Bioinformatics — Lecture 1
Basics of molecular biology and genetics (MM Ch. 1–3)

Krzysztof Bartoszek

Linköping University

krzysztof.bartoszek@liu.se

11 XI 2024 (R36)

Today

Course introduction

DNA, RNA, proteins

Passing on genetic material

The data

Course contents

What is BIOINFORMATICS?

Computational biology? Biomathematics? Biostatistics?

1. Basics of molecular biology and genetics
2. Hidden Markov models, genetic sequence analysis
3. Sequence similarity, sequence alignment
4. Phylogeny reconstruction
5. Quantitative trait modelling (phylogenetic comparative methods)
6. Microarray analysis
7. Network biology

Teaching staff for course

Me: Krzysztof Bartoszek, background

1. MEng in Computer Science, Gdańsk Univ. of Technology 2007
2. MPhil in Computational Biology, Univ. of Cambridge 2008
3. PhD in Statistics, Univ. of Gothenburg 2013
4. Postdoc, Dept. Mathematics Uppsala Univ. 2013–2017
5. Lecturer, STIMA LiU 2017–

Ying Luo

1. Labs
2. Grading
3. Support

Evaluation of 2023 course

1. about 10 students took the course in 2023.
2. 2 submitted an evaluation.
3. Course grade: 4 ± 1.41.
4. Changes for 2024.
   - 4.1 First lab restructured to be open source.
   - 4.2 New teaching assistant.
   - 4.3 Various errors and typos in lectures, labs and exercises fixed.

Course outline

- 7 lectures
- 5 computer labs
- 2 exercise sessions: HMMs, Sequence analysis, Phylogenetics, PCMs (SDEs)

Course literature

MAIN:

EG W.J. Ewens, G.R. Grant. Statistical Methods in Bioinformatics, 2nd ed., New York, 2005. Springer.

MM J. Momand, A. McCurdy. Concepts in Bioinformatics and Genomics, Oxford, 2017. Oxford University Press.

AUXILLARY

More biological

L A.M. Lesk. Introduction to Bioinformatics, Oxford, 2014. Oxford University Press. (focused on proteins)

More mathematical

CHL M. Crochemore, C. Hancart, T. Lecroq. Algorithms on Strings, Cambridge, 2007, Cambridge University Press.

D R. Durrett. Probability Models for DNA Sequence Evolution, 2008, Springer .

DEKM R. Durbin, S. Eddy, A. Krogh, G. Mitchison. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge, 1998, Cambridge University Press.

BE M. Borodovsky, S. Ekisheva. Problems and Solutions in Biological Sequence Analysis, Cambridge, 2006, Cambridge University Press.

Examination

The examination consists of

- Computer labs (need to be passed)
- a written and/or computer exam with max score 20 points and grade limits:
  A : 18p, B: 16p, C: 14p, D: 12p, E: 10p.
  Allowed aids for exam: will be decided on at a later stage
  Materials distributed with exam: will be decided on at a later stage

Bonus points

Active participation in the exercise sessions gives maximum 4 bonus points to the exam.

Active participation means that a student comes prepared to the seminar session with the given day's exercises, correctly solves an exercise on the board, is able to answer questions about the presented solution and is able to give help and comments to the classmates' presented solutions

In the seminars, for each exercise a student will be selected (how depends on the number of students) to present her/his solution.

Course homepage and materials

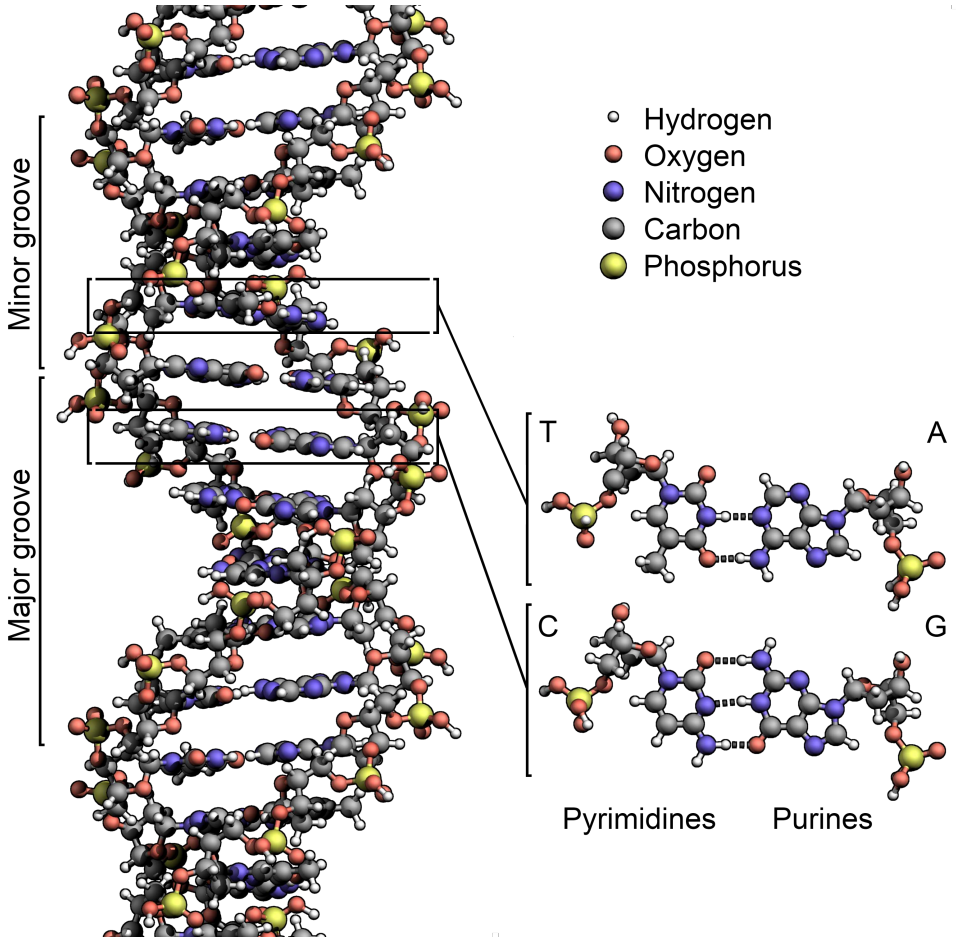
`https://www.ida.liu.se/~732A51/`:
messages and course information (select English)

LISAM (materials and submissions)

Course overview

1. Basics of molecular biology and genetics
2. Genetic sequence analysis (Hidden Markov Models)
3. Sequence similarity and alignments
4. Phylogeny reconstruction
5. Phylogenetic comparative methods
6. Mircorray analysis
7. Network biology

DNA: Deoxyribonucleic acid, double helix

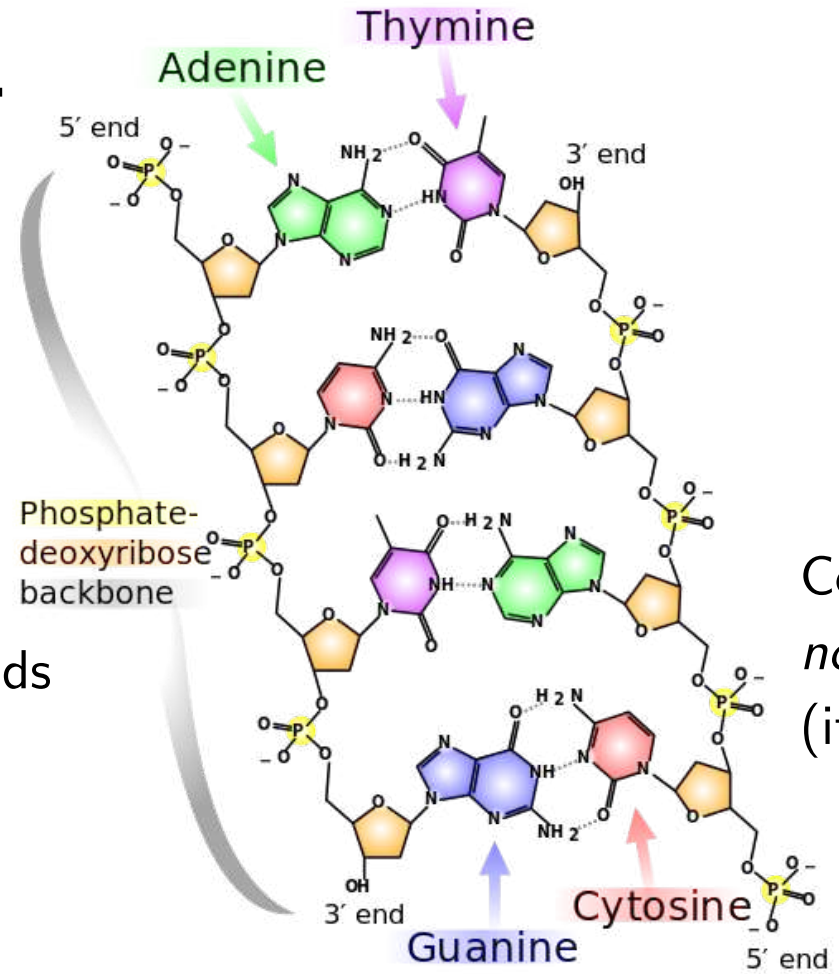


By Zephyris—Own work, CC BY-SA 3.0/GNU Free Documentation License, https://commons.wikimedia.org/w/index.php?curid=15027555

DNA: four base–pairs

$$A \leftrightarrow T$$

$$G \leftrightarrow C$$



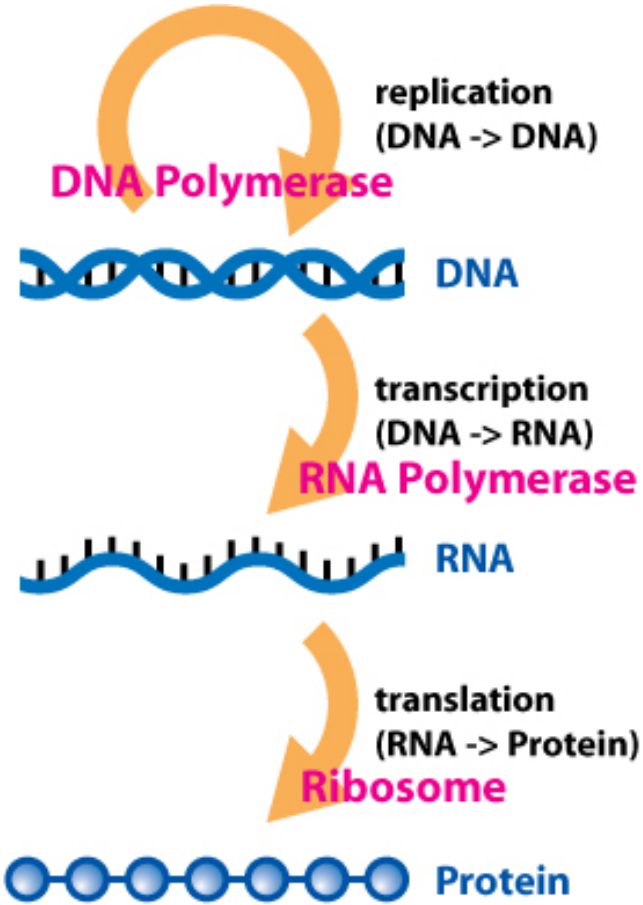
Coding strand corresponds to RNA transcript, 5′ to 3′ direction (if gene)

Complimentary is called *noncoding/template strand* (if gene)

https://en.wikipedia.org/wiki/File:DNA_chemical_structure.svg (graphic by Madeleine Price Ball—CC0 public domain)

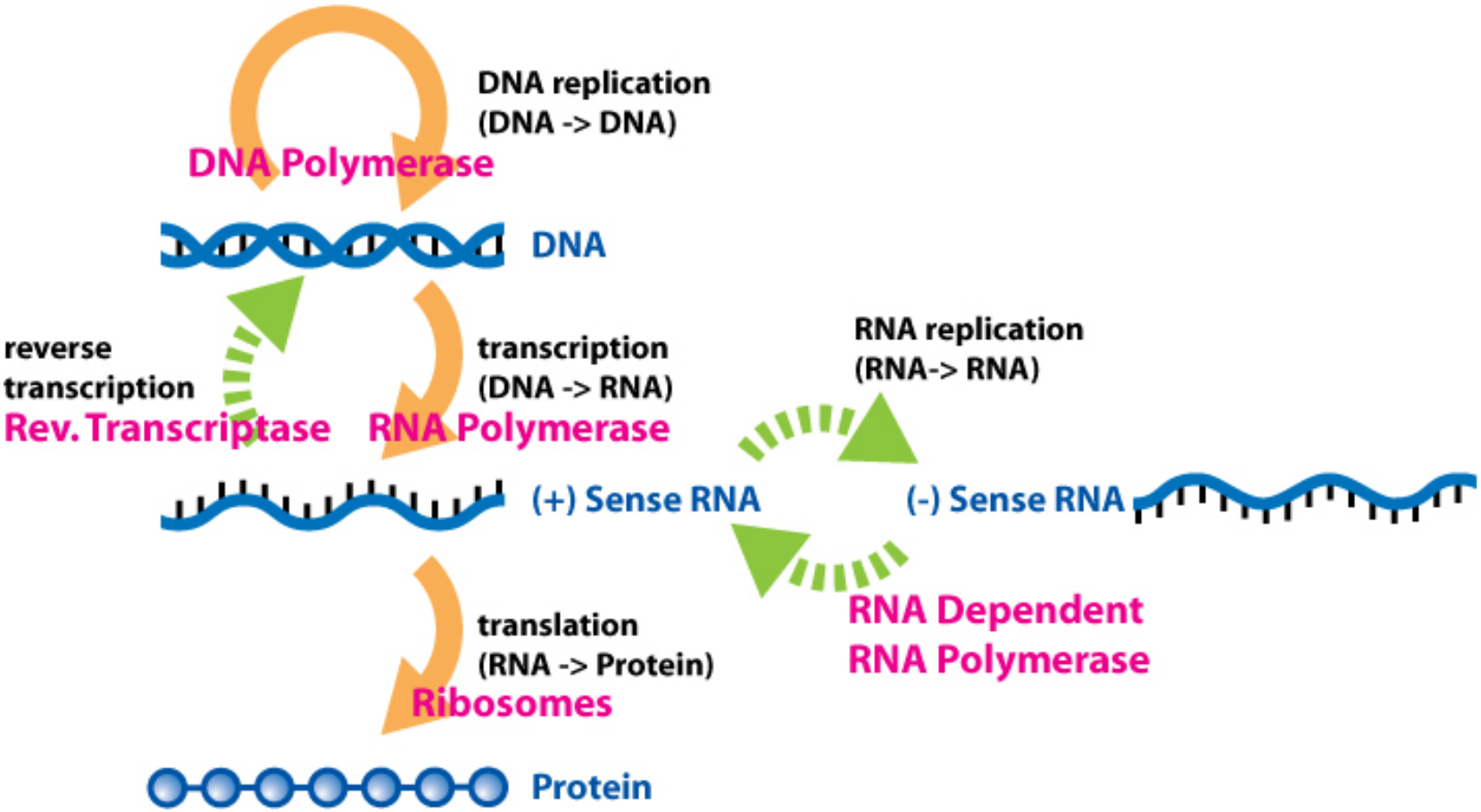
The central dogma

RNA: Ribonucleic acid (*U*racil replaces *T*hymine)



https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology (Dhorspool, Daniel Horspool, CC BY-SA 3.0)

The central dogma



https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology (Dhorspool, Daniel Horspool, CC BY-SA 3.0)

Genetic code: DNA codon table

Amino acids biochemical properties	nonpolar	polar	basic	acidic

Termination: stop codon

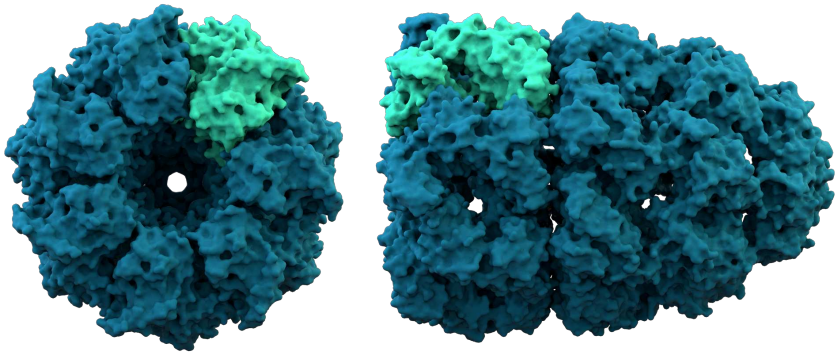
Standard genetic code

1st base	2nd base T		2nd base C		2nd base A		2nd base G		3rd base
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine	TCA		TAA[B]	Stop (*Ochre*)	TGA[B]	Stop (*Opal*)	A
	TTG		TCG		TAG[B]	Stop (*Amber*)	TGG	(Trp/W) Tryptophan	G
C	CTT		CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	ATG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GTG		GCG		GAG		GGG		G

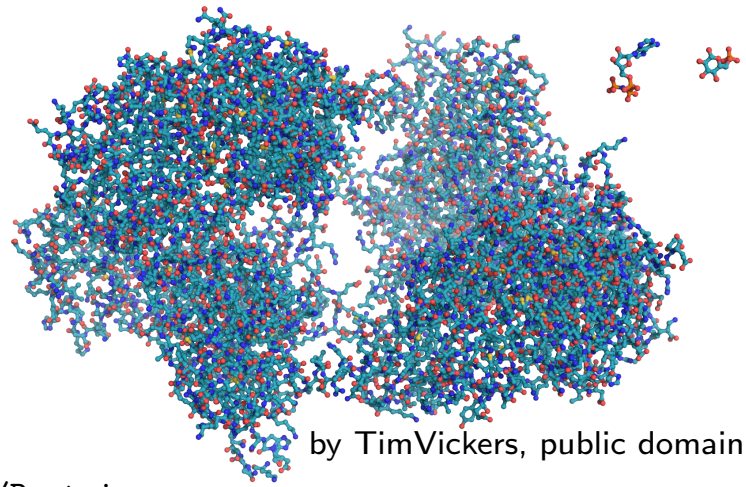
Translation begins at first ATG.

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

Proteins (3D objects)

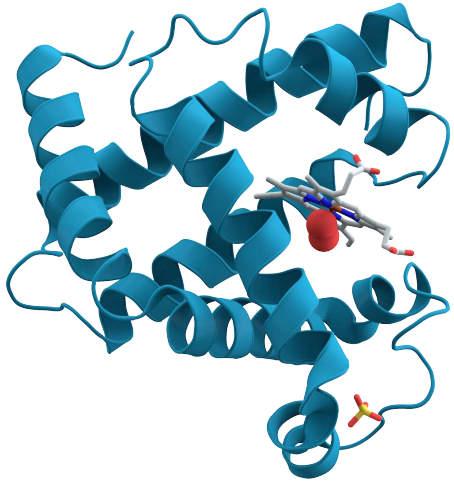


by Thomas Splettstoesser, CC BY-SA 3.0



by TimVickers, public domain

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



by AzaToth, public domain

FASTA format, databases, tools

```
>DNA_SEQUENCE_NAME SEQUENCE DESCRIPTION
AAATGAACGAAAATCTGTTCGCTTCATTCATTGCCCCCACAATCCTAGGCCTACCC
```

Open reading frame (ORF): nucleotide sequence producing a protein sequence without stops

```
>PROTEIN_SEQUENCE_NAME SEQUENCE DESCRIPTION (+1 ATP8)
KWTKICSLHSLPPQS[Stop]
```

```
>PROTEIN_SEQUENCE_NAME SEQUENCE DESCRIPTION (+3 ATP6)
MNENLFASFIAPTILGLP
```

`https://www.ncbi.nlm.nih.gov/search/` `https://www.ebi.ac.uk/services`

`https://www.ebi.ac.uk/Tools/emboss/`

FASTA format (woodmouse.fasta from phangorn)

kbart : bash — Konsole

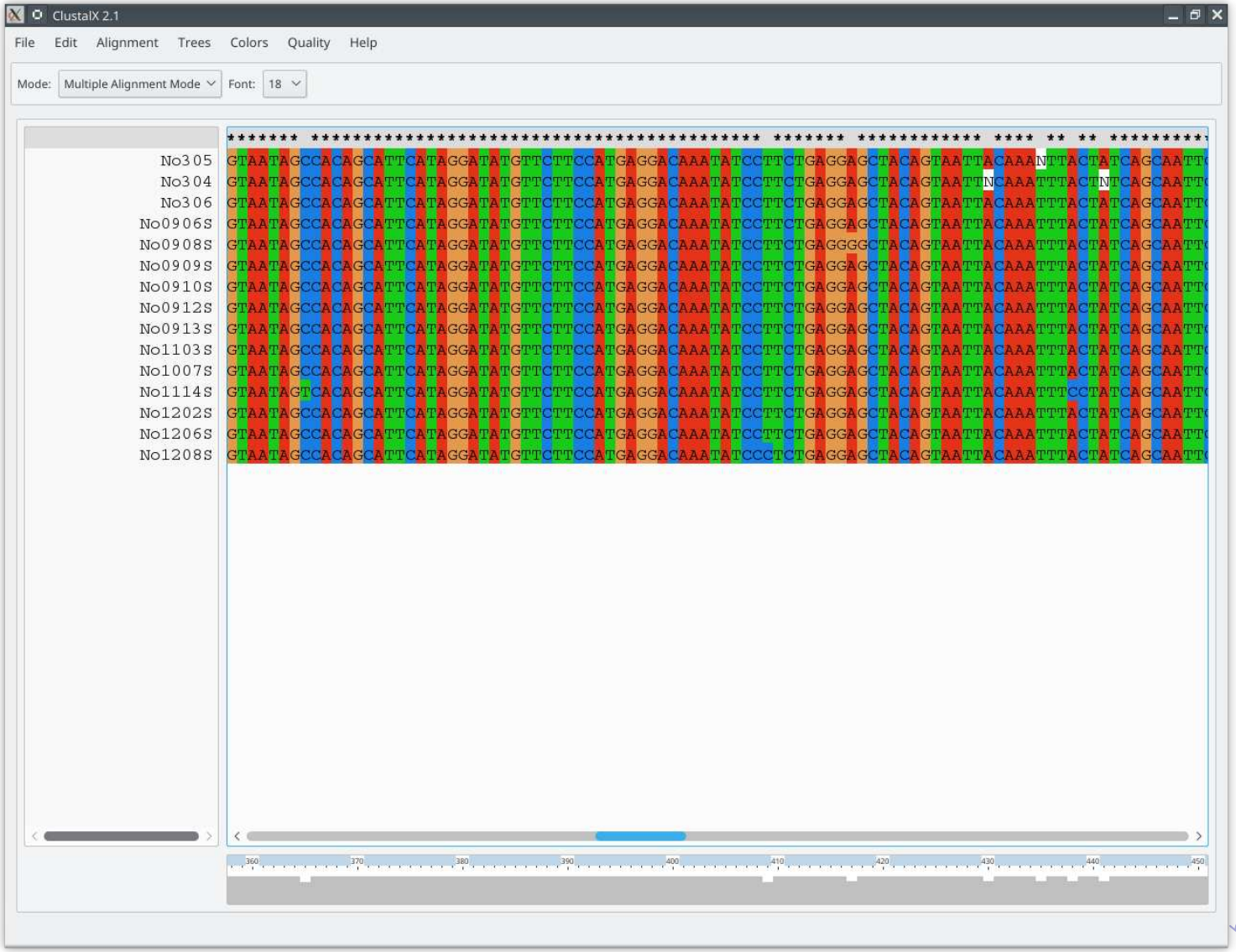
File Edit View Bookmarks Settings Help

```
kbart@linux-vcx4:~> cat woodmouse.fasta
>No305
nttcgaaaaacacacccactactaaaanttatcagtcactccttcatcgacttaccagctccatctaacatttcatcatgatgaaacttcggctcattactagg
cctcacaggcttattcctagcaatacactacacatcagacacaataacagcattctcttcagtaacccatatttgtcgagacgtaaattatggctgactaattc
cctcaatatttttatttgcctatttcttcatgtaggacgaggaatgtattacgggtcatacattttcatagaaacatgaaacatcggtgtggtccttttattc
ataggatatgttcttccatgaggacaaatatccttctgaggagctacagtaattacaaanttactatcagcaattccatacatcggaactaccctagtagagtg
agacaaagccacactgacacgttttttcgcttttcactttatccttccatttatcattgccgccctagtaattgtccacctcttgtttctccatgaaactggat
actcagacgccgataaaatcccatttcacccttactatacaattaaagatattctaggtgtactaataatagtttccttcctaataactttagtccttttcttt
gacaactacatgcctgctaacccacttaacaccccacctcatattaaaccagaatgatatttcctatttgcctatgcaatcctacgatccatccccaataaact
tctatcaatcttaattttagccttattaccatttcttcacacttccaaacaacgcagcctaatattccgcccaattactcagaccctata
>No304
attcgaaaaacacacccactactaaaaattatcaaccactccttcatcgacttaccagctccatctaacatttcatcatgatgaaacttcggctcattactagg
cctcacaggcttattcctagcaatacactacacatcagacacaataacagcattctcttcagtaacccatatttgtcgagacgtaaattatggctgactaattc
cctcaatatttttatttgcctatttcttcatgtaggacgaggaatgtattacgggtcatacattttatagaaacatgaaacatcggtgtagtccttttattc
ataggatatgttcttccatgaggacaaatatccttctgaggagctacagtaattncaaatttactntcagcaattccatacatcggaactaccctagtagaatg
agacaaagctacattgacacgttttttcgcttttcacttcatccttccatttatcattgccgccctagtaattgttcacctcttgtttctccatgaaactggat
actcagacgccgataaaatcccatttcacccttactatacaattaaagatattctaggtgtactaataatagtttccttcctaataactctagtccttttcttt
gacaactacatgcctgctaacccacttaacaccccacctcatattaaaccagaatgatacttcctatttgcctatgcaatcctacggtccatccccaataaact
tctatcaatcttaattttagccttattaccatttcttcacacttccaaacaacgcagcctaatattccgcccaattactcaaaccctgtn
>No306
attcgaaaaacacacccactactaaaaattatcaatcactccttcatcgacttaccagctccatctaacatttcatcatgatgaaacttcggctcattactagg
cctcacaggcttattcctagcaatacactacacatcagacacaataacagcattctcttcagtaacccatatttgtcgagacgtaaattatggctgactaattc
cctcaatatttttatttgcctatttcttcatgtaggacgaggaatgtattacgggtcatacattttatagaaacatgaaacatcggtgtgatccttttattc
ataggatatgttcttccatgaggacaaatatccttctgaggagctacagtaattacaaatttactatcagcaattccatacatcggaactaccctagtagaatg
agacaaagctacattgacacgttttttcgctttcatctccttccatttatcattgccgccctagtaattgttcacctcttgtttctccatgaaactggat
actcagacgccgataaaatcccatttcacccttactatacaattaaagatattctaggtgtactaataatagtttccttcctaataactttagtccttttcttt
gacaactacatgcctgctaacccacttaacaccccacnncatattaaaccagaatgatacttcctatttgcctatgcaatcctacggtccatccccaataaact
tctatcaatcttaattttagccttattaccatttcttcacacttccaaacaacgcagcctaatattccgcccaattactcaaaccctata
>No0906S
attcgaaaaacacacccactactaaaaatcatcaatcactccttcatcgatttaccagctccatctaacatttcatcatgatgaaacttcggctcattactagg
cctcacaggcttattcctagcaatacactacacatcagacacaataacagcattctcttcagtaacccatatttgtcgagacgtaaattatggctgactaattc
cctcaatatttttatttgcctatttcttcatgtaggacgaggaatgtattacggatcatacattttatagaaacatgaaacatcggtatgatccttttattc
ataggatatgttcttccatgaggacaaatatccttctgaggagctacagtaattacaaatttactatcagcaattccatacatcggaactaccctagtagagtg
agacaaagctacattgacacgttttttcgcttttcacttcatccttccatttatcattgccgccctagtaattgttcacctcttgtttctccatgaaactggat
actcagacgccgataaaatcccatttcacccttactatacaattaaagatattctaggtgtactaataataatttccttcctaataactttagtccttttcttt
gacaactacatgcctgctaacccacttaacaccccacctcatattaaaccagaatgatacttcctatttgcctatgcaatcctacgatccatccccaataaact
tctatcaatcttaattttagccttattaccatttcttcacacttccaaacaacgcagcctaatattccgcccaattactcaaatacnnnn
```

kbart : bash

ClustalX (woodmouse.fasta from phangorn)

http://www.clustal.org/



Genome organization

Chromosomes (humans: usually 23 pairs, diploid)

Mitochondrial DNA

Plasmid DNA (usually bacteria but also in Archaea or Eukaryote)

Gene: "a gene is a sequence of DNA or RNA that codes for a molecule that has a function" (`https://en.wikipedia.org/wiki/Gene`)

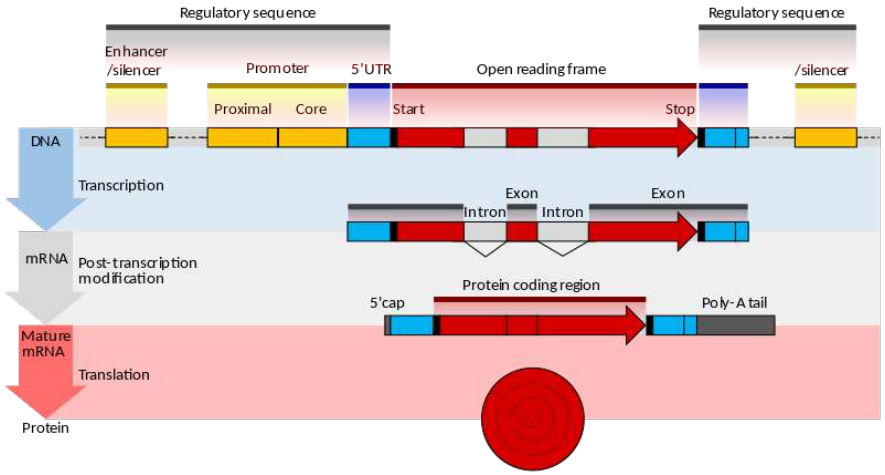
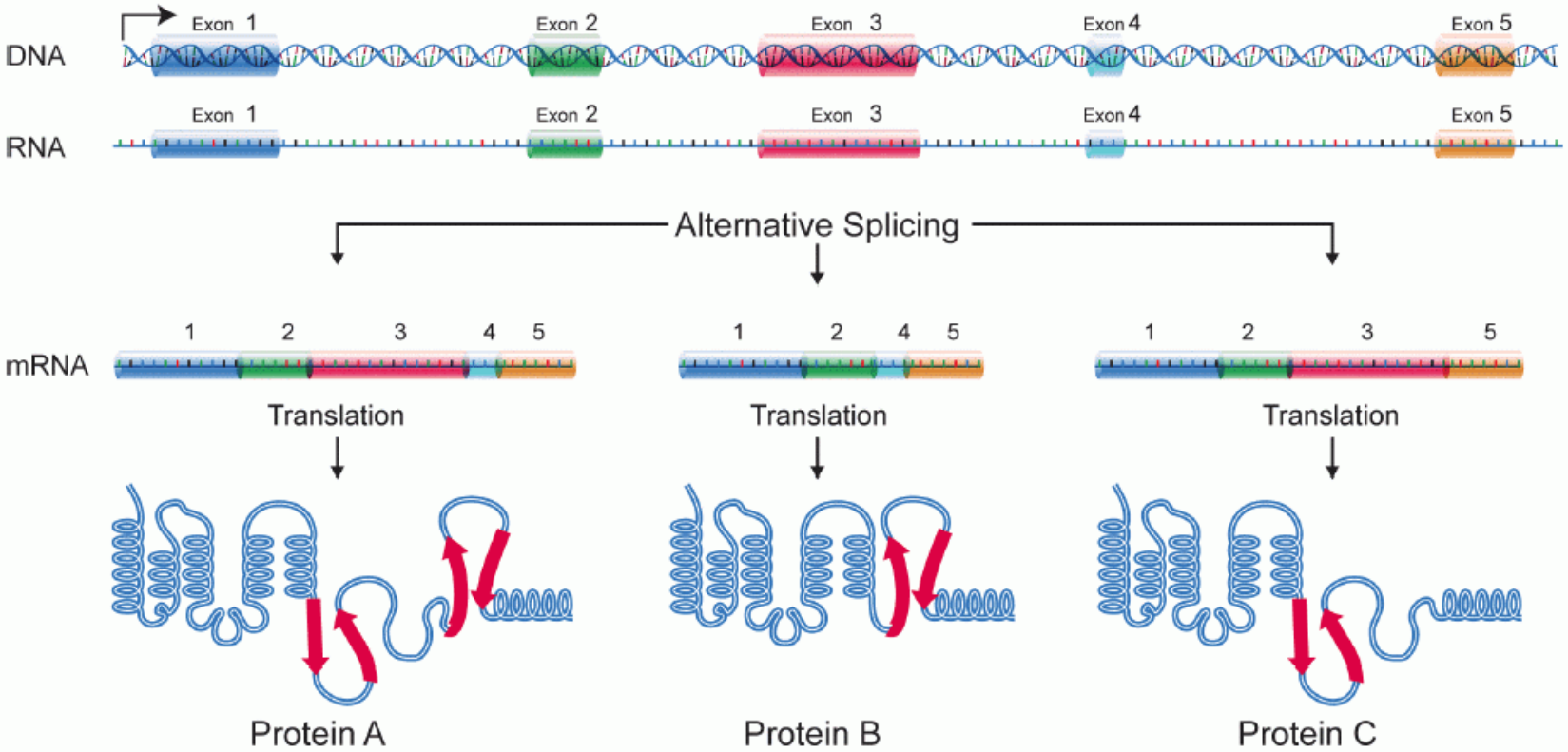


Figure: Eukaryotic protein–coding gene structure.

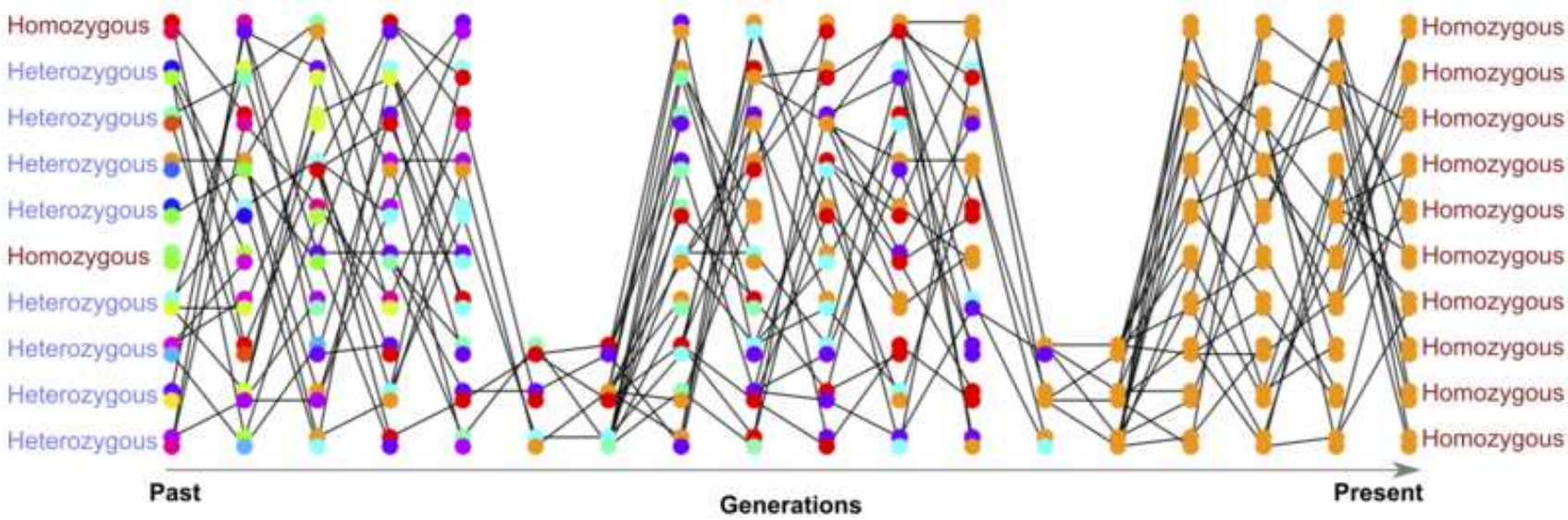
T. Shafee, R. Lowe (2017). "Eukaryotic and prokaryotic gene structure". WikiJournal of Medicine 4(1). DOI:10.15347/wjm/2017.002 CC-BY-SA

Alternative splicing



https://en.wikipedia.org/wiki/Alternative_splicing (public domain)

Evolution

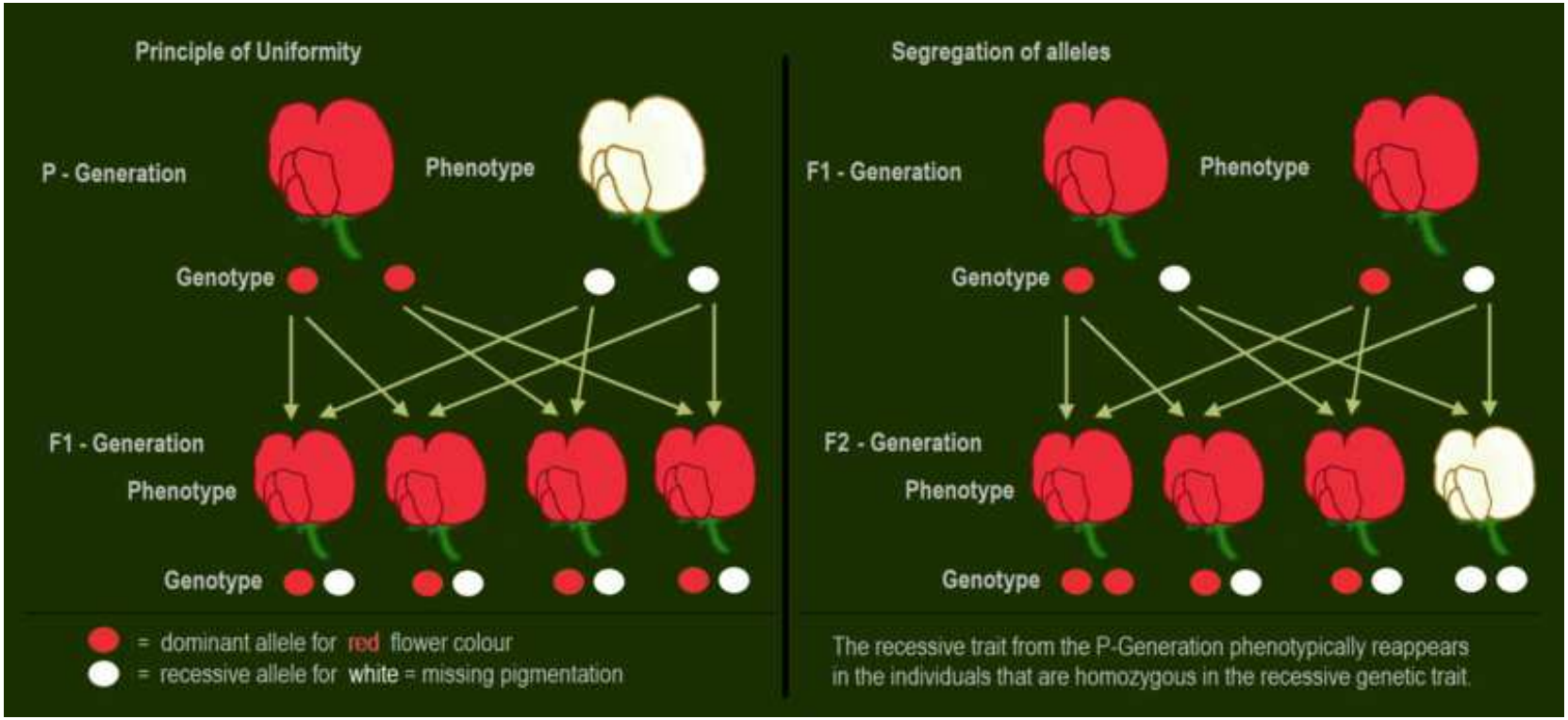


A population with bottlenecks

https://en.wikipedia.org/wiki/File:Loss_of_heterozygosity_over_time_in_a_bottlenecking_population_with_label.png
by Graham Coop, CC BY-SA 3.0

https://cooplab.github.io/popgen-notes/

Gregor Mendel (XIX century Moravian monk)



https://en.wikipedia.org/wiki/File:Dominant-recessive_inheritance_-_flowers_of_pea_plants.png
by Sciencia58, public domain

Nomenclature (diploid organisms)

Two copies of each gene, chromosome pairs (cf with *polyploid*)
locus: fixed position on a chromosome
allele: "alternative sequence variant that occurs at a certain locus"
e.g. blue, green, brown eye colour (multiple alleles)
homozygote: both alleles the same
heterozygote: alleles different
dominant allele: expressed in the phenotype in a heterozygote
recessive allele: only expressed in the phenotype in a homozygote
co–dominance: in heterozygote contribution of both alleles visible (e.g. AB blood type)

	A	a
A	AA	Aa
a	aA	aa

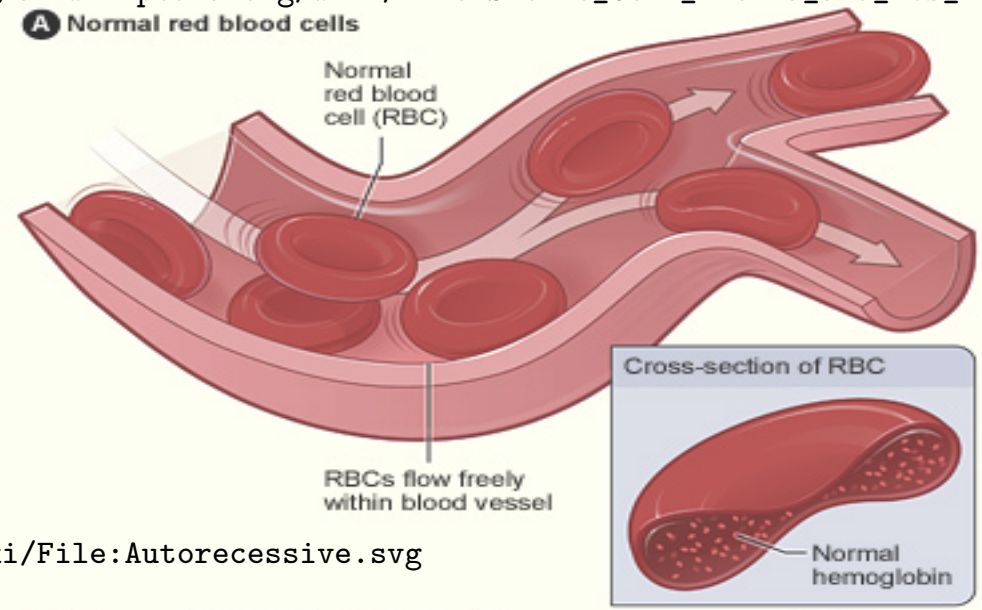
Single–nucleotide polymorphism (SNPs)

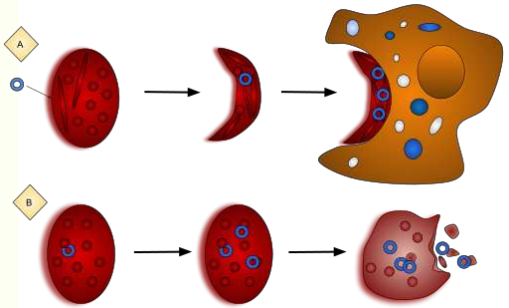
AAATGAACGAAAATCTGTTCGCTTCATTCATTGCCCCCAC
AAA**A**GAACGAAAATCTGTTCGCTTCATTCATTGCCCCCAC
AAATGAACGAAAATCTGTTCGCTTCATTCATTGCCCCCAC
Present in visible (e.g. $> 1\%$) part of population

Example: Sickle cell disease, SNP in β–globin gene
GAG$\rightarrow$GTG, glutamic acid (E)$\rightarrow$valine(V),
homozygotes: pain, anemia, swelling, infections, stroke
heterozygotes: problems minor, hinders malaria parasite reproduction

Sickle cell disease

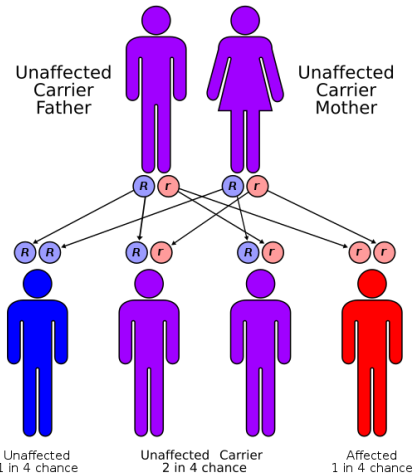
https://en.wikipedia.org/wiki/File:Sickle_Cell_Anemia_and_its_Protection_Against_Malaria.svg





by Dsheffer02, CC BY-SA 4.0

https://en.wikipedia.org/wiki/File:Autorecessive.svg

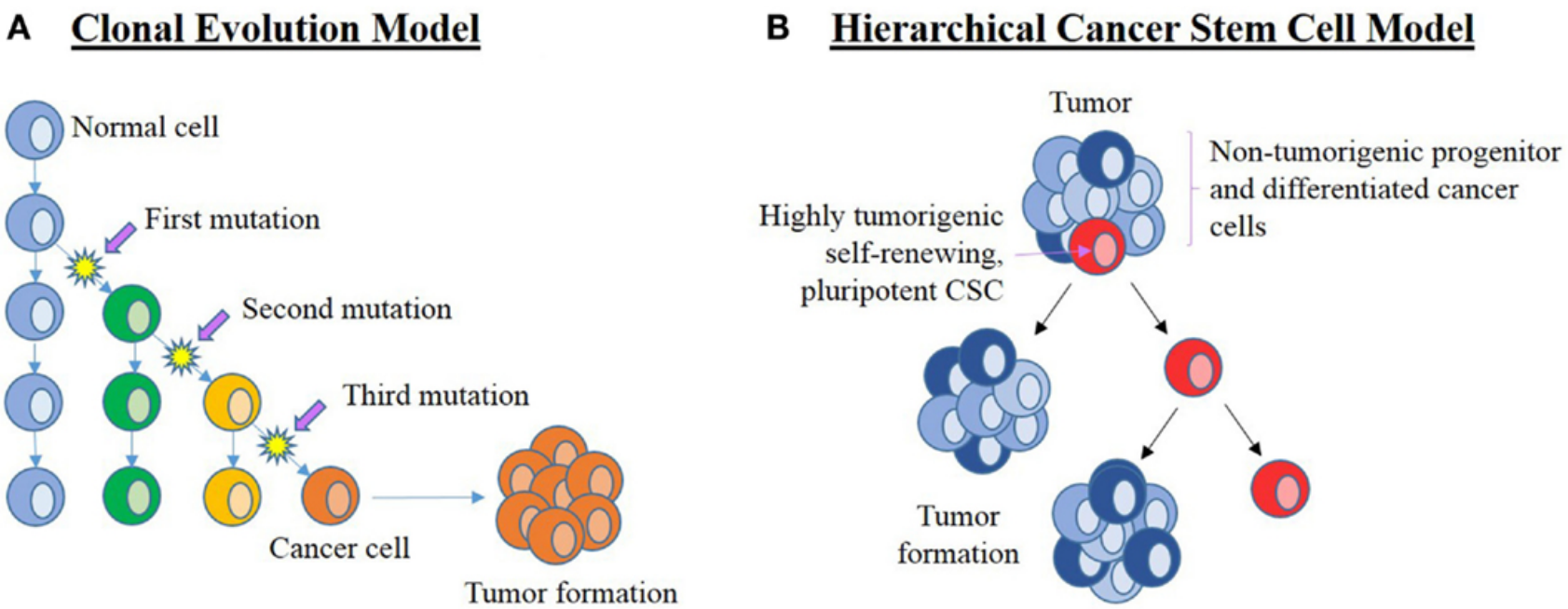


by Cburnett, GNU Free Documentation License

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

The National Heart, Lung, and Blood Institute (NHLBI), Public domain

Evolution: cancer



A. Bradshaw, A. Wickremsekera, S. T. Tan, L. Peng, P. F. Davis and T. Itinteang (2016)
Cancer Stem Cell Hierarchy in Glioblastoma Multiforme. Front. Surg. 3:21. doi:10.3389/fsurg.2016.00021, CC BY

Wright–Fisher model

Two alleles (A and a)
neutral evolution (mutation does not change fitness)
diploid individuals (N individuals treated as $2N$ copies of locus)
X_n: number of As in generation n, binomial transition:

$$p(X_{n+1} = j | X_n = i) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}$$

0 and $2N$ are absorbing states and with probability 1 the process will eventually end up in one of them

$$P(X_\infty = 2N) = X_0/(2N)$$

DNA: Sequencing

Single gene (genome fragments), whole genome
DNA fragmented: need sequence assembly algorithms
de–novo assembly: "Given a set of sequence fragments, the object is to find a longer sequence that contains all the fragments."
Shortest common superstring (NP–complete)
mapping assembly: "assembling reads against an existing backbone sequence, building a sequence that is similar but not necessarily identical to the backbone sequence"

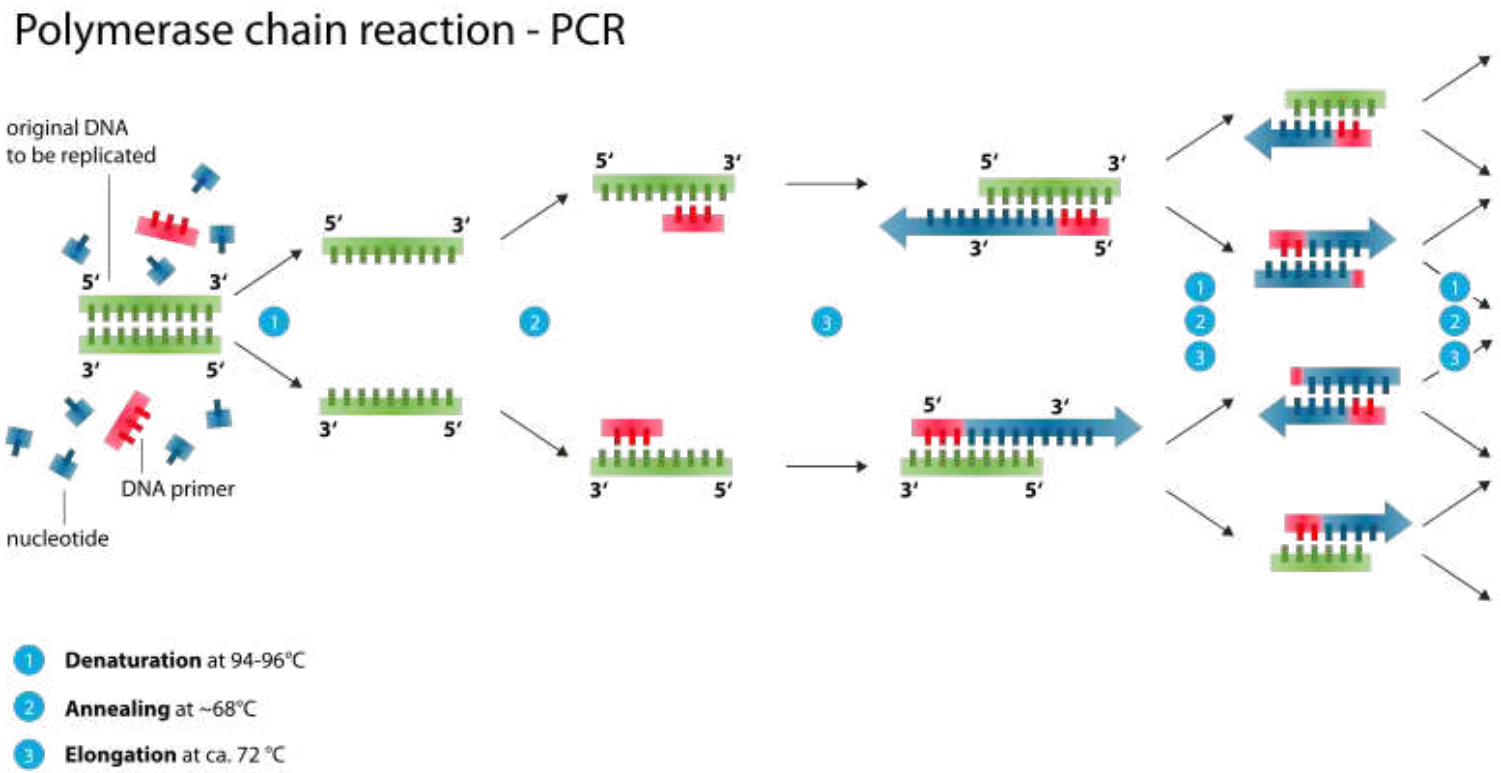
`https://en.wikipedia.org/wiki/Sequence_assembly`

read coverage: "number of unique reads that include a given nucleotide in the reconstructed sequence"

`https://en.wikipedia.org/wiki/Coverage_(genetics)`

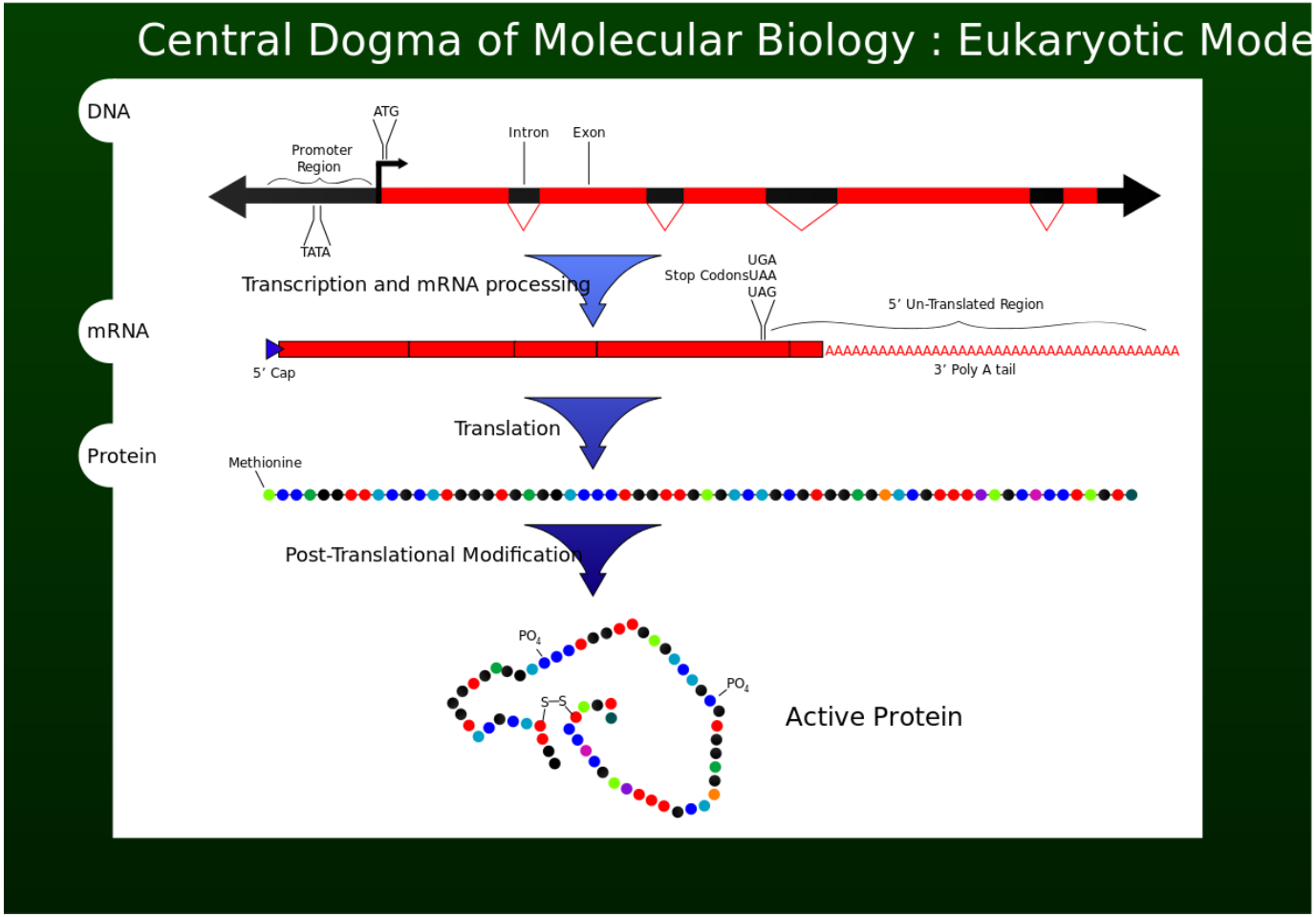
Errors in reading (look at FASTAQ format)
Errors in assembly

DNA: PCR



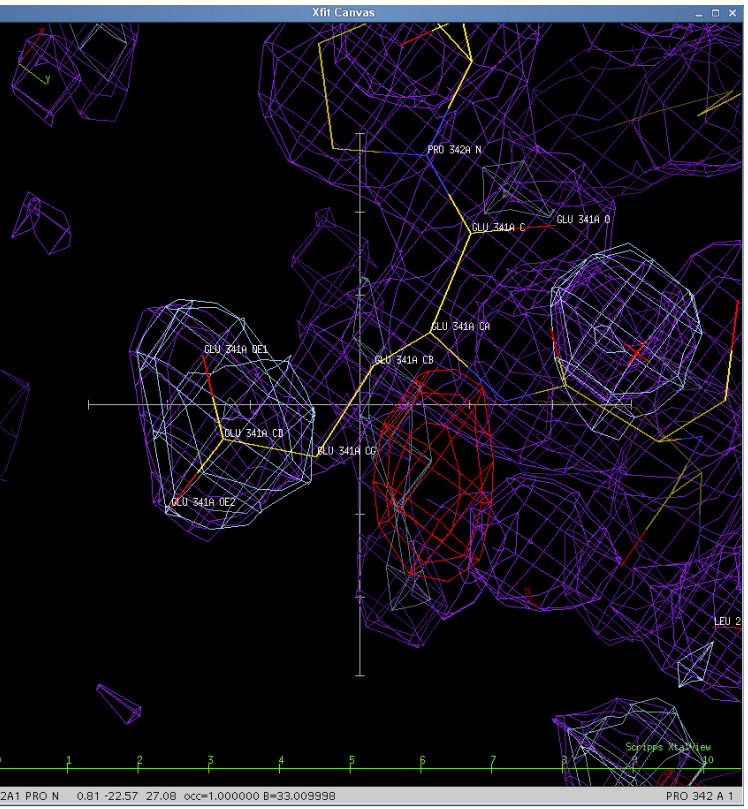
https://en.wikipedia.org/wiki/Polymerase_chain_reaction (by Enzoklop, CC BY-SA 4.0)

Proteins



https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology (by Mike Jones, CC BY-SA 2.5)

Proteins



X–ray crystallography

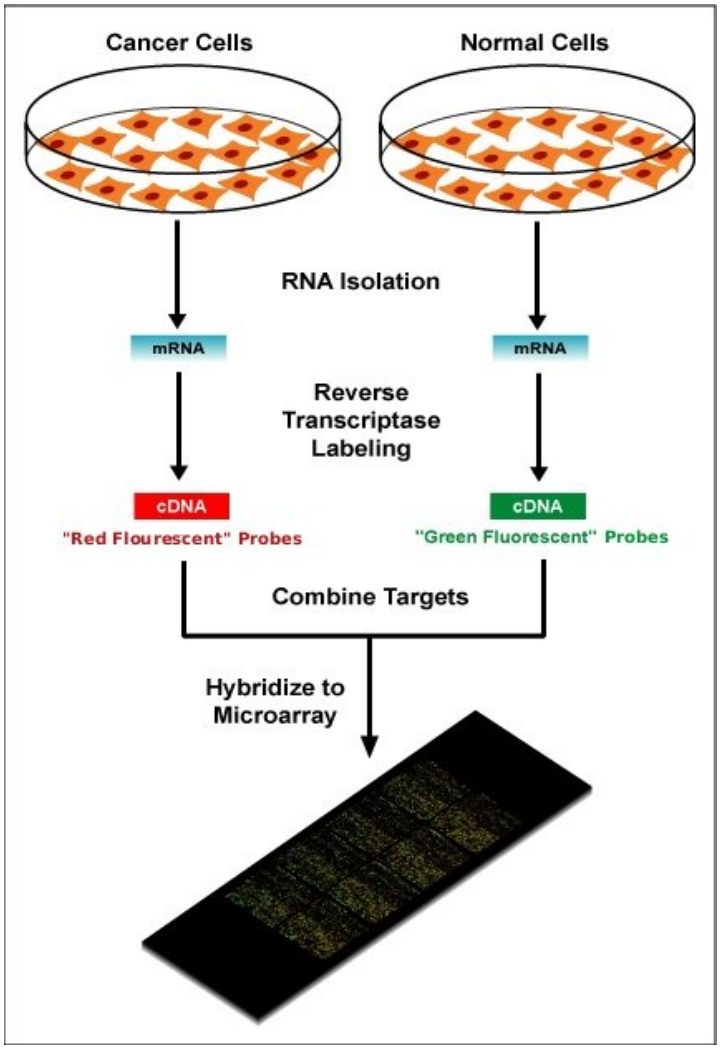
Nuclear magnetic resonance (NMR)

Computational methods:
given protein sequence find 3D structure
with minimal free energy (stability of system)
Knowledge–based methods,
Homology based methods (50% identity),
Threading: (if low identity) structure alignment

Proteins are dynamic!

Visualization by XFIT of output of X–ray cystallography
(from my hand–in assignment for Structural Biology course, Univ. Cambridge)

Genes



https://en.wikipedia.org/wiki/DNA_microarray (by Larssono, public domain)

What we observe

Collections of sequences:

- DNA (genes, regulatory regions),
- RNA,
- protein

3D protein structures

Phenotypic measurements:

- "usual" traits (e.g. body size),
- gene expression levels,
- health status (e.g. cancer)

What we want

Genotype–phenotype map

Protein structure prediction

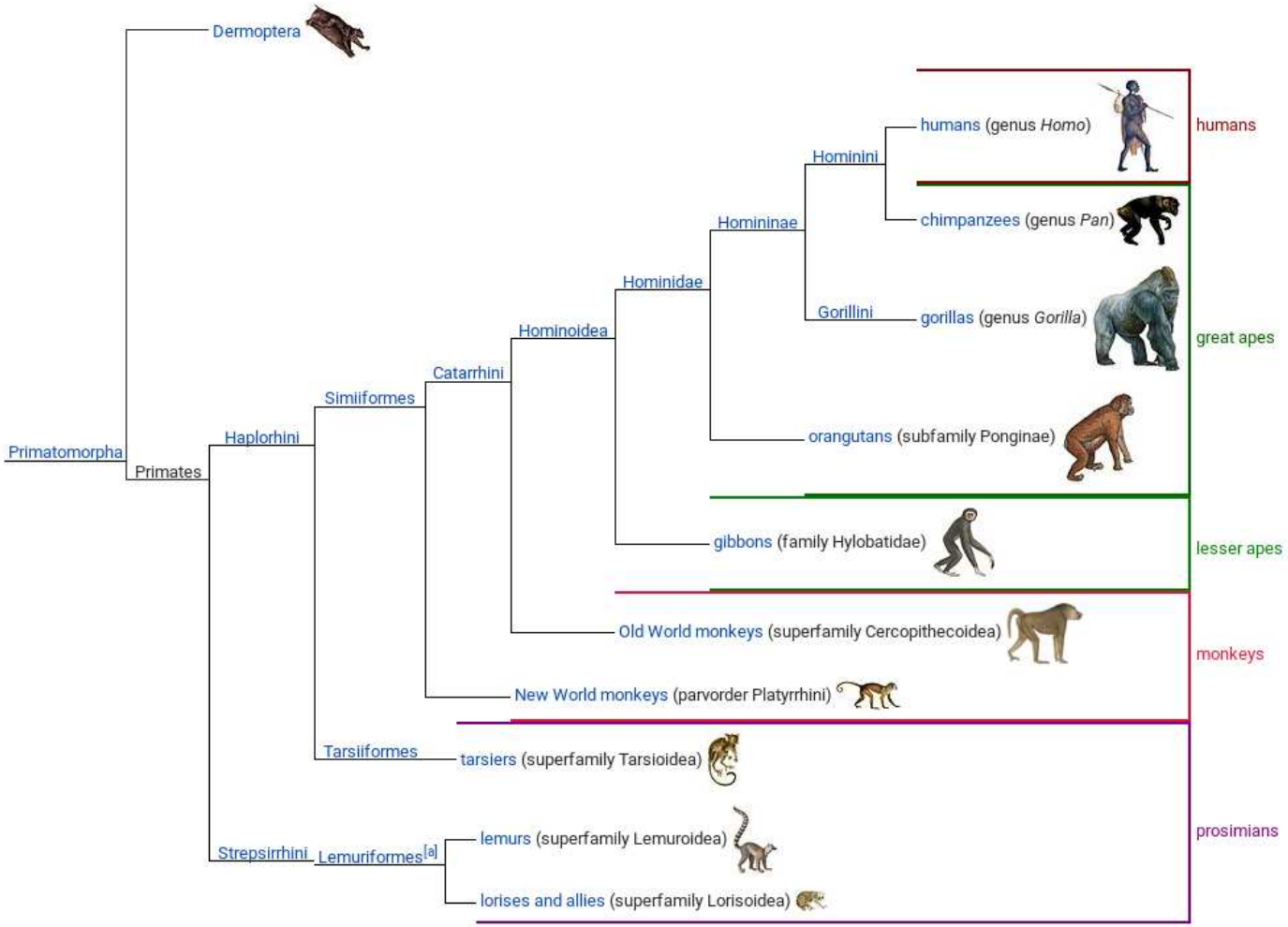
In silico drug design

Phylogeography: tracking epidemics
Phylogenetics: tracking(/predicting?) species, tumour evolution

`https://evograd.wordpress.com/2017/07/31/applications-of-phylogenetics-in-medicine-and-public-health/`

Phylogenetic linguistics

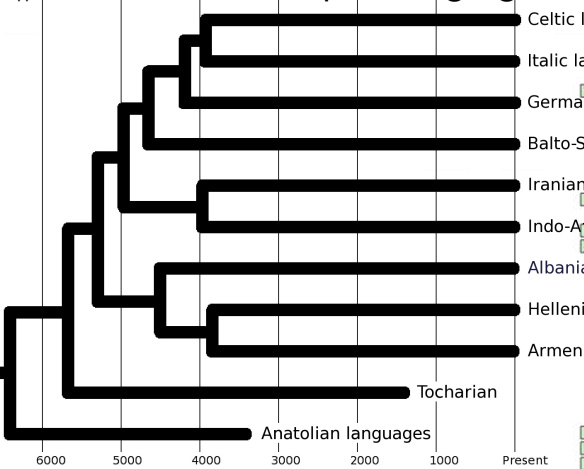
Primate phylogeny



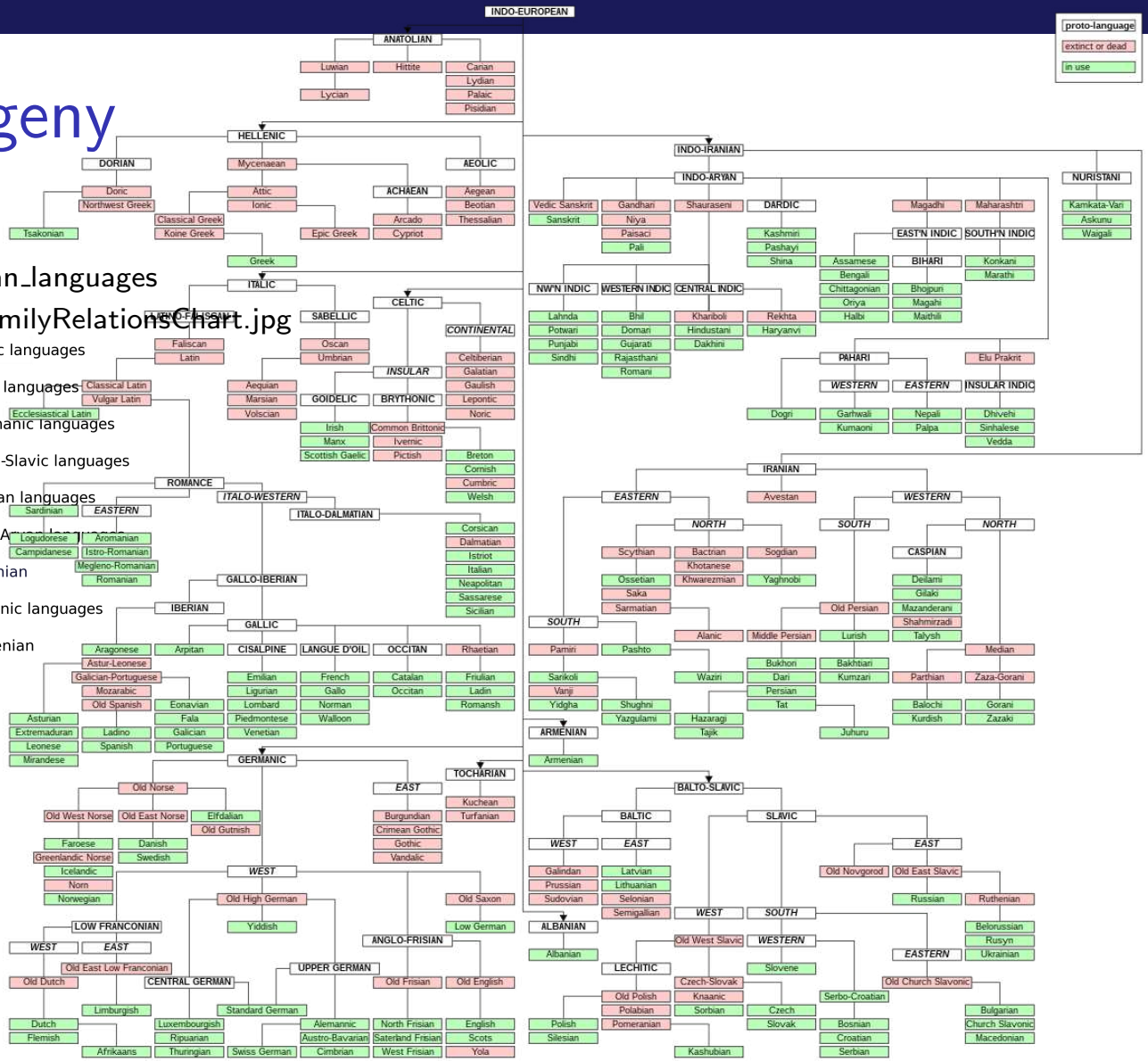
https://en.wikipedia.org/wiki/Primate

Language phylogeny

https:en.wikipedia.orgwikiIndoEuropean_languages
#mediaFile:IndoEuropeanLanguageFamilyRelationsChart.jpg



by Ordinary Person, CC BY-SA 4.0



https://en.wikipedia.org/wiki/Indo-European_languages#/media/File:IndoEuropeanTree.svg

by multiple authors, first version by Mandrak, CC BY-SA 3.0

Are phylogenetic methods useful in linguistics?

http://www.replicatedtypo.com/phylogenetics-in-linguistics-the-biggest-intellectual-fraud-since-chomsky/10854.htm

Software

R packages

BioPerl

Biopython

Stand–alone programs

Questions?