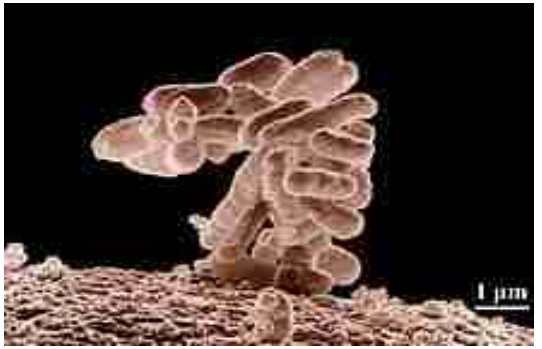
- A species that is extensively studied/sequenced to understand particular biological phenomena, with the expectation that discoveries made for the model organism will provide insight into the workings of other organisms.
- Selection criteria:
  - easy experimental manipulation
  - ease of genetic manipulation
  - easy to grow

    → short life-cycle/generation times
  - easy to extract DNA data
  - Economical importance → rice
- Often researchers reverse-engineer organisms
- Full list of model organisms: http://www.life.umd.edu/labs/mount/Models.html

Some Model Organisms

- *Escheria coli*

  gut bacterium → can cause food poisoning, grows fast, inexpensive to cultivate



- *Drosophila Melanogaster*

  fruit fly → breeds quickly

- *Arabidopsis Thaliana*

  flowering plant → small genome

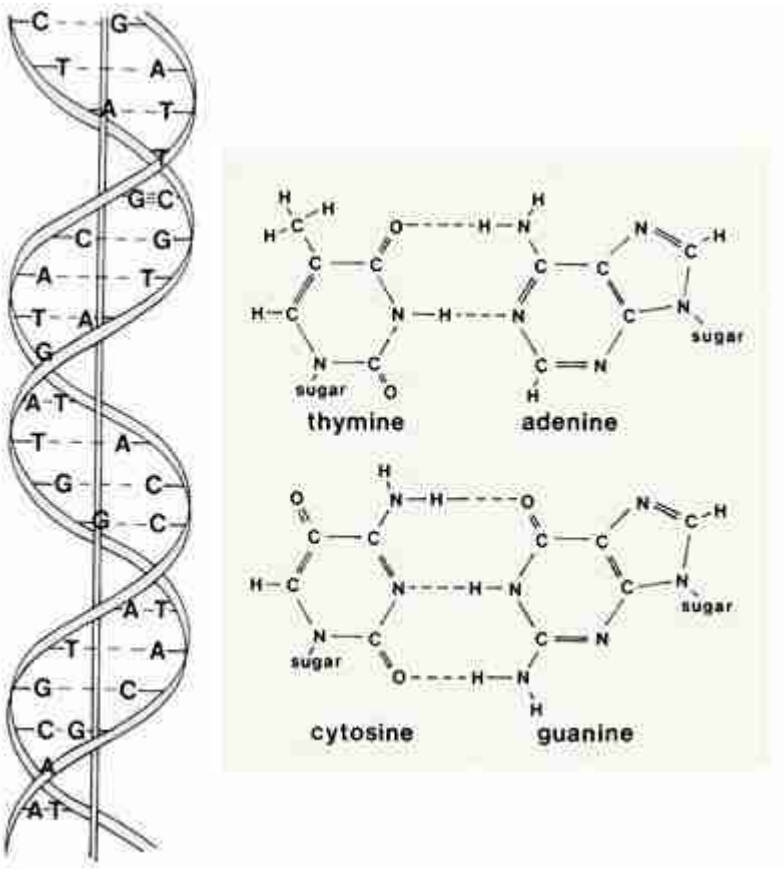
Back to DNA

- What's a base pair?
- Pairing of A with T *or* C with G in double-stranded DNA

AATTGGC

TTAACCG

complement



Sloppy terminology

- The # of base pairs is frequently used as synonym for the # of nucleotides in a single-strand sequence
- The following sequence has 5 nucleotides: *ACGGT*
- We can also say that it has 5 base pairs
- As in CS we use kilo, giga, etc for sequence lengths
  - kb → kilo-bases
  - Mb → Mega-bases
  - Gb → Giga-bases

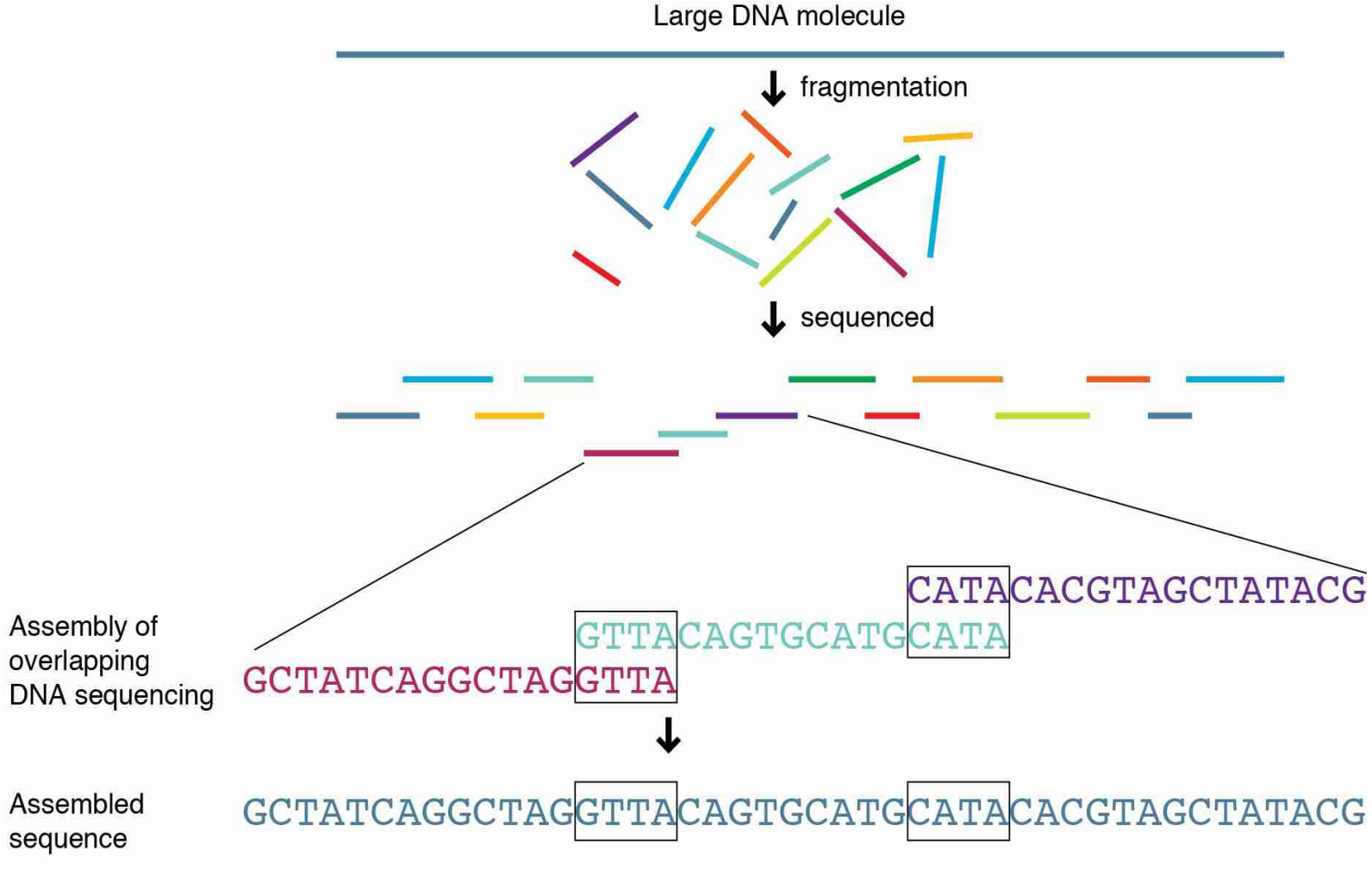
Genome

- The full genetic information of an organism
  - Contains all chromosomes
  - Comprises the coding & non-coding sequence data of the organism
  - Coding sequence data → part of the genome that encodes proteins
  - Non-coding (in earlier days: junk) DNA → part of the genome that does not encode proteins but still has a function
    - The function of non-coding DNA is only partially known
    - Non-coding DNA regulates protein processes

Genome Size

- Not necessarily correlated with organism complexity
- *Homo Sapiens*: 3.2 Gb (Giga-bases)
- *Marbled lungfish*: 130 Gb (Giga-bases)
- Plants often have very large genomes → partially due to redundant information caused by hybridization

Shotgun Sequencing



Shotgun Sequencing

- We can read fragments up to a length of ≈ 1000 bp

  → 1000 bp correspond roughly to the length of an average gene

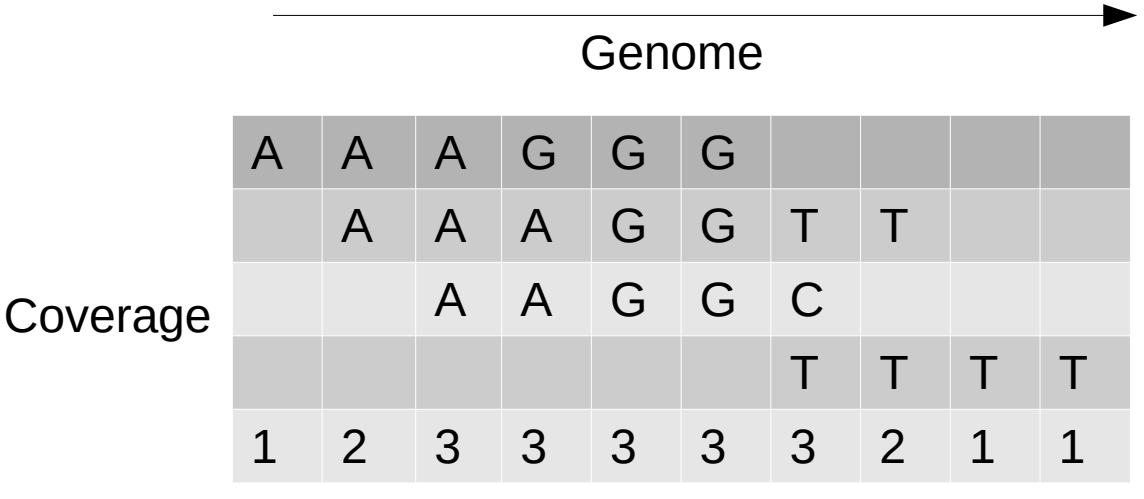
- What do we do for reading *genomes*?

  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

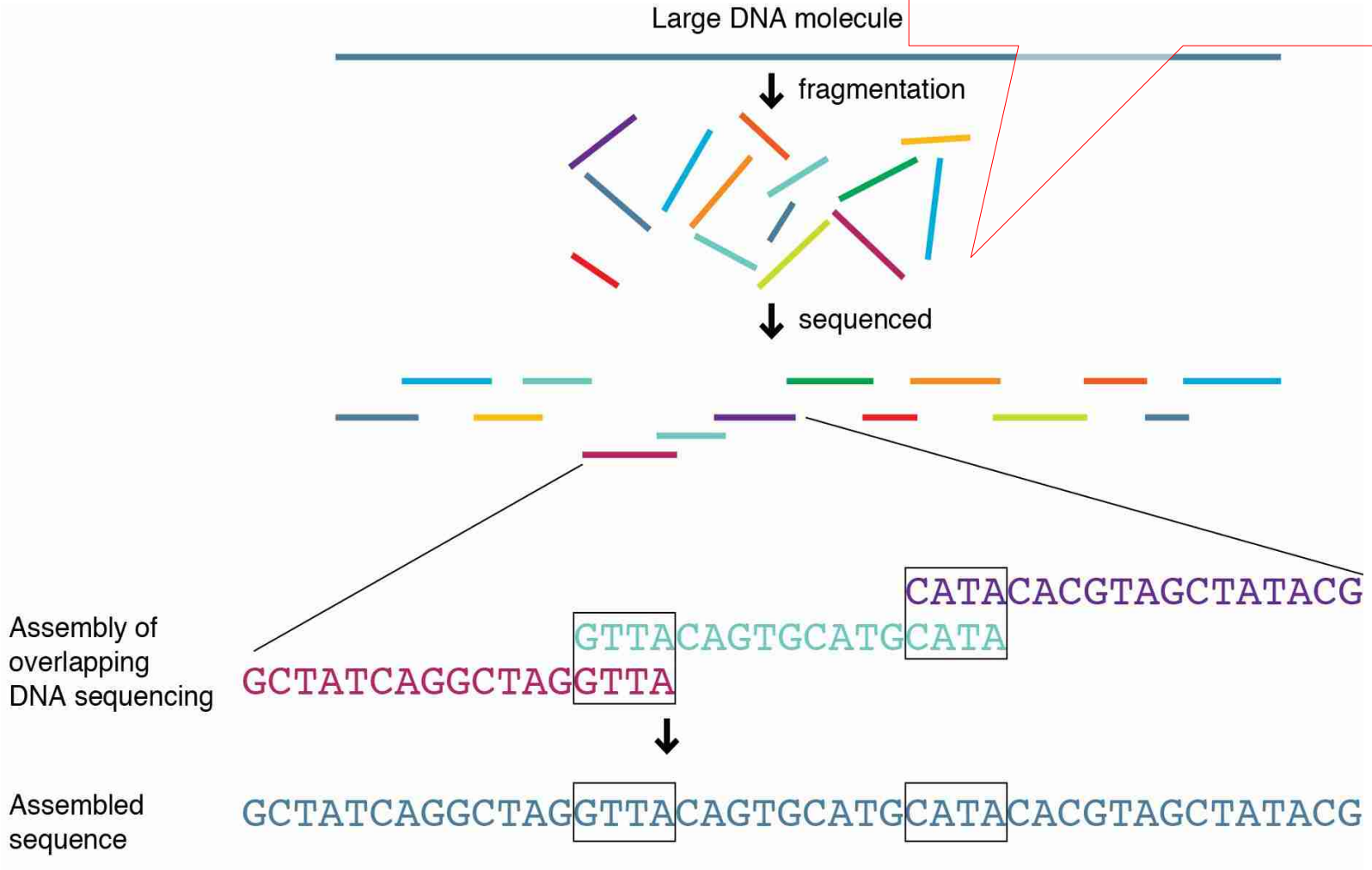


Shotgun Sequencing

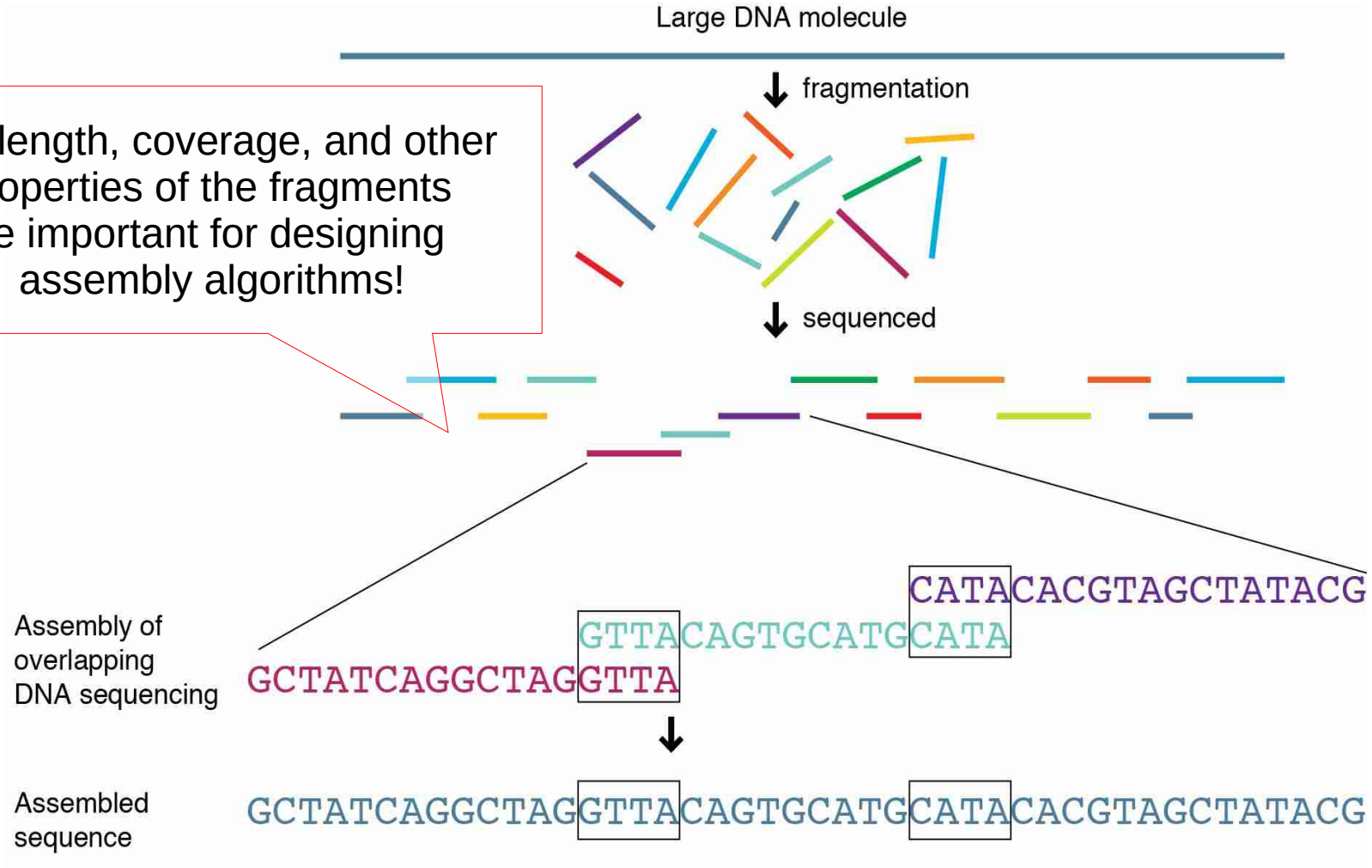
- We can read fragments up to a length of ≈ 1000 bp (Sanger Sequencing)
- What do we do for reading genomes?
  1) Break up genome randomly into fragments
  2) Read fragments
  3) Assemble fragments into a genome with computers
- Important characteristics:
  - *Coverage*
  - *Fragment* length
  - *Paired-end*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

Shotgun Sequencing

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



Shotgun Sequencing



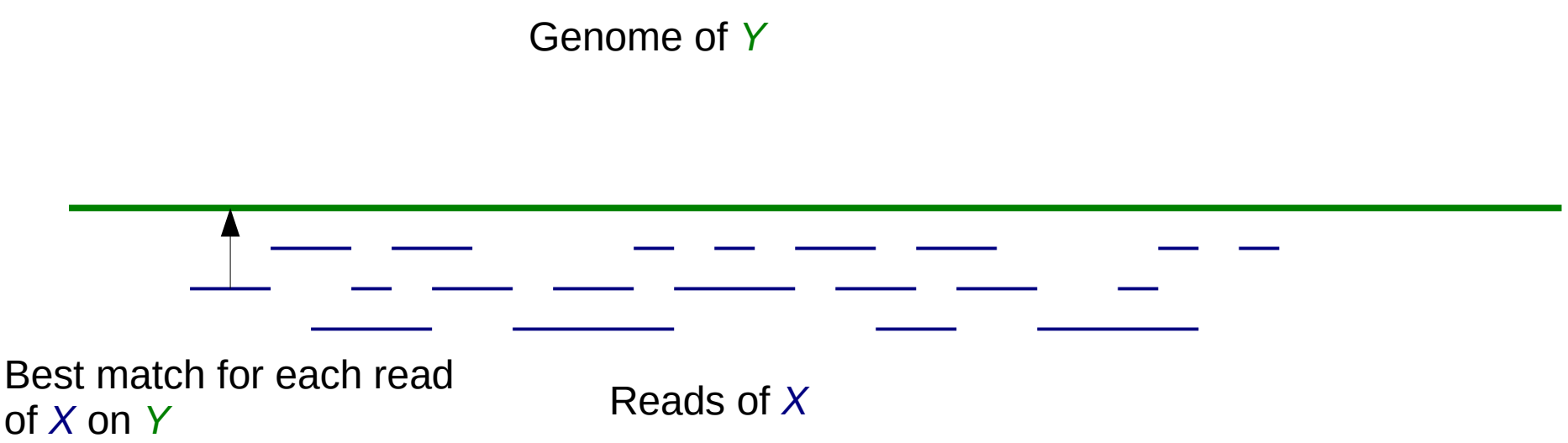
De novo versus *by reference* assembly

- 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



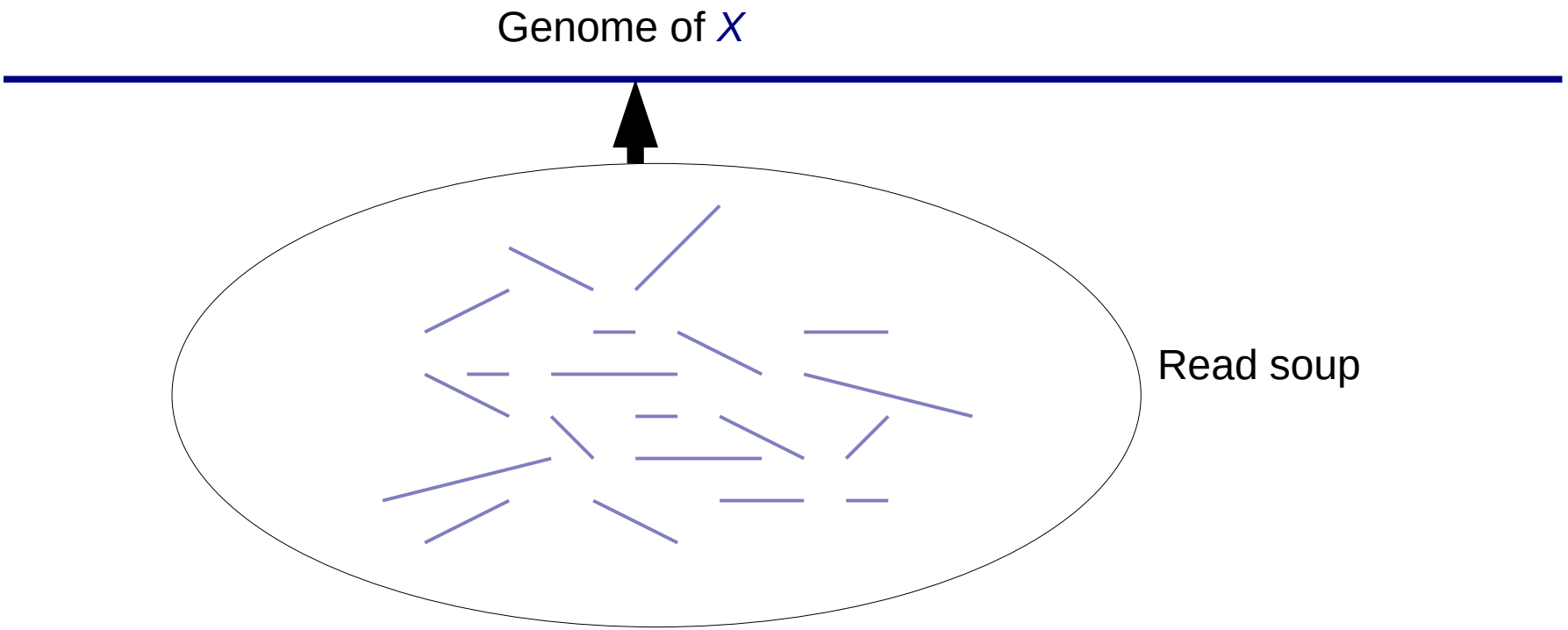
De novo versus *by reference* assembly

- 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

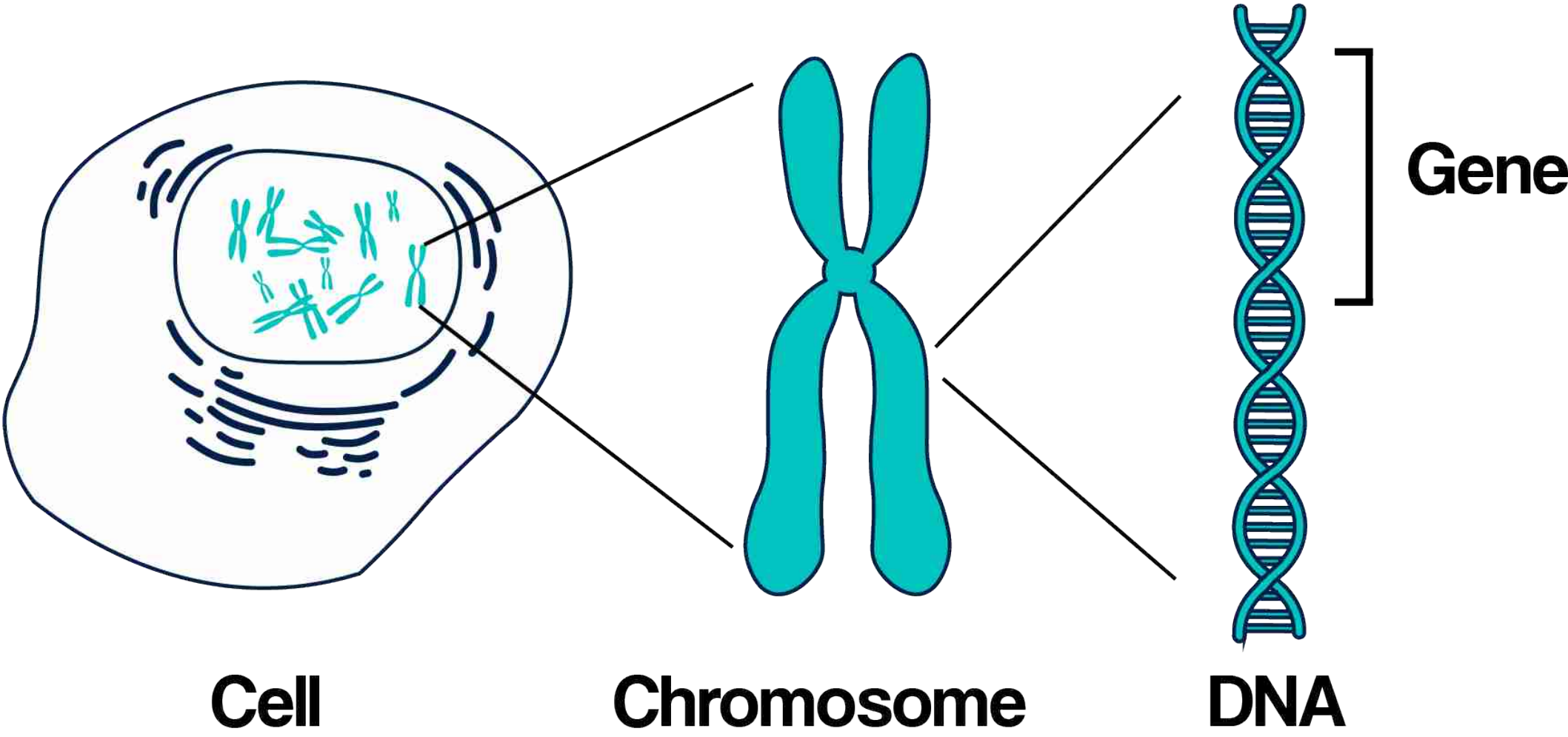
  → contains additional information

  → 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 previously

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)

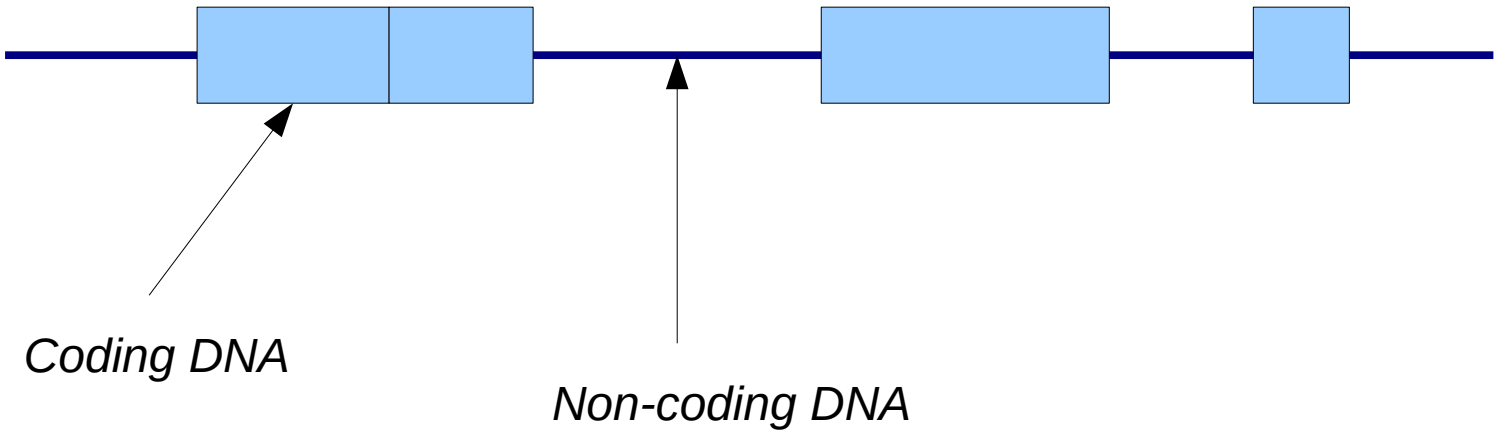
Translating DNA ↔ Protein data

- 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*

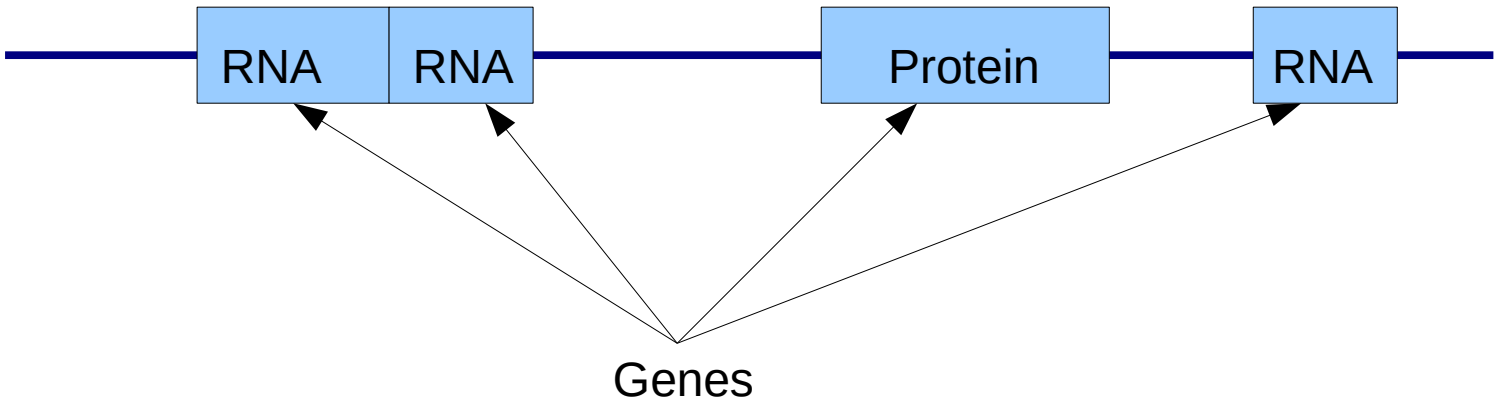
Top-level view

Chromosome: a long DNA molecule



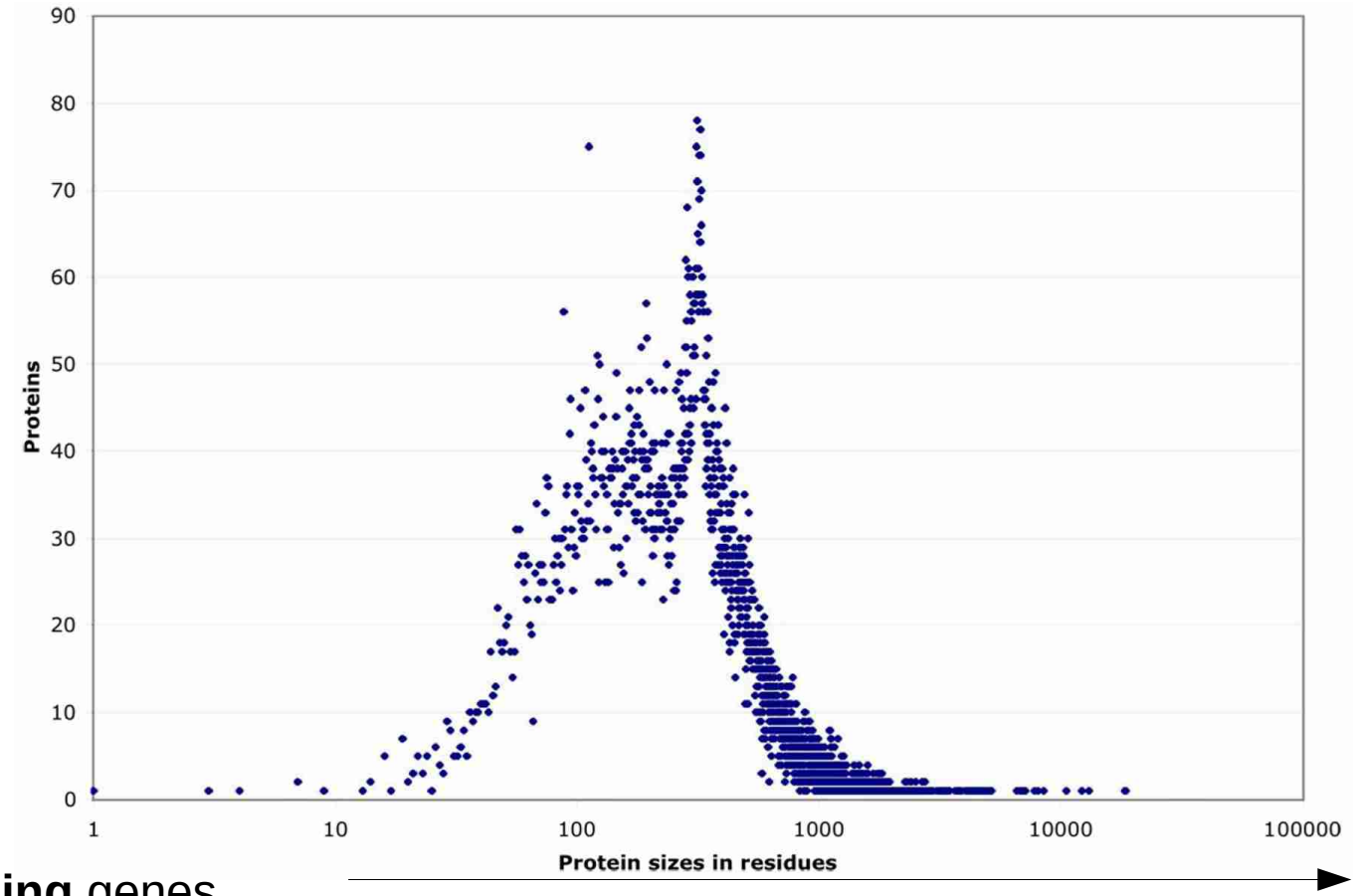
Top-level view

Chromosome: a long DNA molecule



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

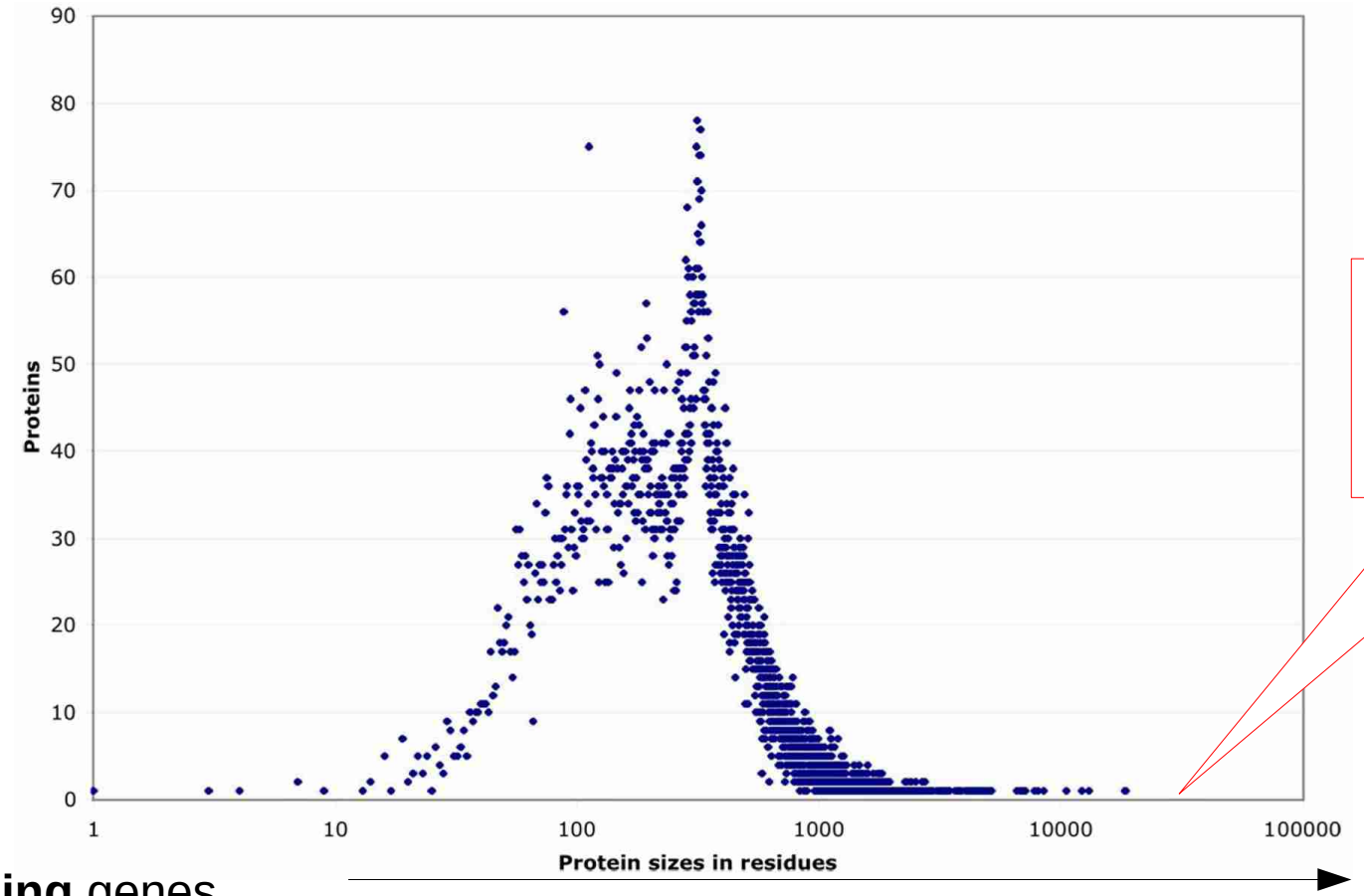
Average Protein gene Lengths



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



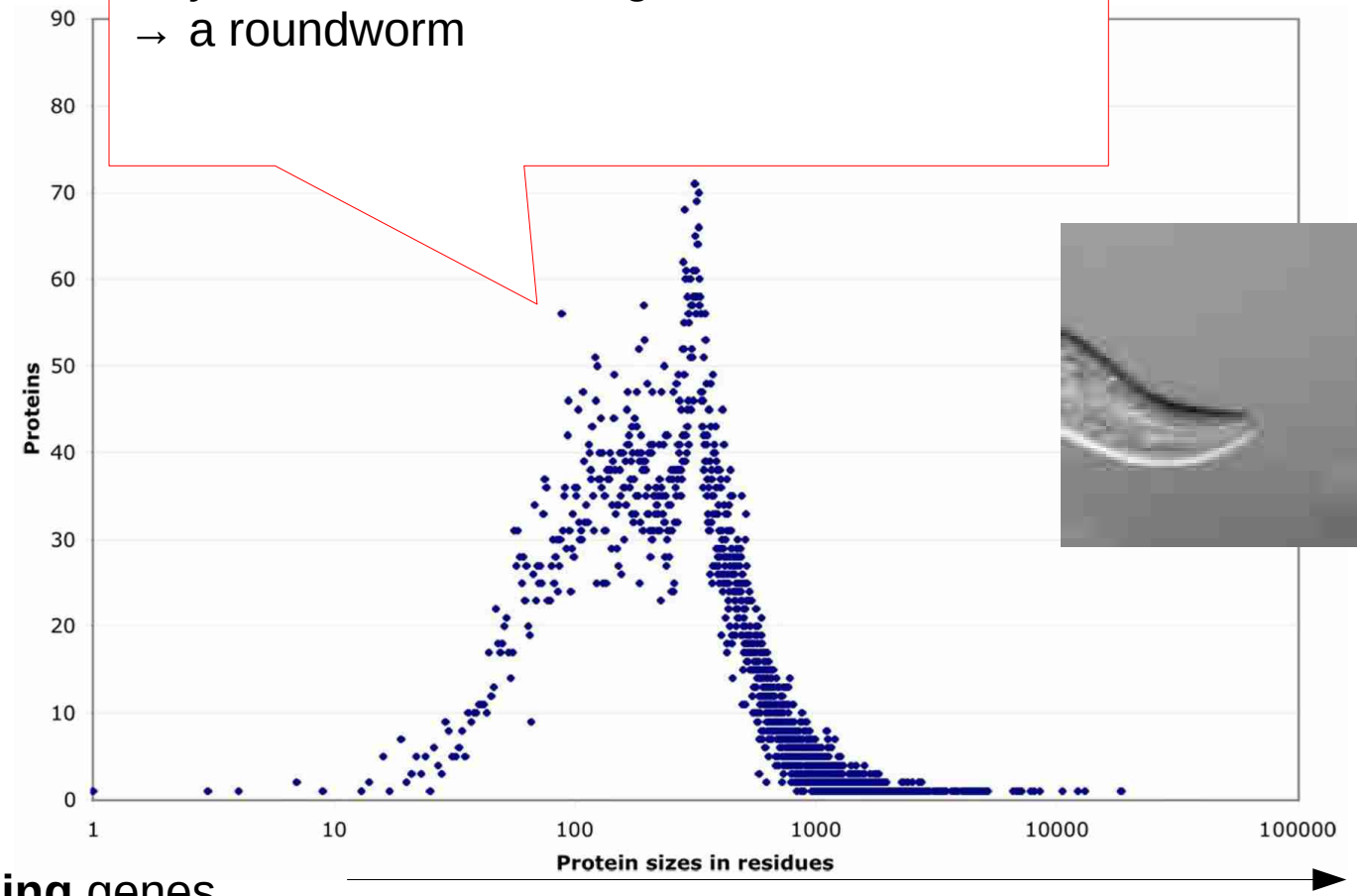
Logarithmic scale!

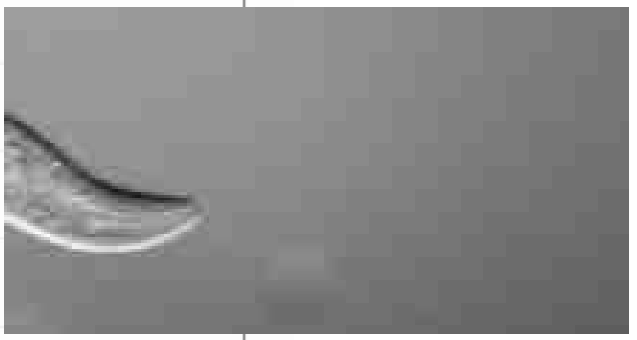
Number of **Protein-coding** genes

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

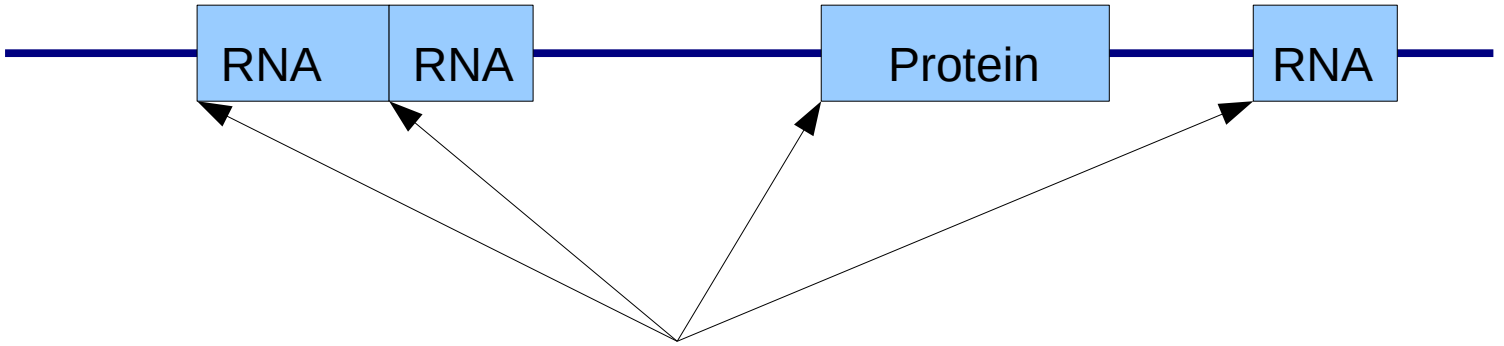




Number of **Protein-coding** genes

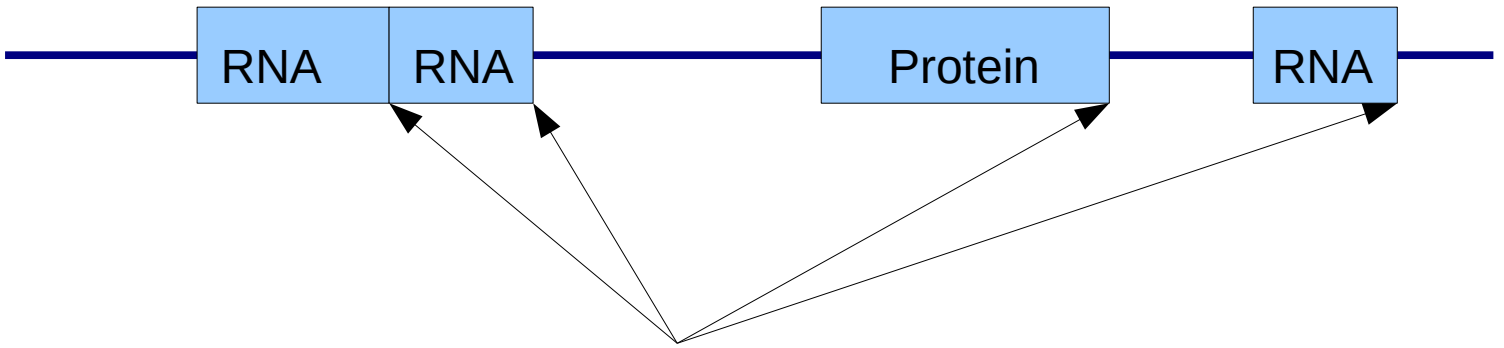
Protein sequence length → this is counted in # amino acid characters, not nucleotides, multiply by three to obtain DNA length!

Top-level view



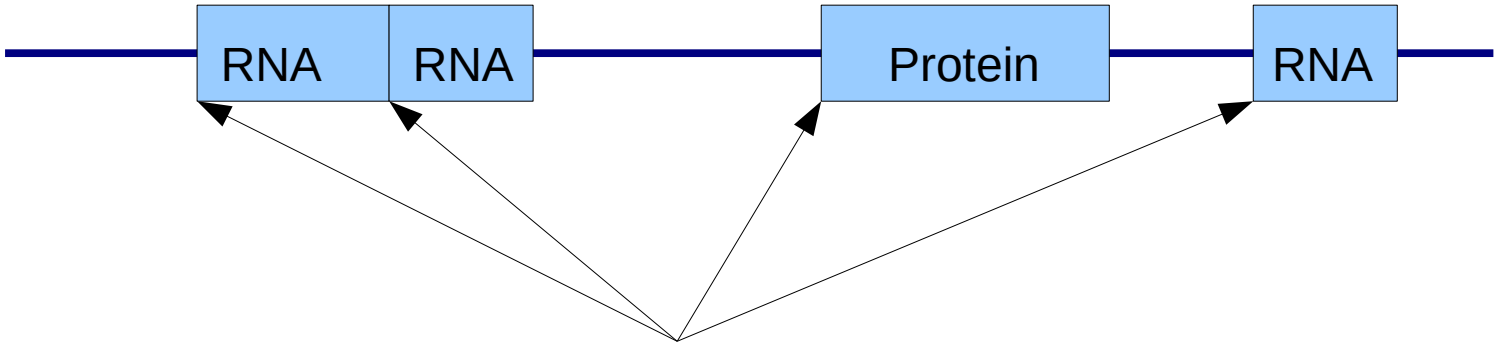
How do we know where genes start?

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	

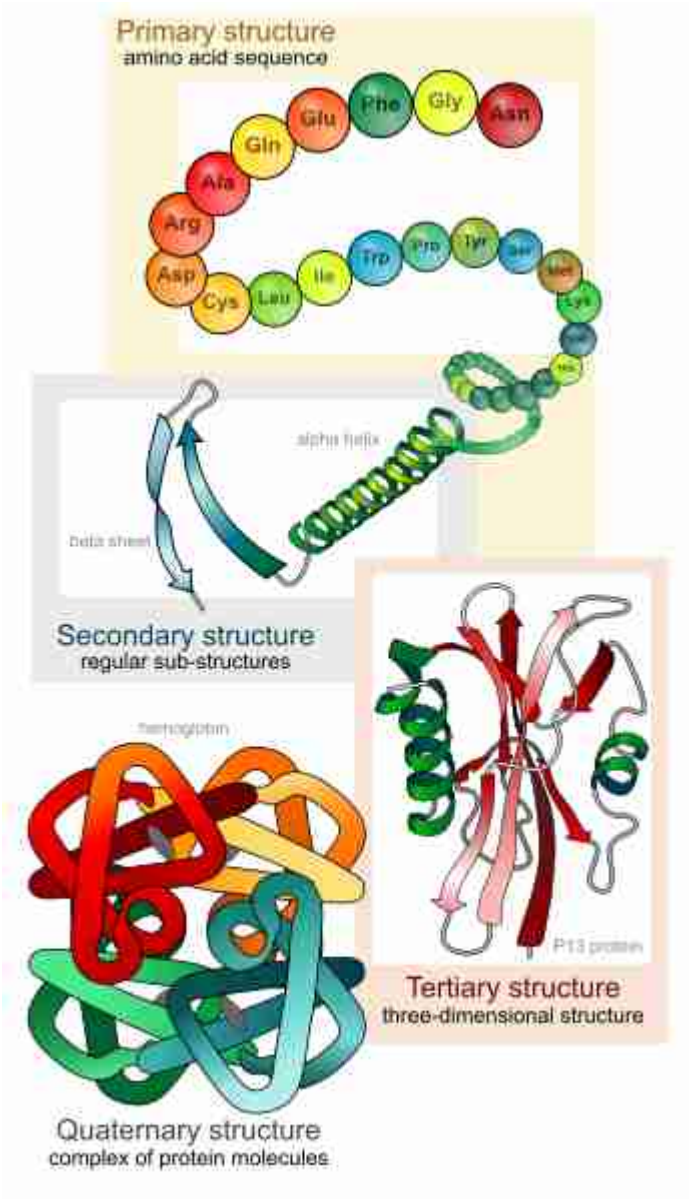
Amino acid	Codons	Compressed
Leu/L	TTA, TTG, CTT, CTC, CTA, CTG	YTR, CTN
Lys/K	AAA, AAG	AAR
Met/M	ATG	
Phe/F	TTT, TTC	TTY
Pro/P	CCT, CCC, CCA, CCG	CCN
Ser/S	TCT, TCC, TCA, TCG, AGT, AGC	TCN, AGY
Thr/T	ACT, ACC, ACA, ACG	ACN
Trp/W	TGG	
Tyr/Y	TAT, TAC	TAY
Val/V	GTT, GTC, GTA, GTG	GTN
STOP	TAA, TGA, TAG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

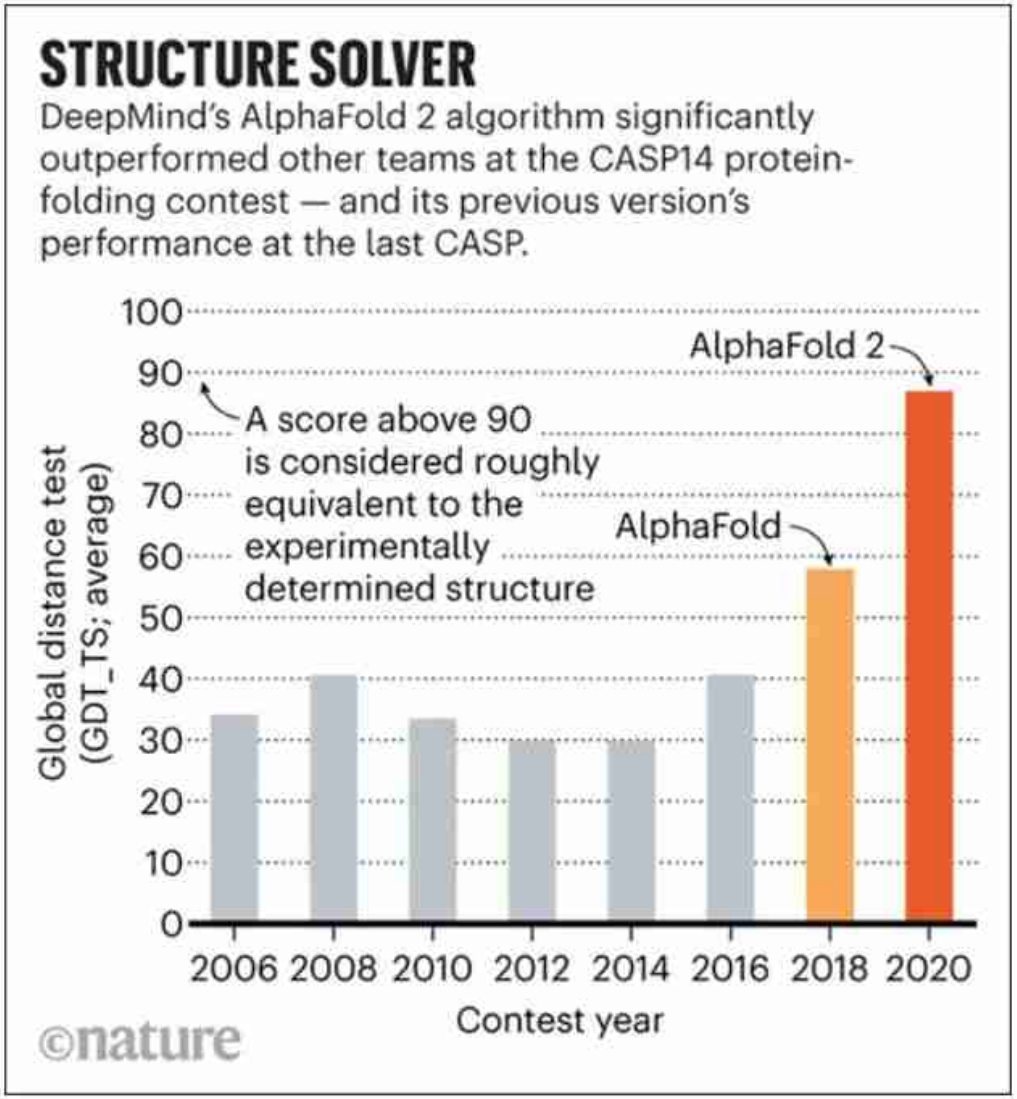
  1) Secondary
  2) Tertiary
  3) Quaternary

  structure!
- The structure determines the function/effect of a protein
- One would like to predict the structure from the protein sequence (primary structure)
- Used to be a challenging problem until AlphaFold came
- 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 (**CASP**) www.predictioncenter.org
- Blind testing and benchmarking of programs

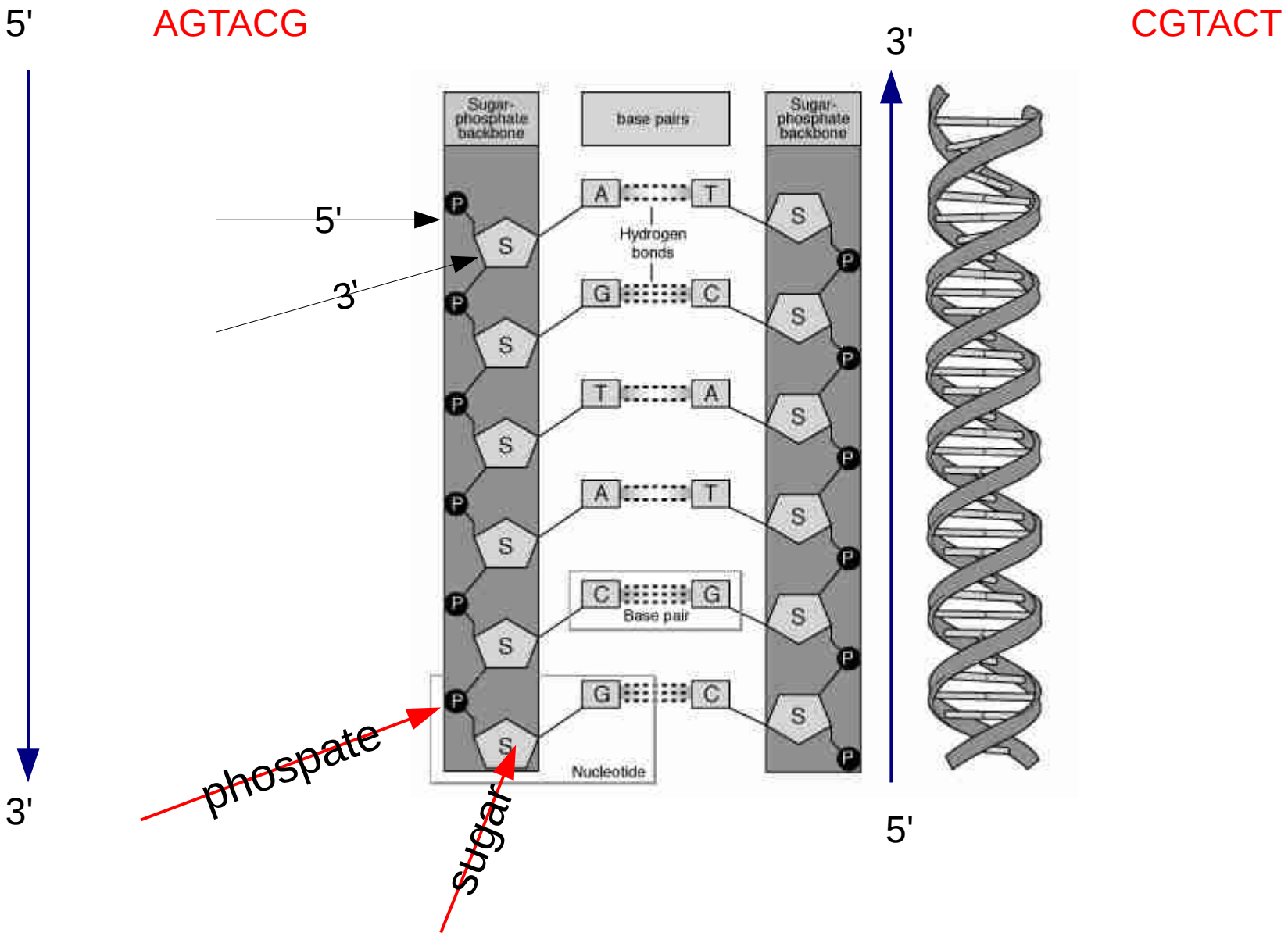
Alpha Fold


STRUCTURE SOLVER
DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.
Global distance test (GDT_TS; average)
100
90
80
70
60
50
40
30
20
10
0
A score above 90 is considered roughly equivalent to the experimentally determined structure
AlphaFold
AlphaFold 2
2006
2008
2010
2012
2014
2016
2018
2020
Contest year
©nature

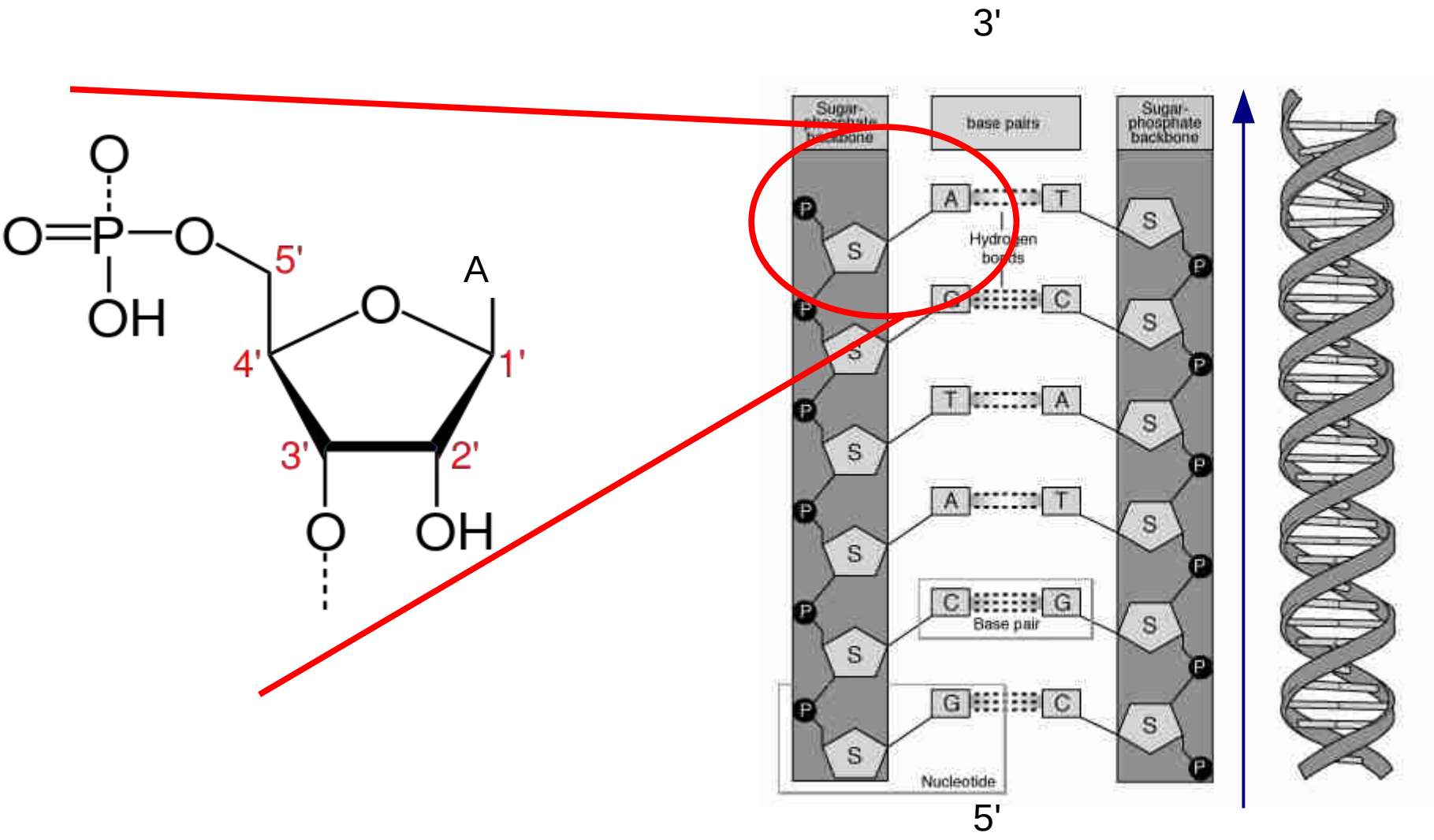
Another challenging problem

- 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'

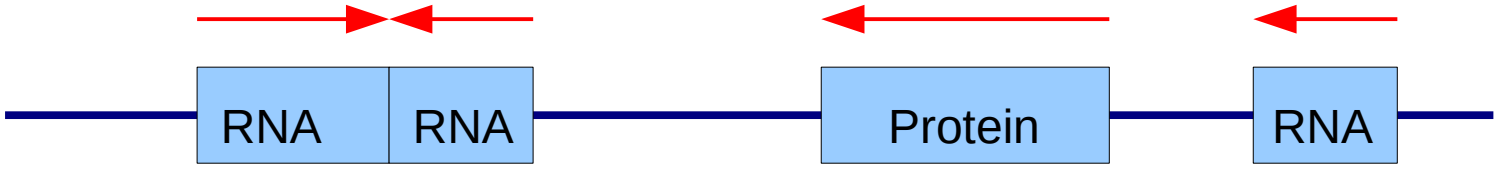
3' and 5'


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
C
G
Hydrogen bonds
Base pair
Nucleotide
P
S
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'*

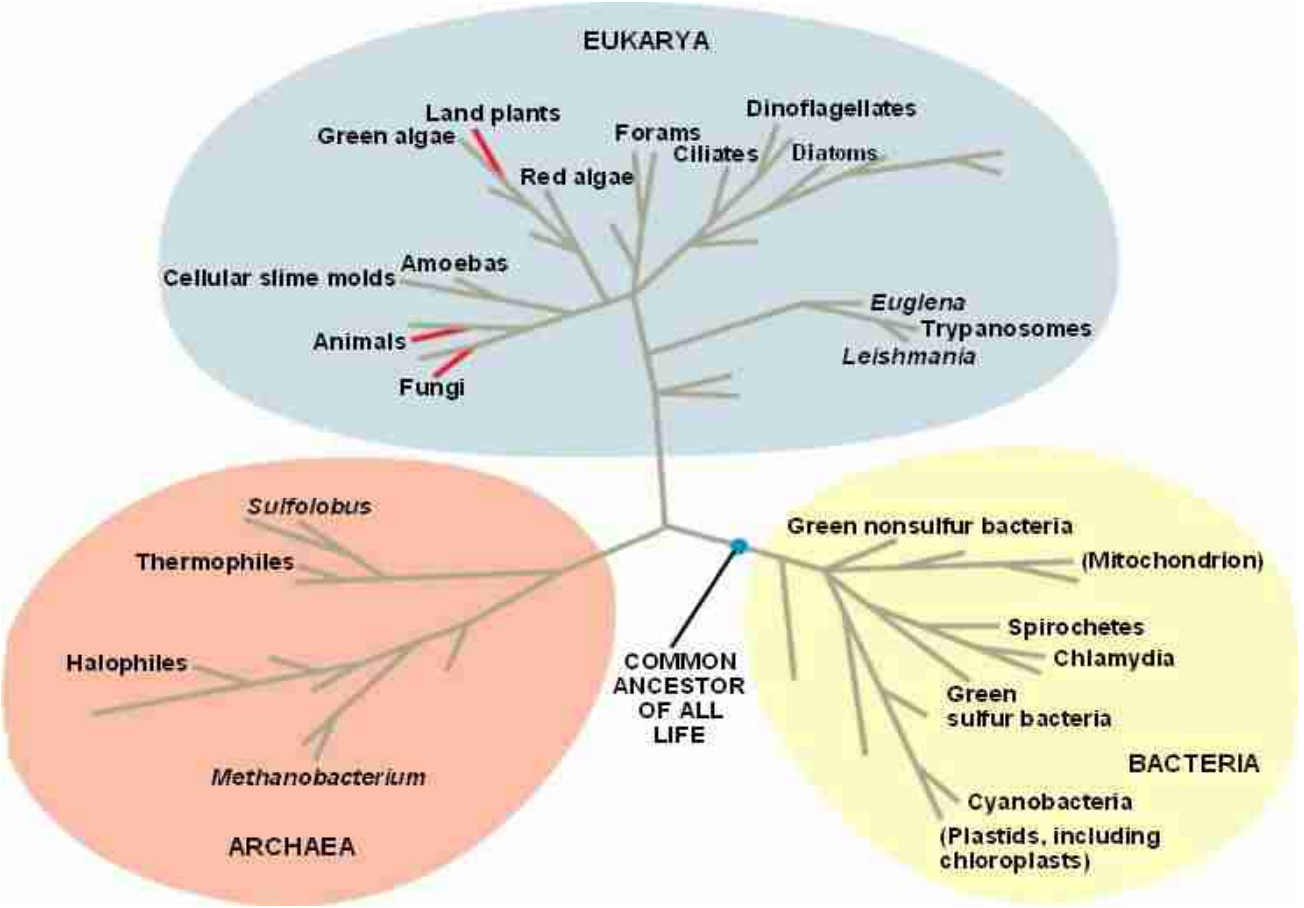
Top-level view



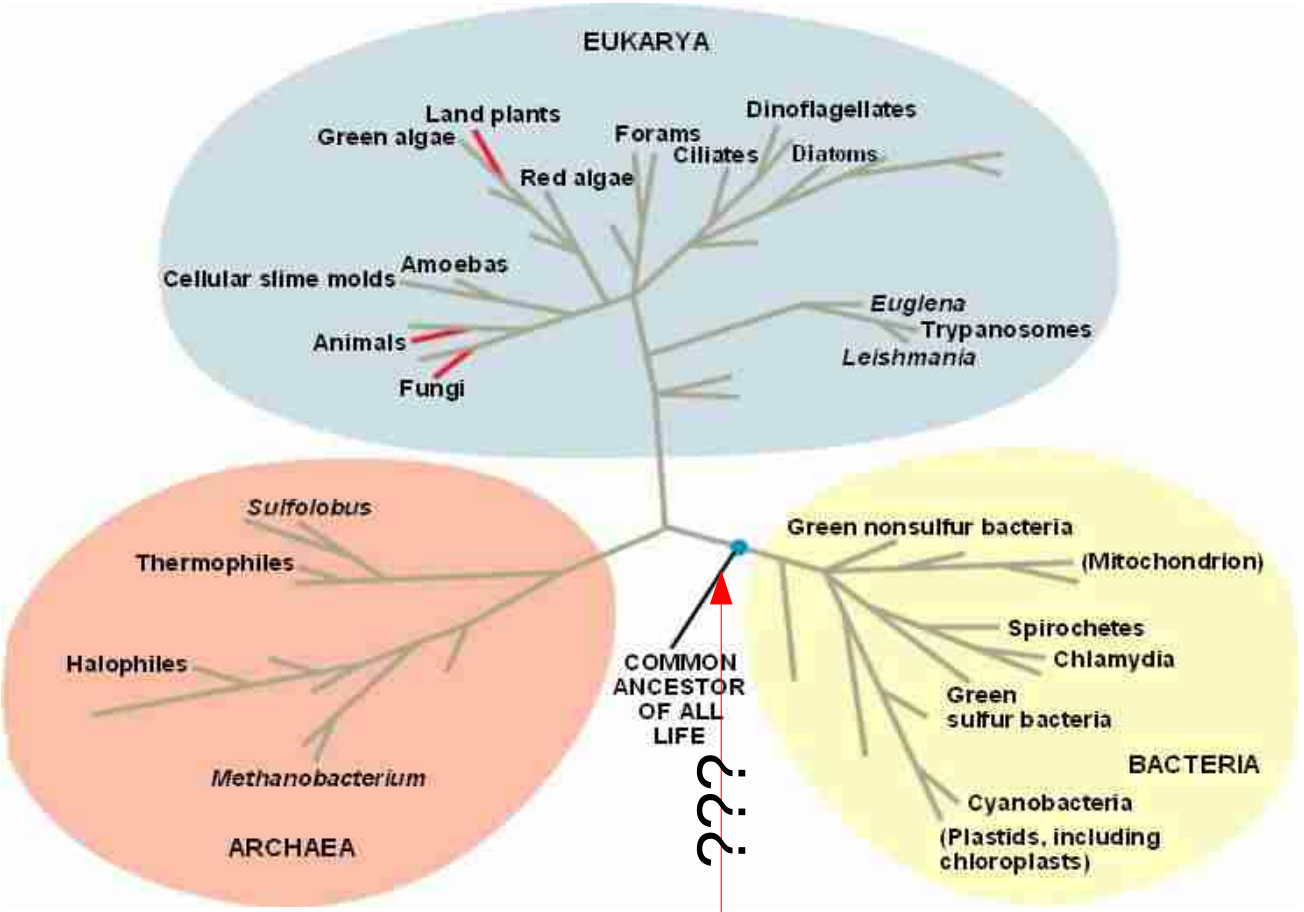
→ Genes have a direction!
→ depending on which strand of the double helix encodes the gene
They must be read from the correct side to be recognized!

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

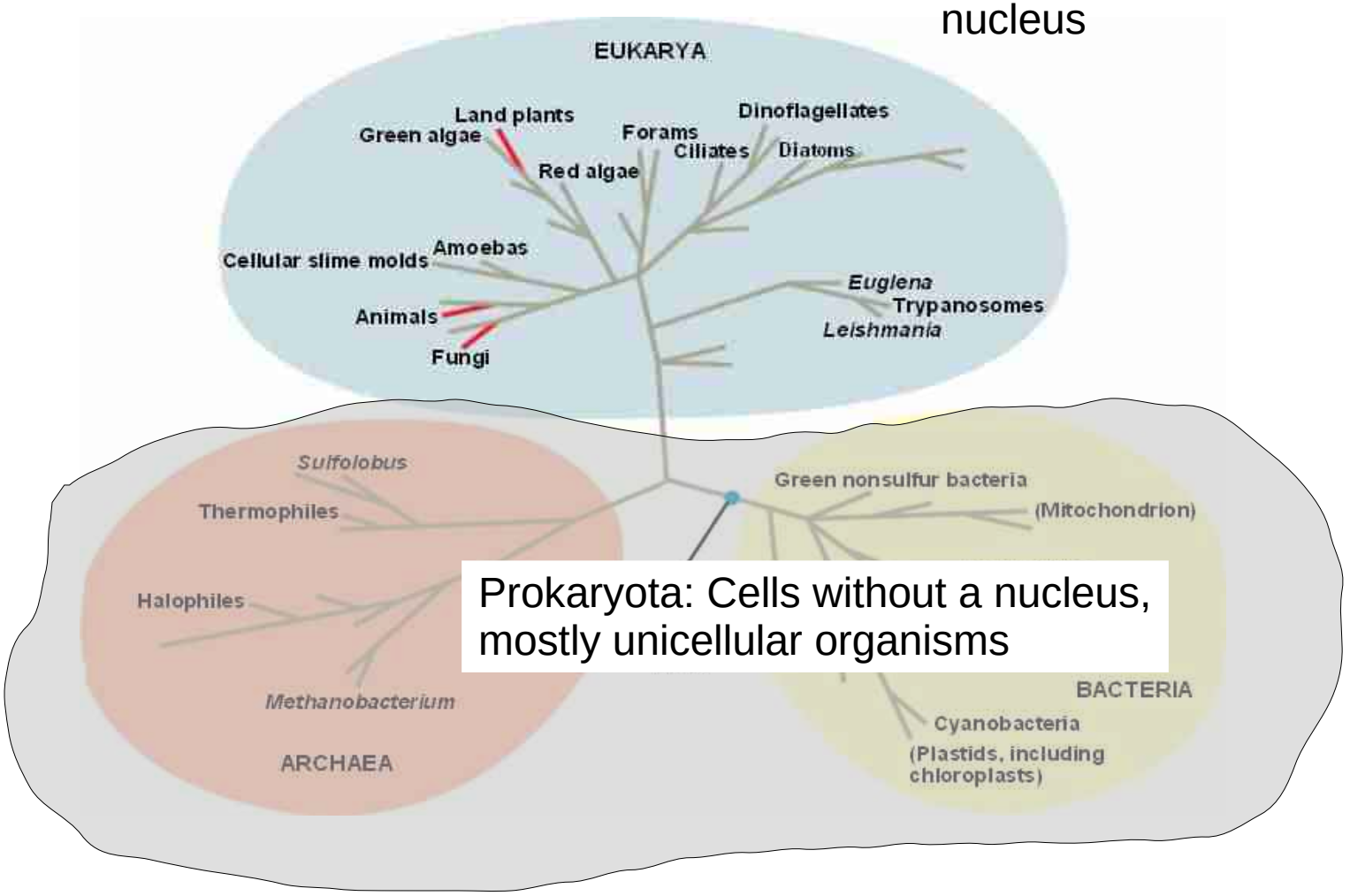


Salty environments
Hot environments

Where is the common ancestor?

The domains of life

Eukaryota: organisms with a cell nucleus



Prokaryota: Cells without a nucleus, mostly unicellular organisms

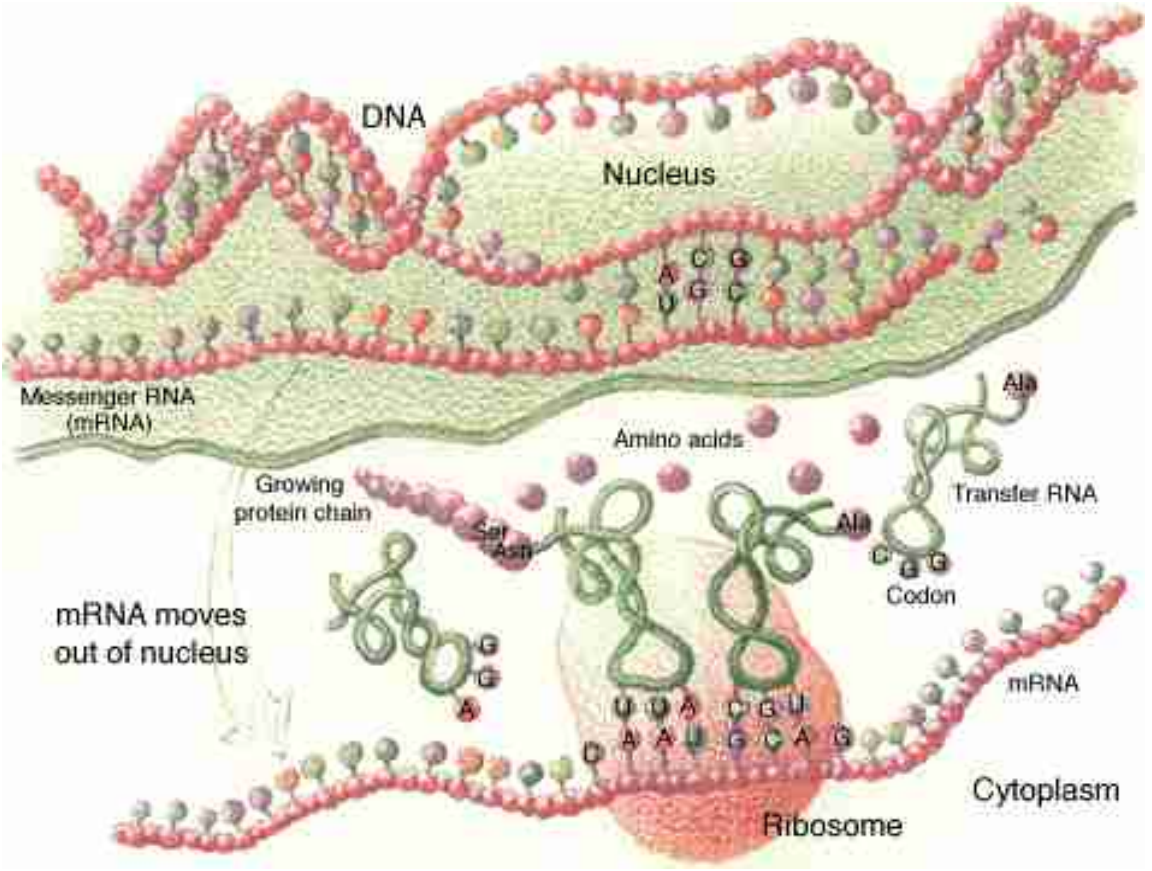
More about genes

- *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
Ala
Transfer RNA
Codon
mRNA moves
out of nucleus
mRNA
Cytoplasm
Ribosome

RNA

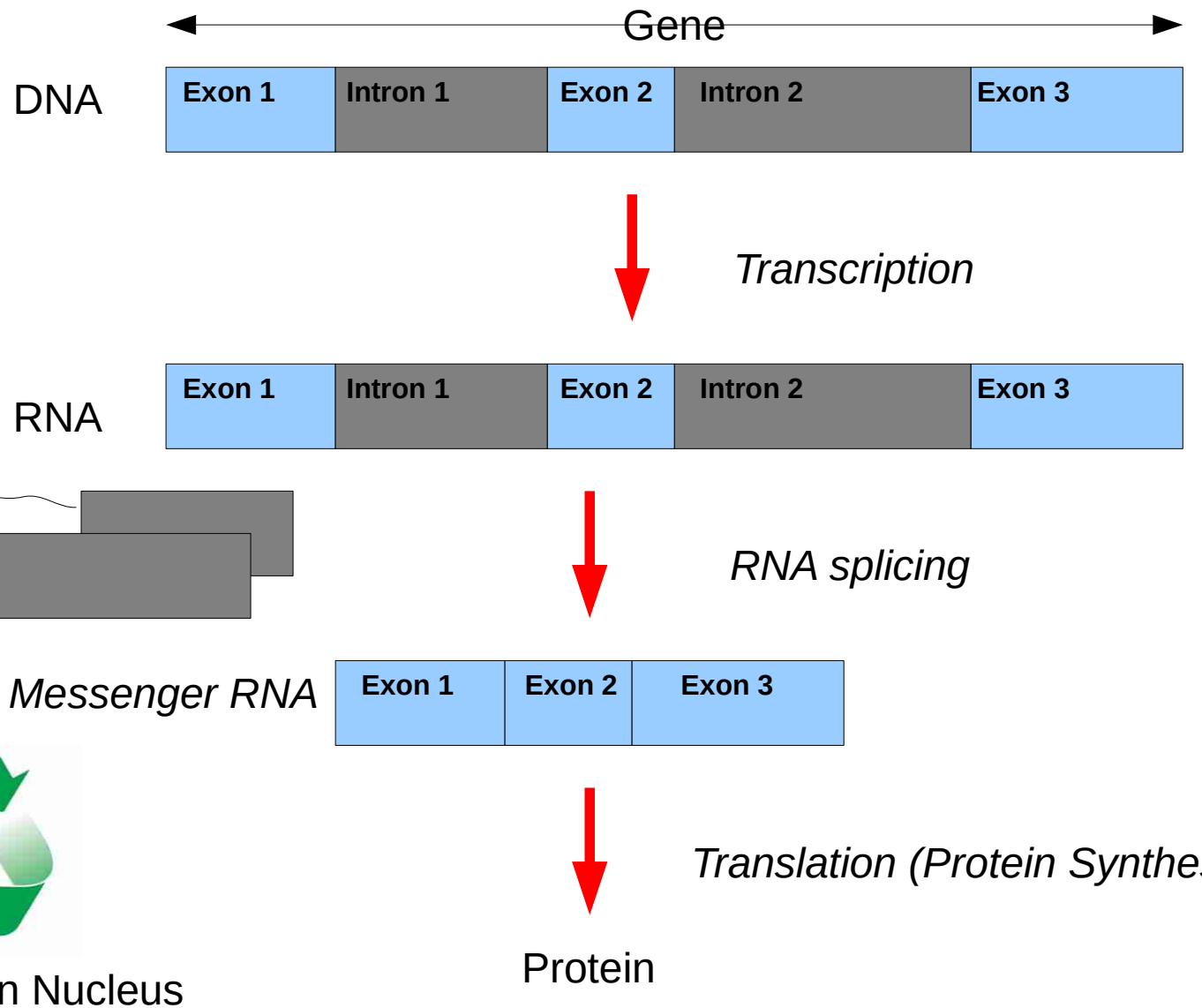
- RNA is involved in the process of DNA *Transcription*
- RNA is a copy of a coding DNA strand (a gene)
- And involved in the process of Transcription to construct either:

  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

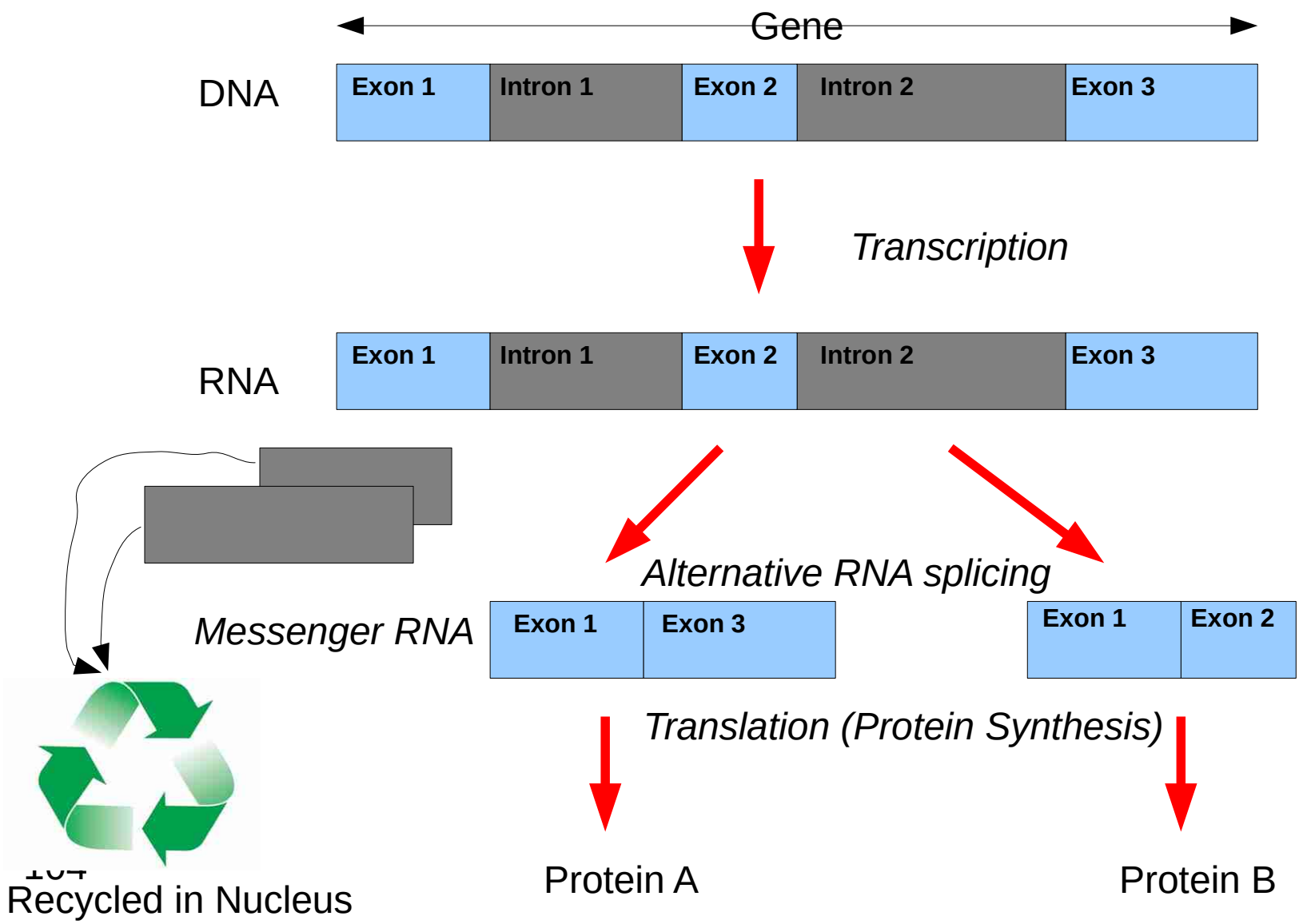
RNA Splicing *Eukaryota*



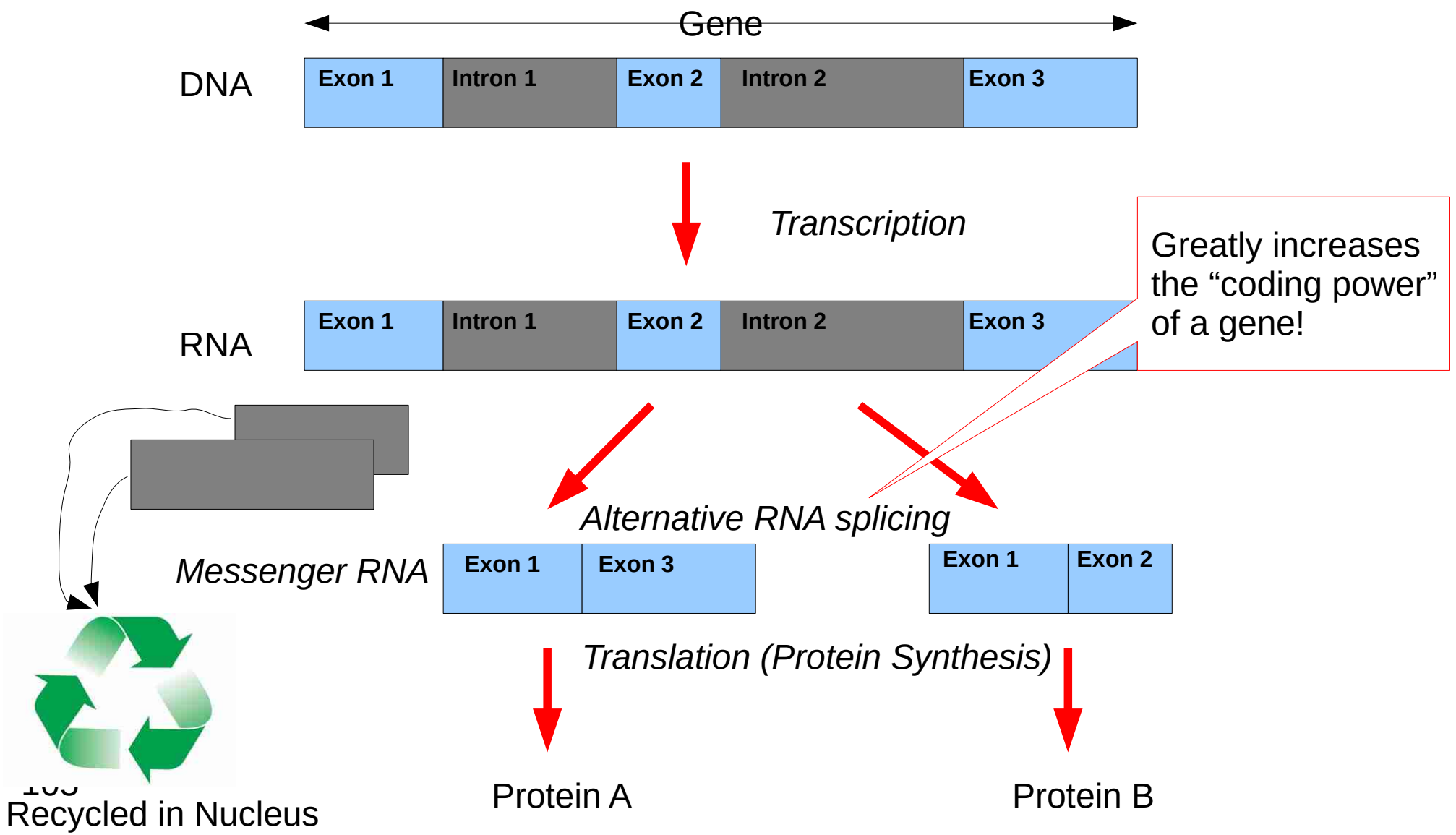
Eukaryotic RNA

- 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

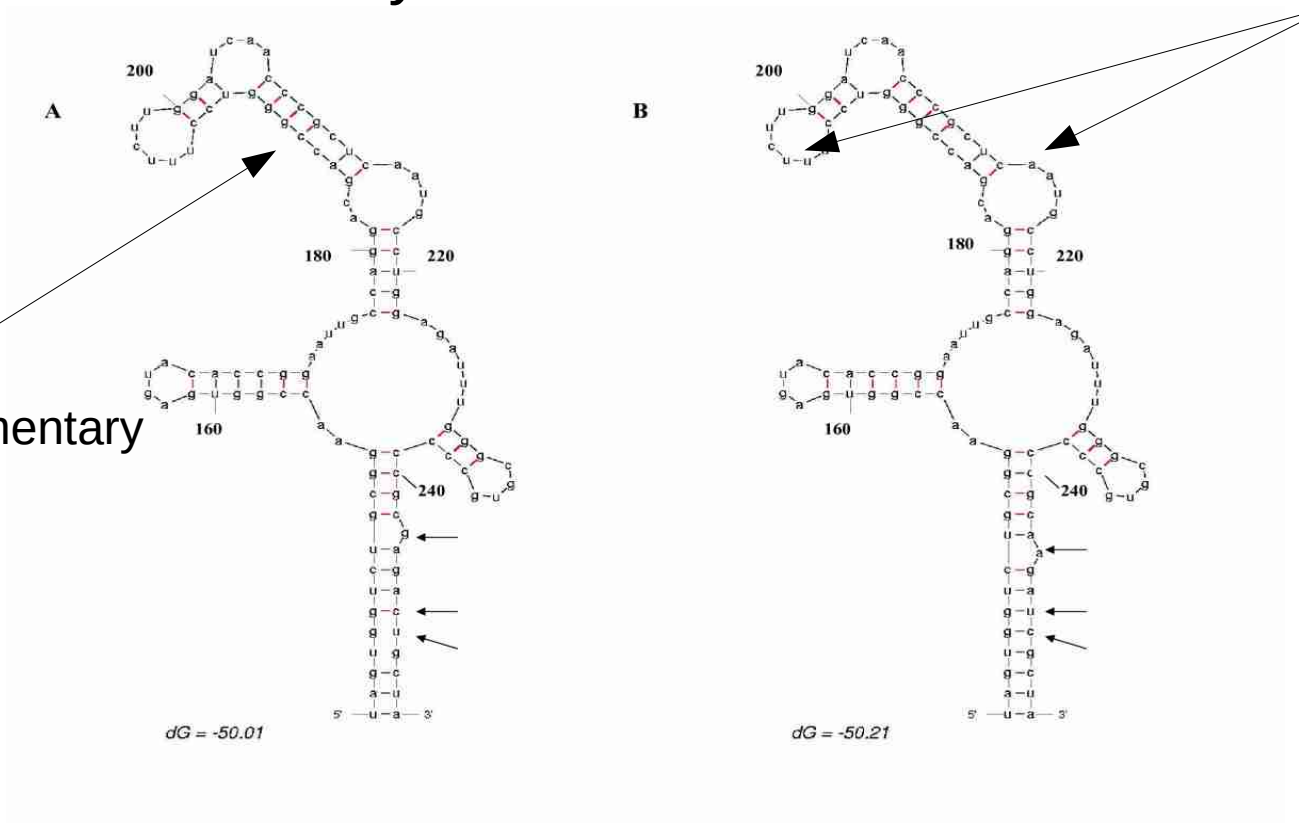
  → brings in the amino acids

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

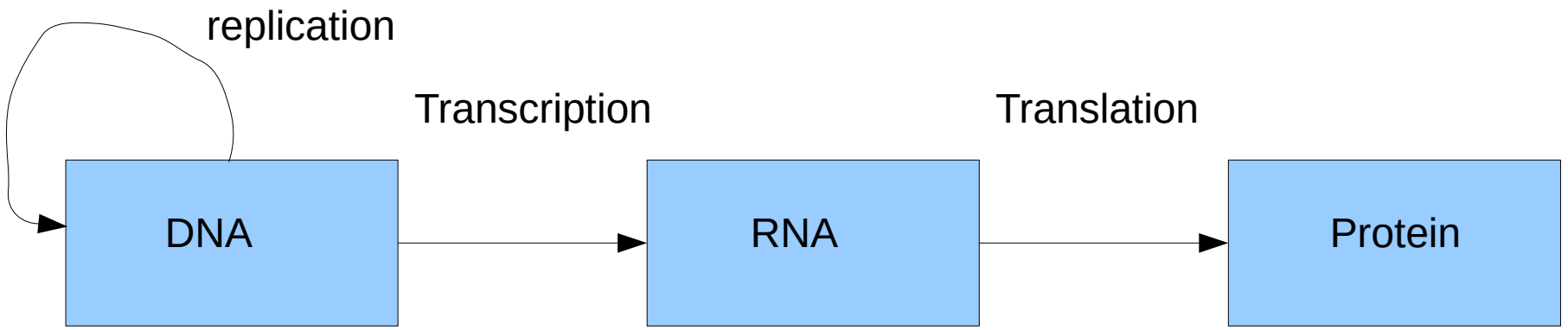
C ↔ G

RNA Secondary Structure

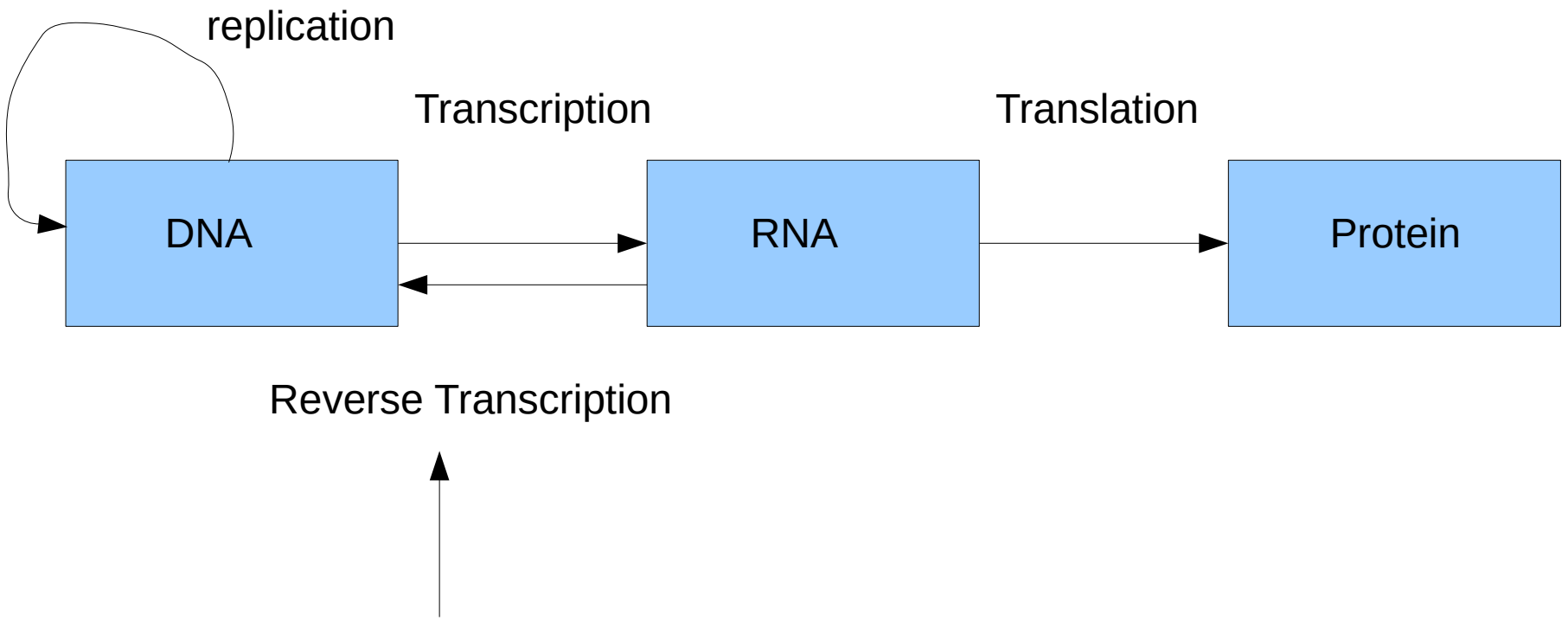
- Importance for RNA evolution

  → 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*?

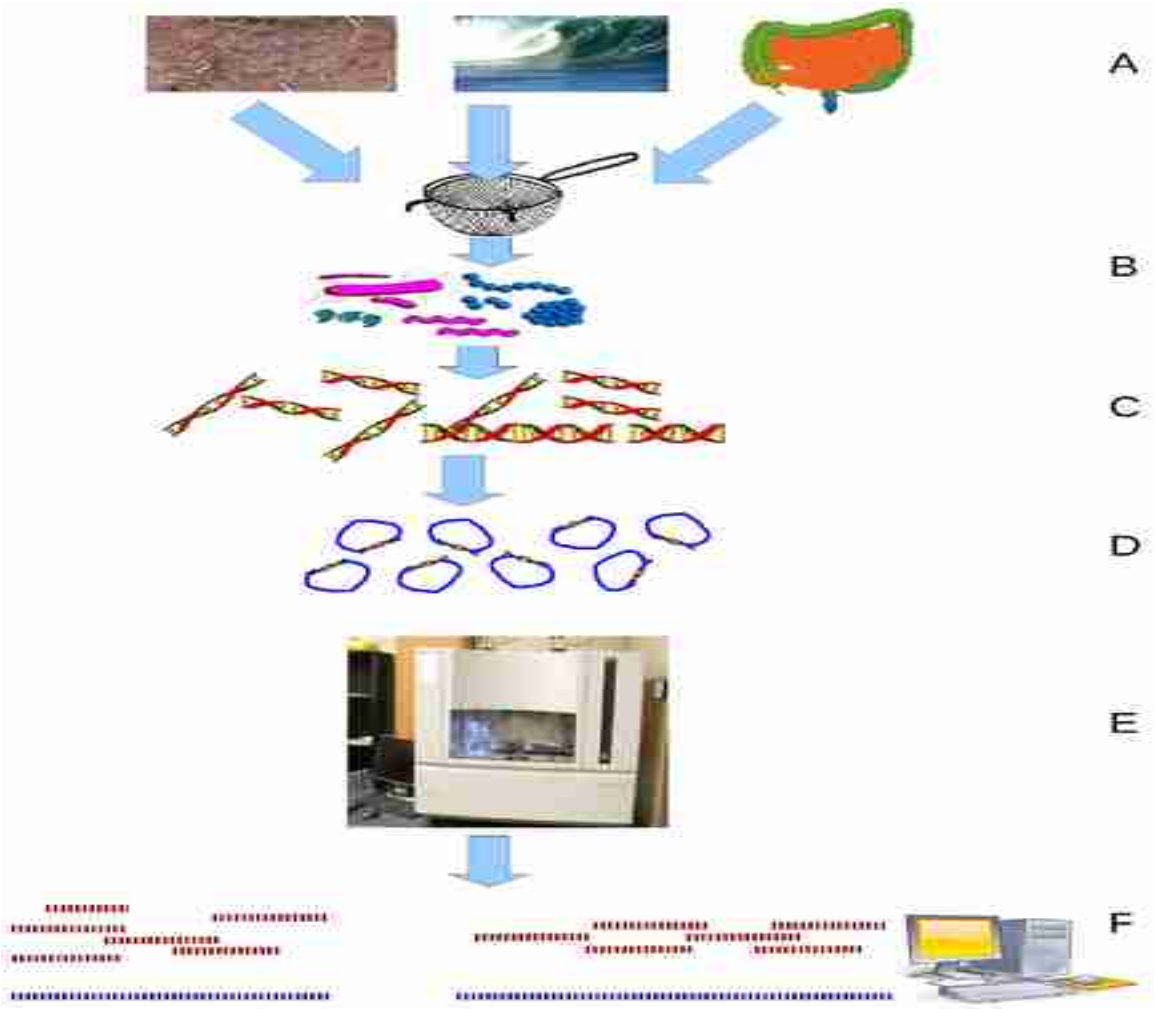
- The set/entirety of all RNA (mRNA, tRNA, rRNA) molecules in a cell
- In contrast to a genome, the transcriptome reflects the activity in a cell!

  → 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
- *1000* insect transcriptomes project 1KITE www.1kite.org

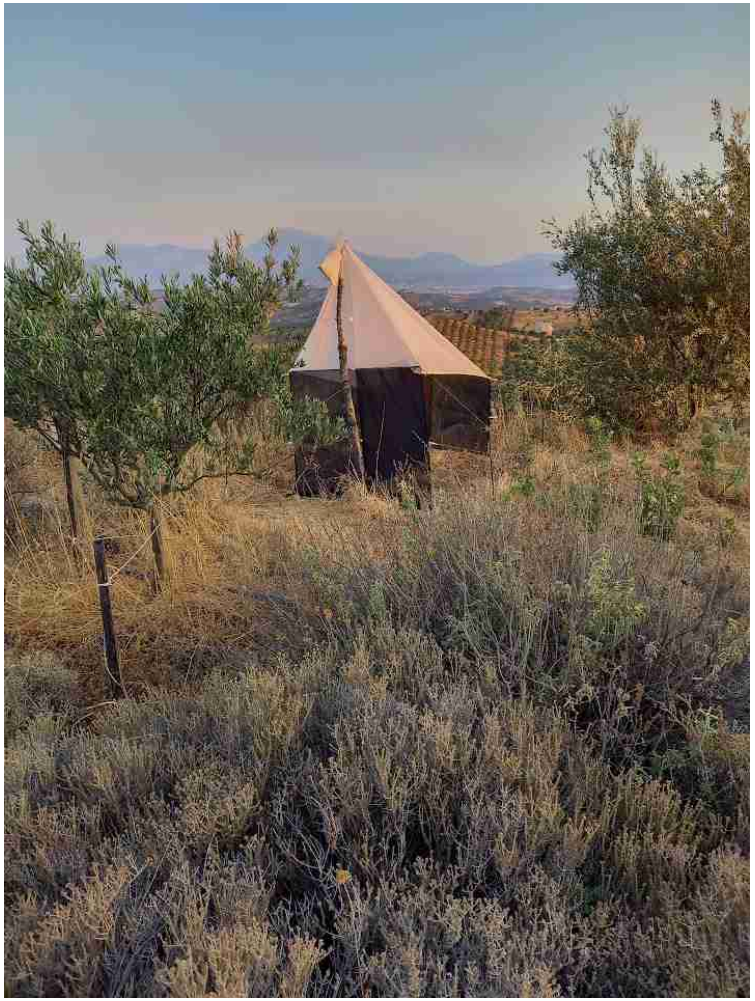
What is a *Meta-Genome*?



The Meta-Genome

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

Field Work
Insect Metagenetics



Malaise trap for insect biodiversity monitoring
→ the island of Crete is a Biodiversity hotspot
→ high levels of endemism

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 (paternal/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*

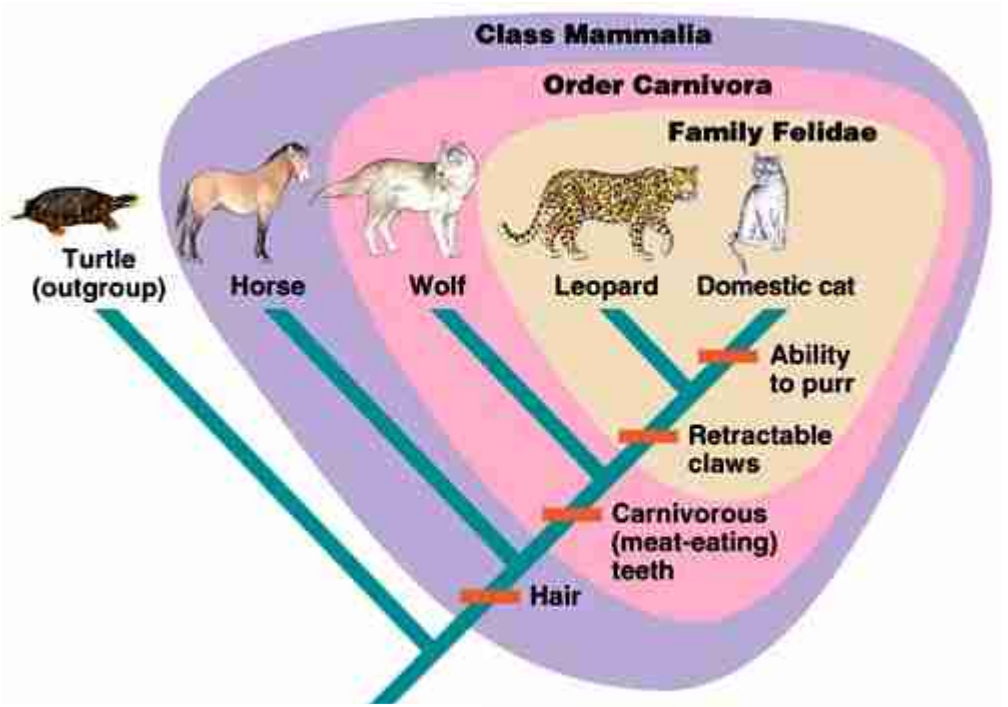
What's a species?

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

"Coalescent-based species delimitation in an integrative taxonomy"

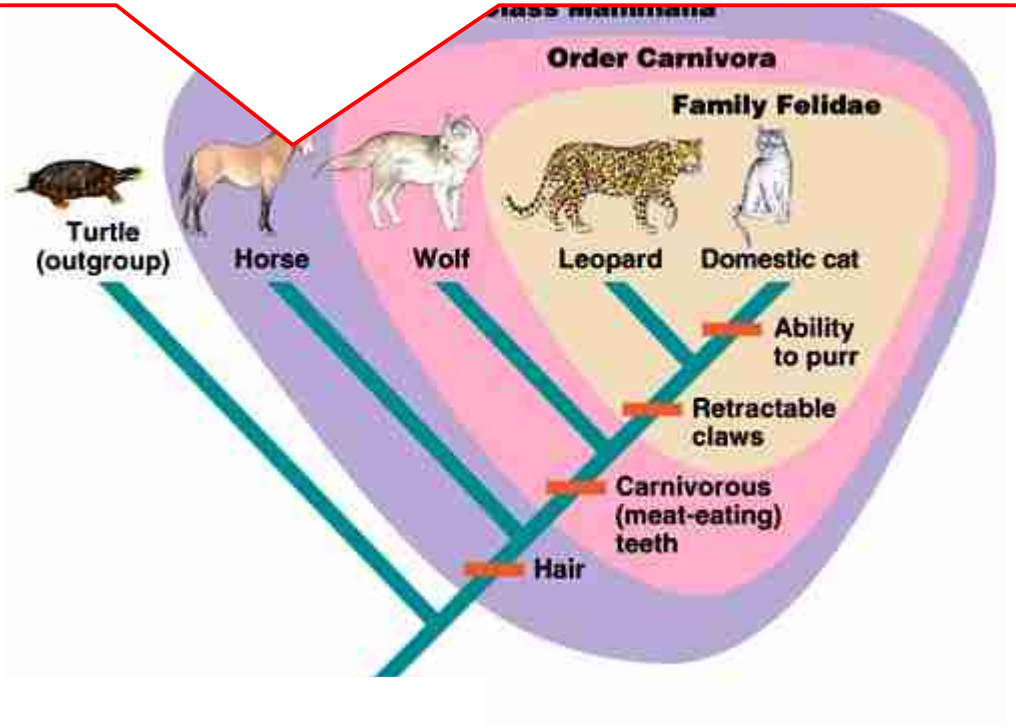
- Tricky que
- Different c
  - → gener
  - → more than *30* definitio
- By reproduction
  - → two species that can reproduce
  - → what about bacteria/viruses ????
- Evolutionary species concept
  - → via ancestral descent in an evolutionary tree
- General lineage (Abstammung/Verzweigung) concept
  - → 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

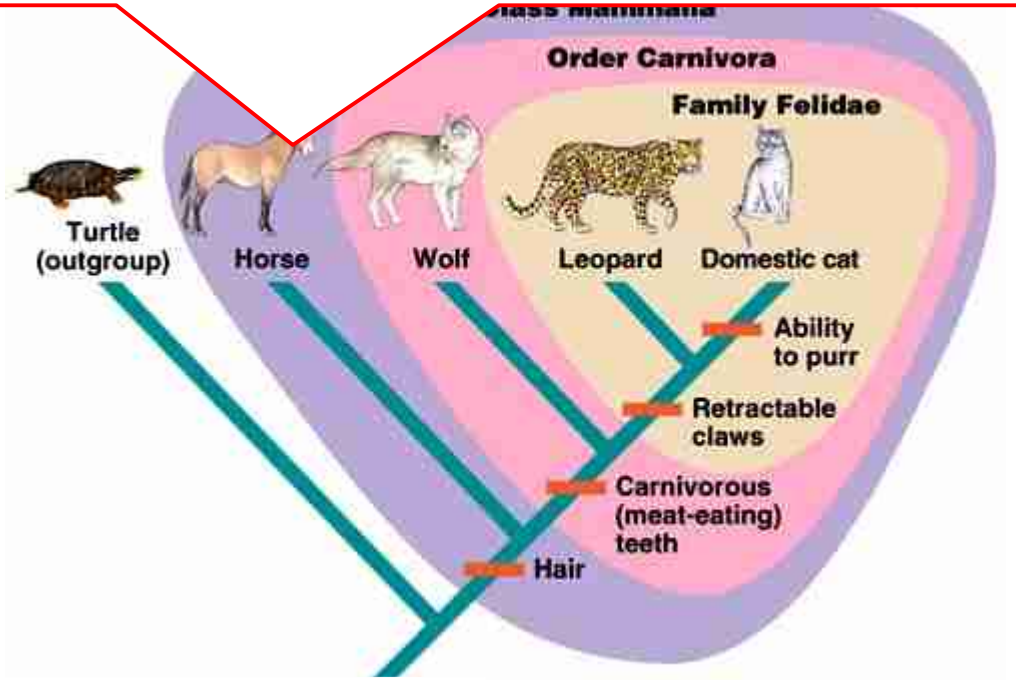
First systematic classification of living beings by Aristotele *384 -382* BC
Some terms still in use today, e.g., classification of animals into *Vertebrates* versus *Invertebrates*



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*

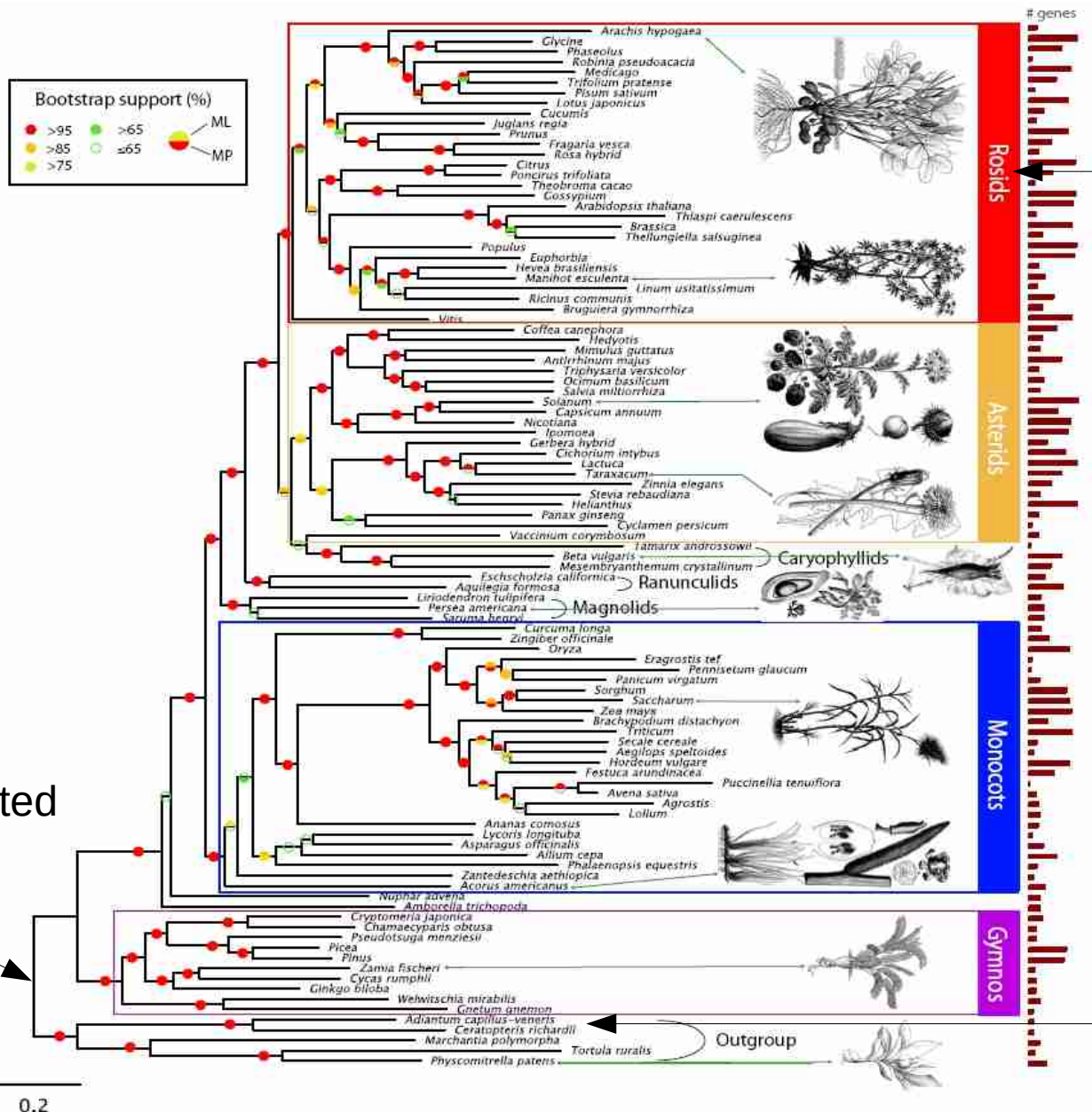
Wirbeltiere
Σπονδυλωτά



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

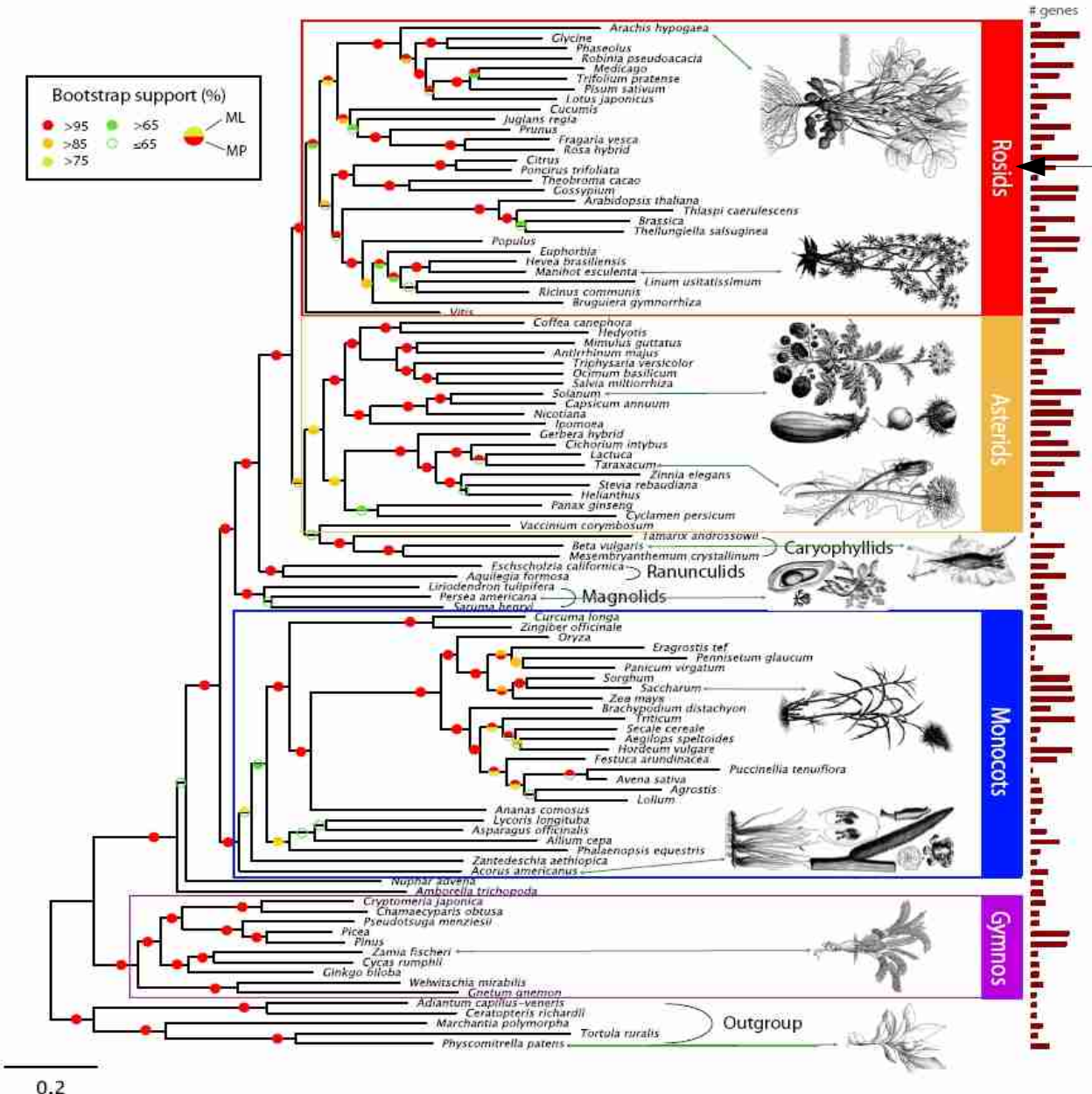
The ingroup

This tree is unrooted

124

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/εναπομείναντα” (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” – Ukranian and American evolutionary biologist Theodosius Dobzhansky

Next Lecture – Live at KIT

- Lukas Hübner
  - Comparing sequences computationally
  - Algorithms on strings of DNA
- Alexey Kozlov
  - The famous BLAST algorithm
  - Genome Assembly

Drop me an Email!

- Alexandros.Stamatakis@kit.edu

Backup Slide: *The Human Genome Project*

- 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