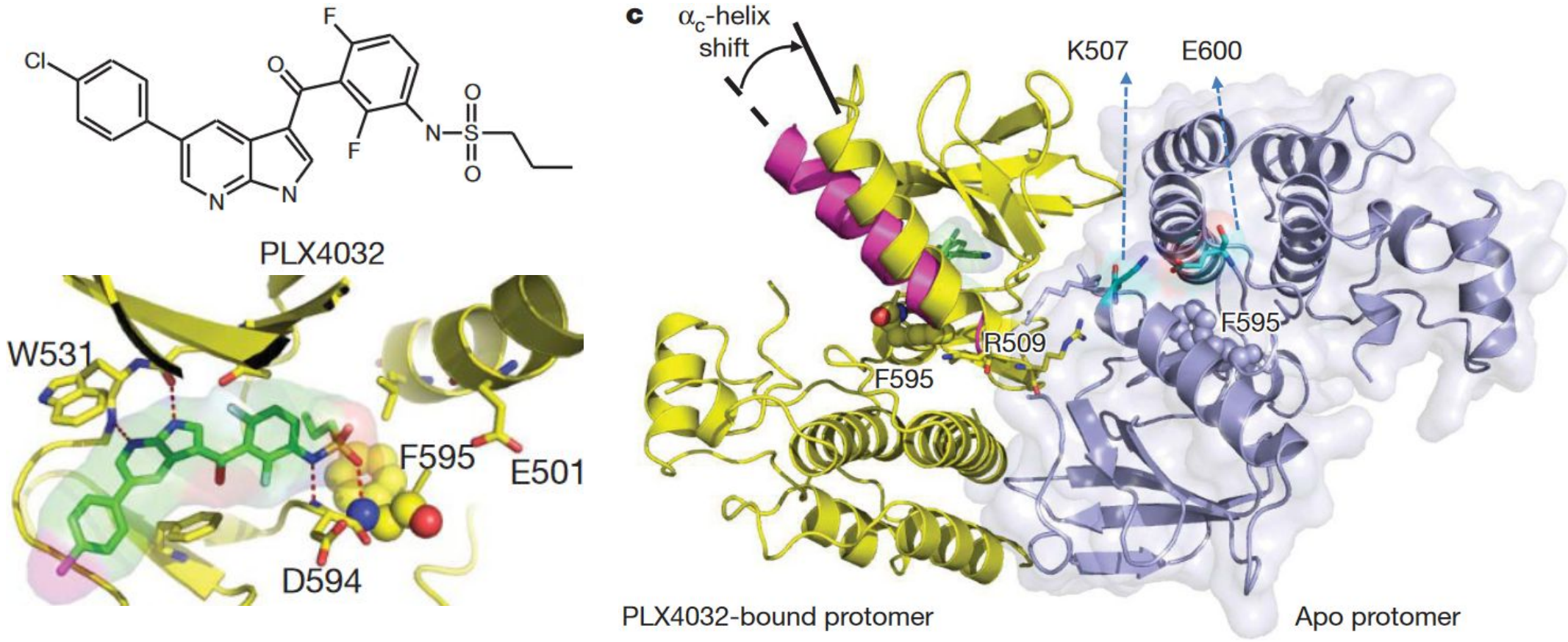
AMIDD Lecture 3: Biological Sequence Analysis (I)



Dr. Jitao David Zhang, Computational Biologist
[1] Pharmaceutical Sciences, Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche
[2] Department of Mathematics and Informatics, University of Basel

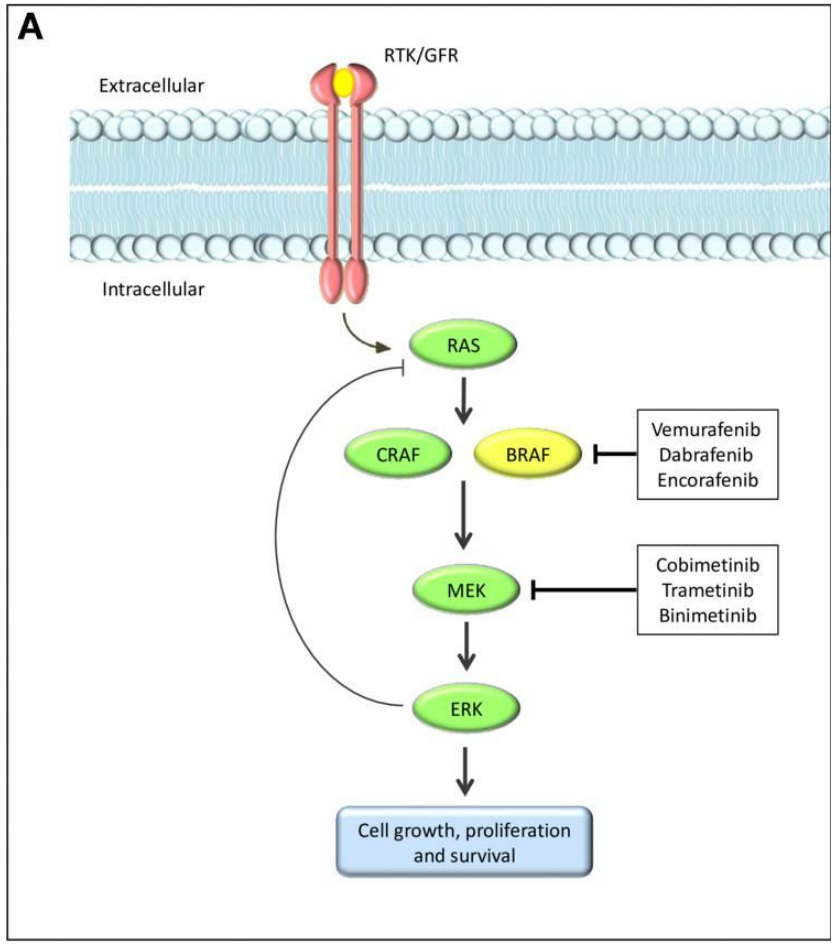
Today's goals

- Learning drug-discovery concepts with the help of Bollag *et al.*
- Biological and mathematical foundations of biological sequence analysis
  - Biological functions and mathematical modelling of mutations
  - Dynamic programming
- Selected applications of biological sequence analysis for drug discovery

Questions for the abstract of Bollag et al., 2010

1. What is the **indication** of PLX4032? BRAF-mutant melanoma
2. What is the **gene target** of PLX4032? The mutant form of BRAF, primarily V600E. The authors also suggested an activity against the V600K mutation.
3. The malignancy depends on which **biological pathway**? The RAF/MEK/ERK pathway.
4. What is the **Mechanism of Action** of PLX4032? PLX4032 inhibits the kinase activity of mutant BRAF, which inhibits ERK phosphorylation and blocks the RAF/MEK/ERK pathway in BRAF mutant cells.
5. What **went wrong** in the first **Phase I clinical trial**? And how was it **solved**? Patients did not respond, i.e. doctors observed no tumour regressions. The drug developer changed the formulation from crystalline to amorphous. The new formulation allowed higher drug exposures, which lead to high response rate.
6. What was the **dosing regimen** in the final **Phase I** clinical trial, and what is the response rate? Oral dose, 960 mg twice every day (bid, latin *bis in die*); Response rate: 81%.

The RAF/MEK/ERK pathway, also known as the MAPK pathway



RTK=receptor tyrosine kinase;
GFR=growth factor receptor
RAS/RAF/MEK/ER=MAP kinases

(A) normal pathway; **(B)** the most common resistance mechanisms, among others BRAF amplification and alternative splicing. Reference: Tanda *et al.* "Current State of Target Treatment in BRAF Mutated Melanoma." Frontiers in Molecular Biosciences 7 (2020): 154.

Amorphous and crystalline formulations

- Crystalline and amorphous formulations of the same drug have different physicochemical properties, which directly impact drug **exposure**, *i.e.* amount of drug achieving in the body per time unit. Exposure belongs to the field of study of pharmacology, more specifically **pharmacokinetics** (PK), which studies *what body does to the drug*.
- Having a good target and a good molecule is not enough - important other properties include PK and PD (**pharmacodynamics,** *what drug does to the body*).

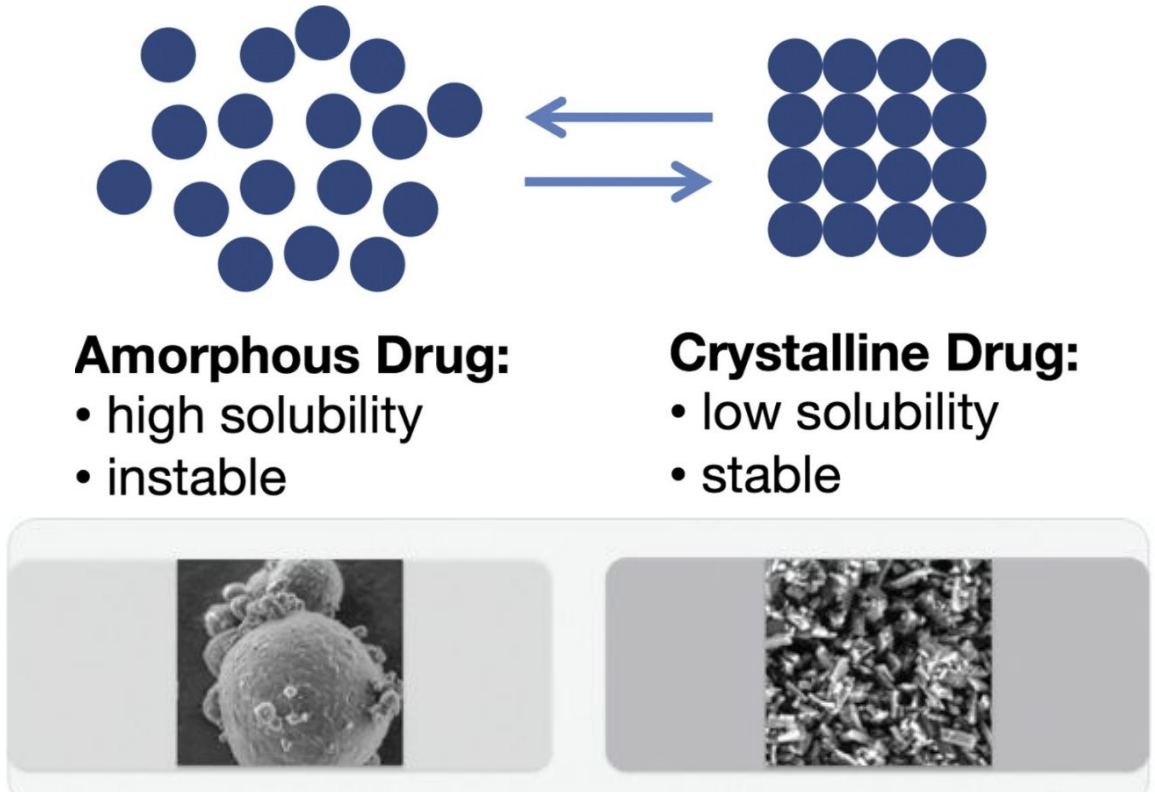


Figure sources: Dengal *et al.*, Advanced Drug Delivery Reviews, 2016; American Pharmaceutical Review

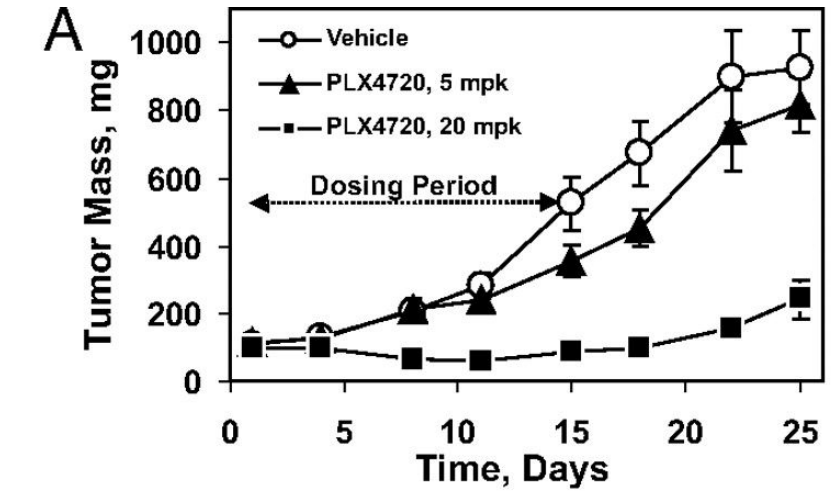
Questions for Bollag et al., 2010 (I)

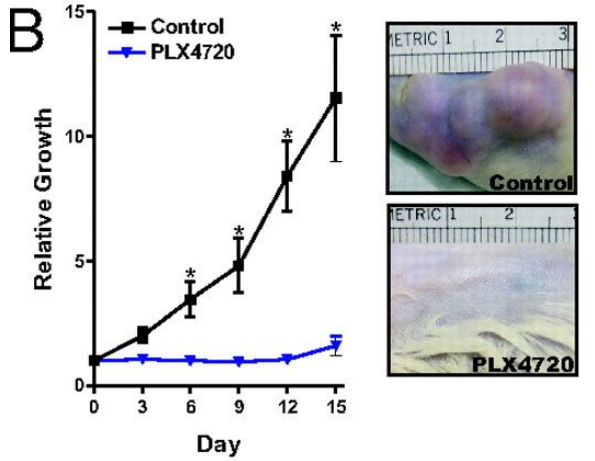
1. We learned that many drugs target one of the four protein types: GPCRs, ion channels, kinases, and nuclear receptors. Which type does the target of PLX4032 belong to? Kinases
2. How was the efficacy of PLX4032 tested? Cell lines, xenografts (mouse), and finally patients. Beagle dogs, cynomolgus monkeys, and rats were used to test safety, not efficacy.
3. Why was PLX4032 chosen for further development, but not PLX4720? Better scaling of pharmacokinetic (PK) properties.
4. How was the exposure of PLX4032 in the blood quantified? Which mathematical operation was used? Area under of curve of plasma concentrations - integration.
5. How was the final dosing regimen (960-mg BID) determined? It was the maximum tolerated dose, toxicities detected in higher-dose groups.

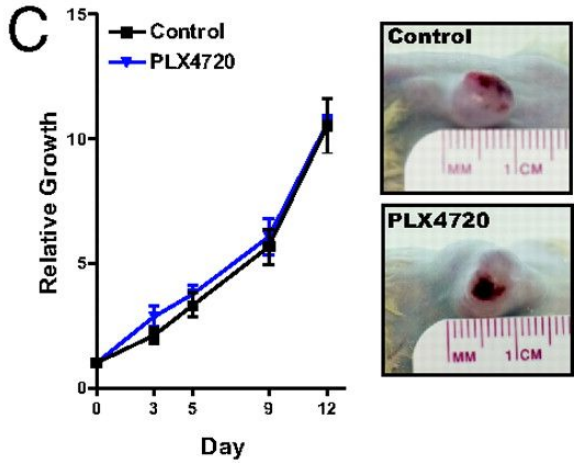
Efficacy observed in the xenograft model

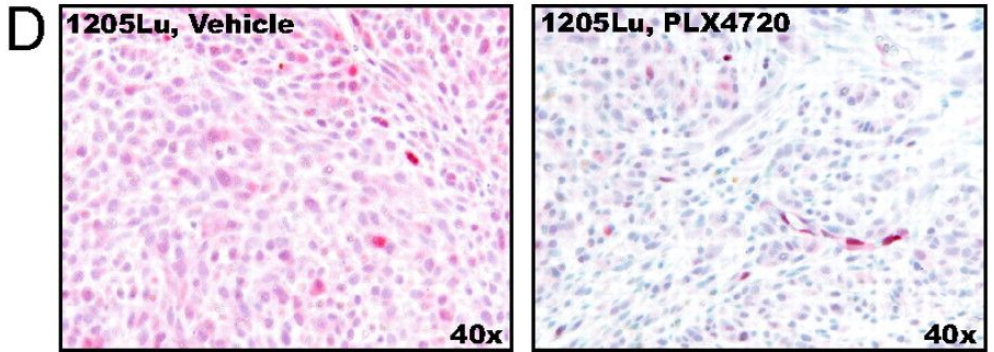
(*A*) Tumor volume measurements of xenograft tumors treated with 5 or 20 mg/kg PLX4720 by oral gavage or treated with vehicle. Dosing occurred from days 1 to 14. (*B* and *C*) Two million cells [1205Lu (*B*); C8161 (*C*)] were s.c. injected into SCID mice. After reaching sufficient size, mice were treated by oral gavage with vehicle control (*Left*) or 100 mg/kg PLX4720 (*Right*) twice daily for the indicated times. (D) 1205Lu xenograft tumors were extracted, fixed in formalin, and paraffin embedded. Vehicle- (*Left*) and PLX4720- (*Right*) treated samples were immunostained for phospho-ERK.

Figure 4 from Tsai, et al. “Discovery of a Selective Inhibitor of Oncogenic B-Raf Kinase with Potent Antimelanoma Activity.” PNAS (2008): 3041–46. https://doi.org/10.1073/pnas.0711741105.









Questions for Bollag et al., 2010 (II)

6. How did patients with the V600K mutation in BRAF respond? They responded better, with 71% and 100% reduction in tumour dimensions.

7. What measures were taken to demonstrate the effect of BRAF inhibition in patient biopsies? Phosphorylated-ERK and Ki67 levels were measured. Values from both measurements were found to be decreased in the later measurements, showing the reduction in ERK pathway activity.These are biomarkers, i.e. measurements that correlates drug treatment with efficacy.

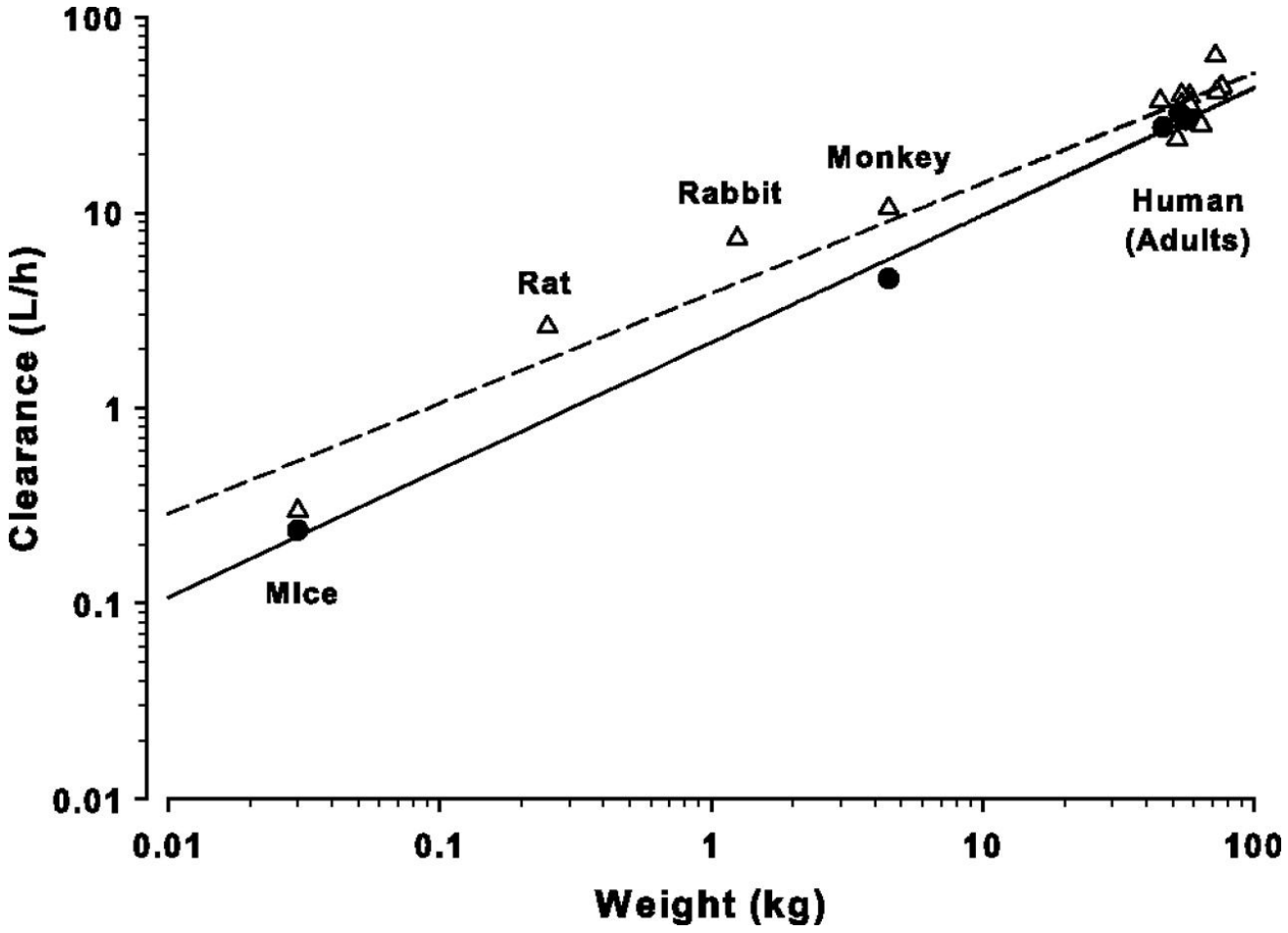
8. What side effects of PLX4032 were reported? Besides fatigue, rash, and joint pain, 31% patients treated at the maximum tolerated dose (MTD) developed skin lesions described as cutaneous squamous cell carcinomas, keratoacanthoma type.

9. What measures were taken against side effects and safety concerns of PLX4032? **Limiting the dose**, resection of the lesion, dermatological monitoring.

10. Where do you think mathematics and informatics is used in the discovery and development of PLX4032? *Almost every step*, especially X-ray data analysis, data summary and modelling.

Further questions

- What about using Raf inhibitors in tumors lacking Raf mutations? About 50% melanoma patients have BRAF mutations. Patients with wild-type BRAF often receive cancer immunotherapy (anti-PD-1 and anti-CTLA-4).
- What does *scaling* mean in Q3? Scaling means inter-species dose extrapolation, i.e. estimation of dose in a species (say monkey) based on data collected in another species (say dog).
- Reserving room in Biozentrum to follow the course together. I fully support it!



An example of scaling: Figure source: https://doi.org/10.1128/AAC.00067-11

A single-amino-acid difference in BRAF gene may mean longer survival of melanoma patients given the correct treatment

McArthur, Grant A., Paul B. Chapman, Caroline Robert, James Larkin, John B. Haanen, Reinhard Dummer, Antoni Ribas, *et al.*

Safety and Efficacy of Vemurafenib in BRAFV600E and BRAFV600K Mutation-Positive Melanoma (BRIM-3): Extended Follow-up of a Phase 3, Randomised, Open-Label Study

The Lancet Oncology 15, Nr. 3 (1. März 2014): 323–32. https://doi.org/10.1016/S1470-2045(14)70012-9.

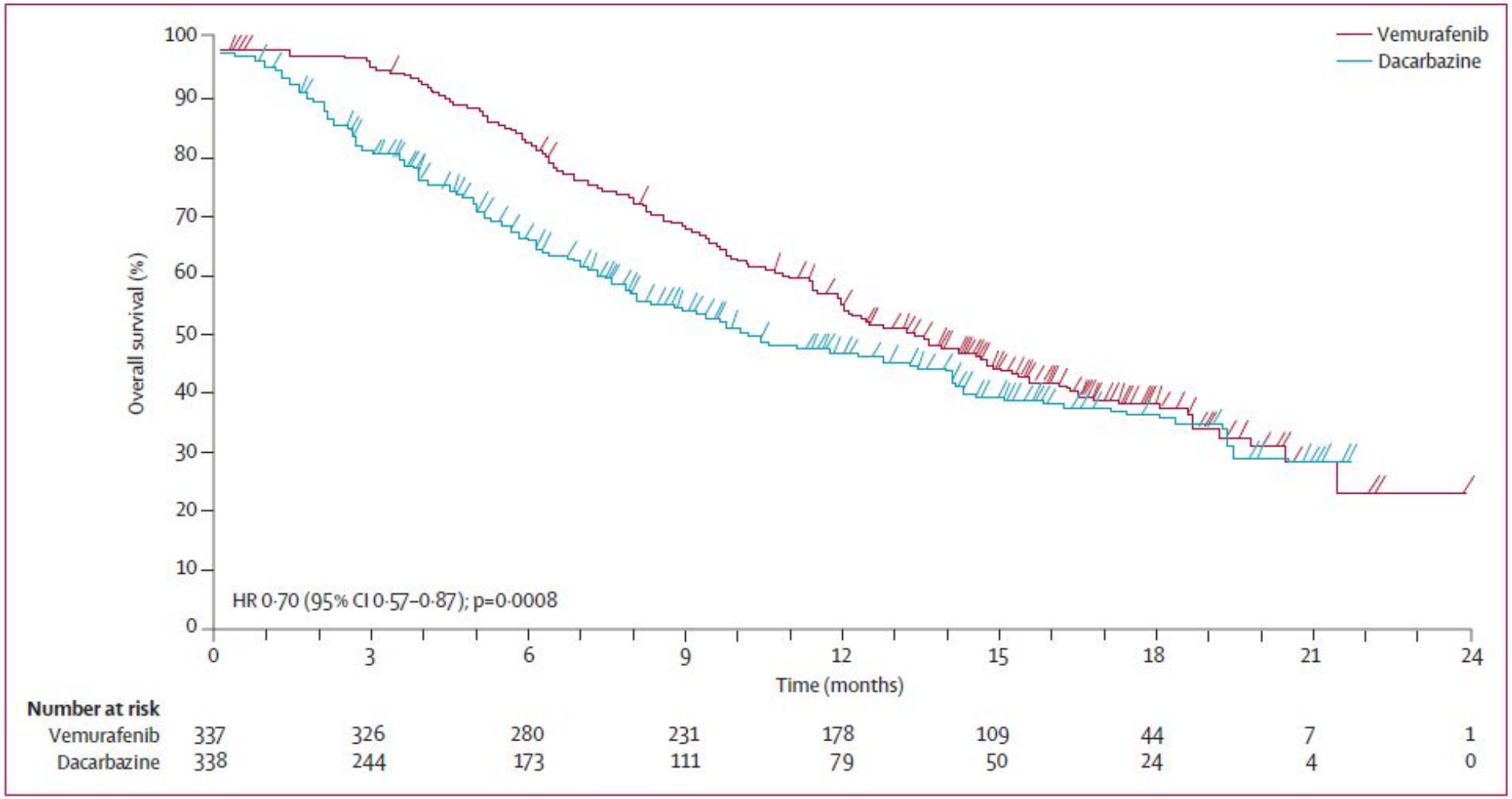
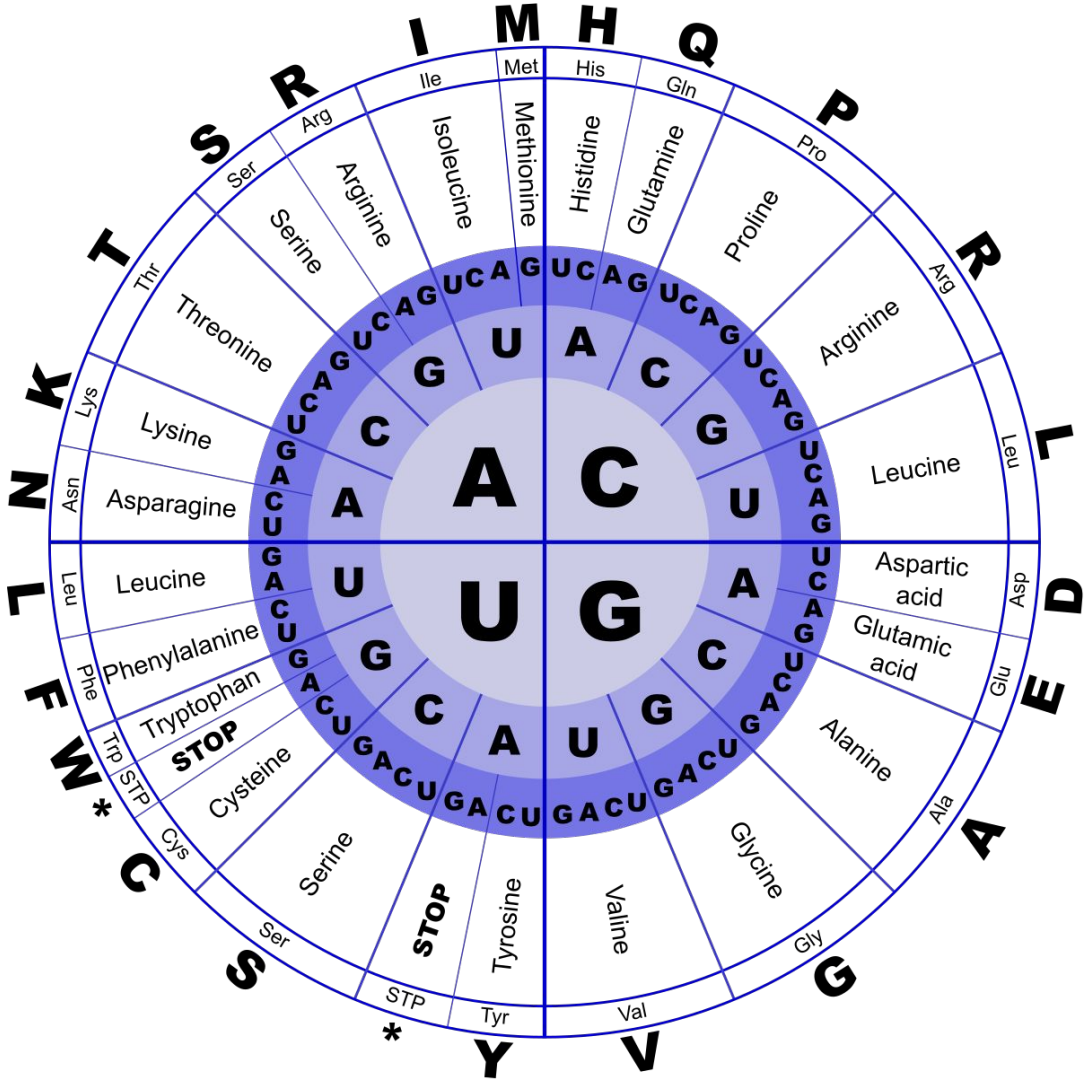
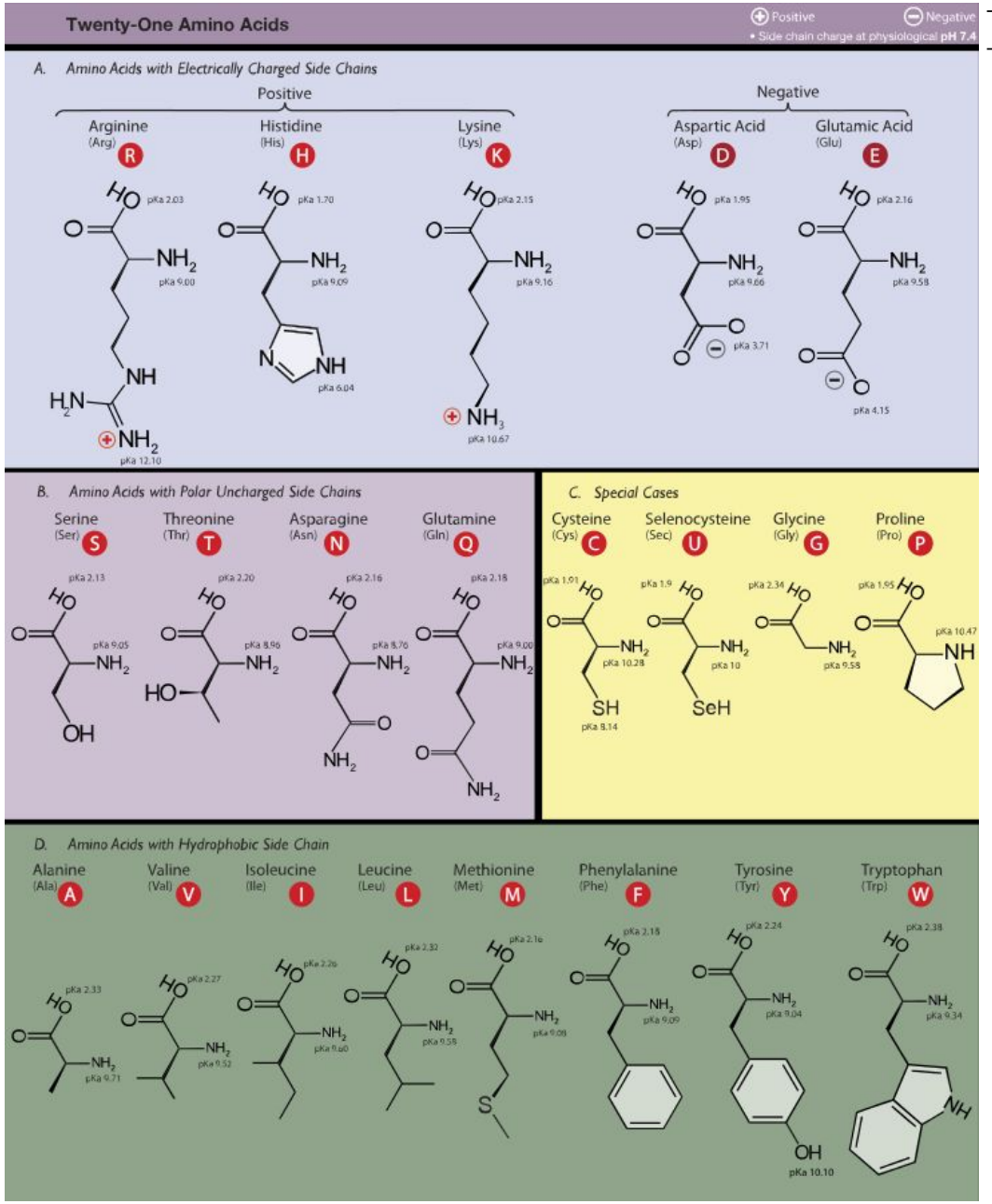


Figure 2: Overall survival (randomised population; censored at crossover) for patients randomly assigned to vemurafenib or to dacarbazine (cutoff Feb 1, 2012)

Codes of the central dogma





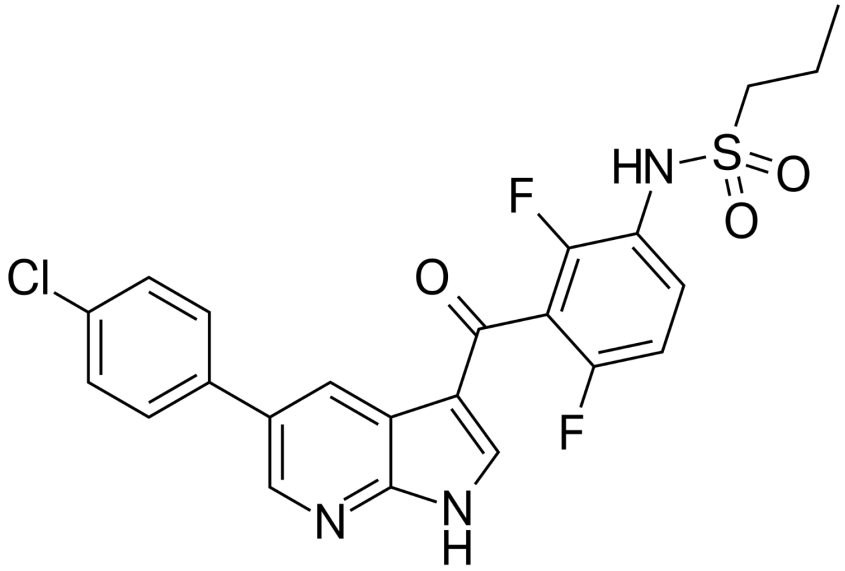
Left: Genetic codes (OpenChart, public domain); Right:Dancojocari, CC BY-SA 3.0, via Wikimedia Commons

Vemurafenib (Zelboraf, PLX4032)
V600E mutated BRAF inhibition

- V600E: Valine (V) on the amino-acid position 600 is substituted by glutamic acid (E).
- **Question: what mutations in DNA would cause the change in the amino-acid sequence?**

```
                560         570         580         590         600
IKLIDIARQT AQGMDYLHAK SIIHRDLKSN NIFLHEDLTV KIGDFGLATV
                610         620         630         640         650
```

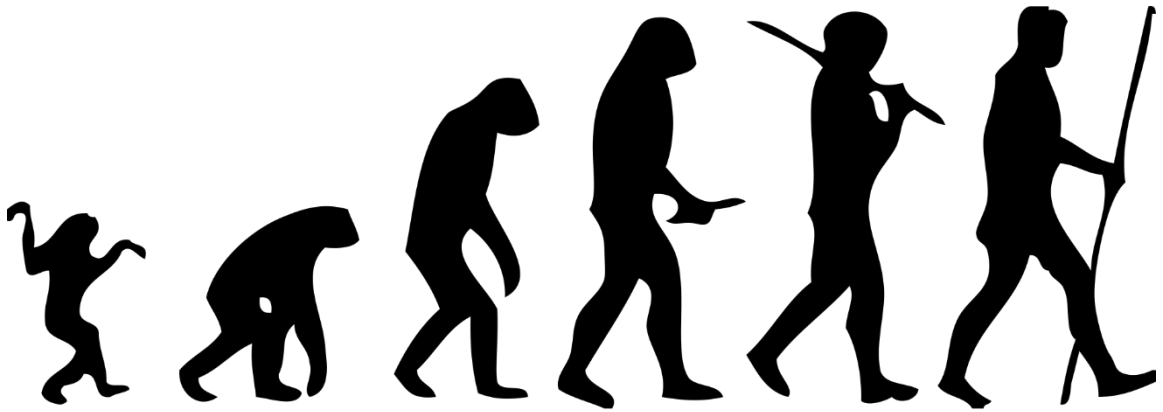
Fragment of BRAF protein. Source: UniProtKB, P15056 (BRAF_HUMAN)



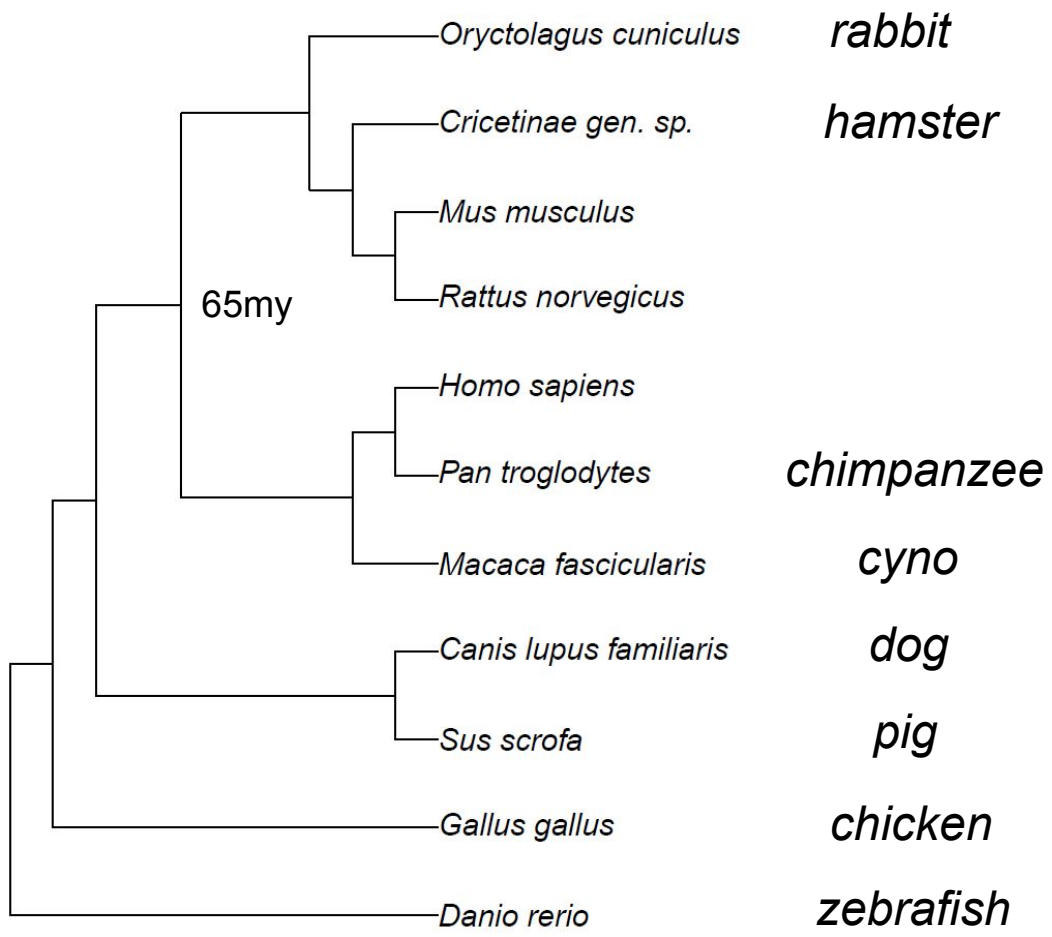
- View the 3D structure of the molecule at PDB ligand database
- View the X-ray structure of BRAF in complex with PLX4032 on PDB: accession number 3OG7.

Source: https://commons.wikimedia.org/wiki/File:Vemurafenib_structure.svg

Evolution: what is wrong with this figure?

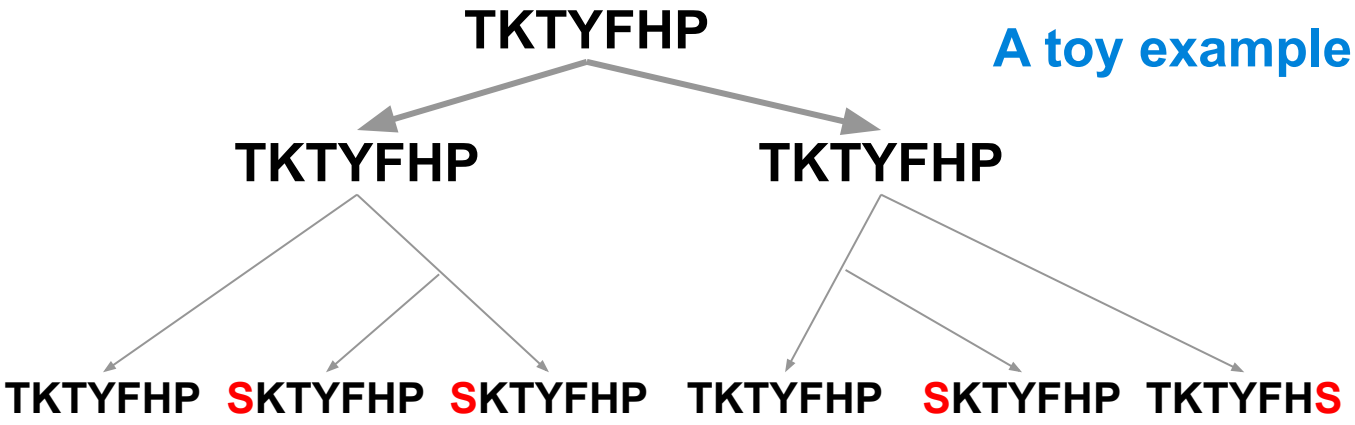


Phylogeny of commonly used species for animal studies

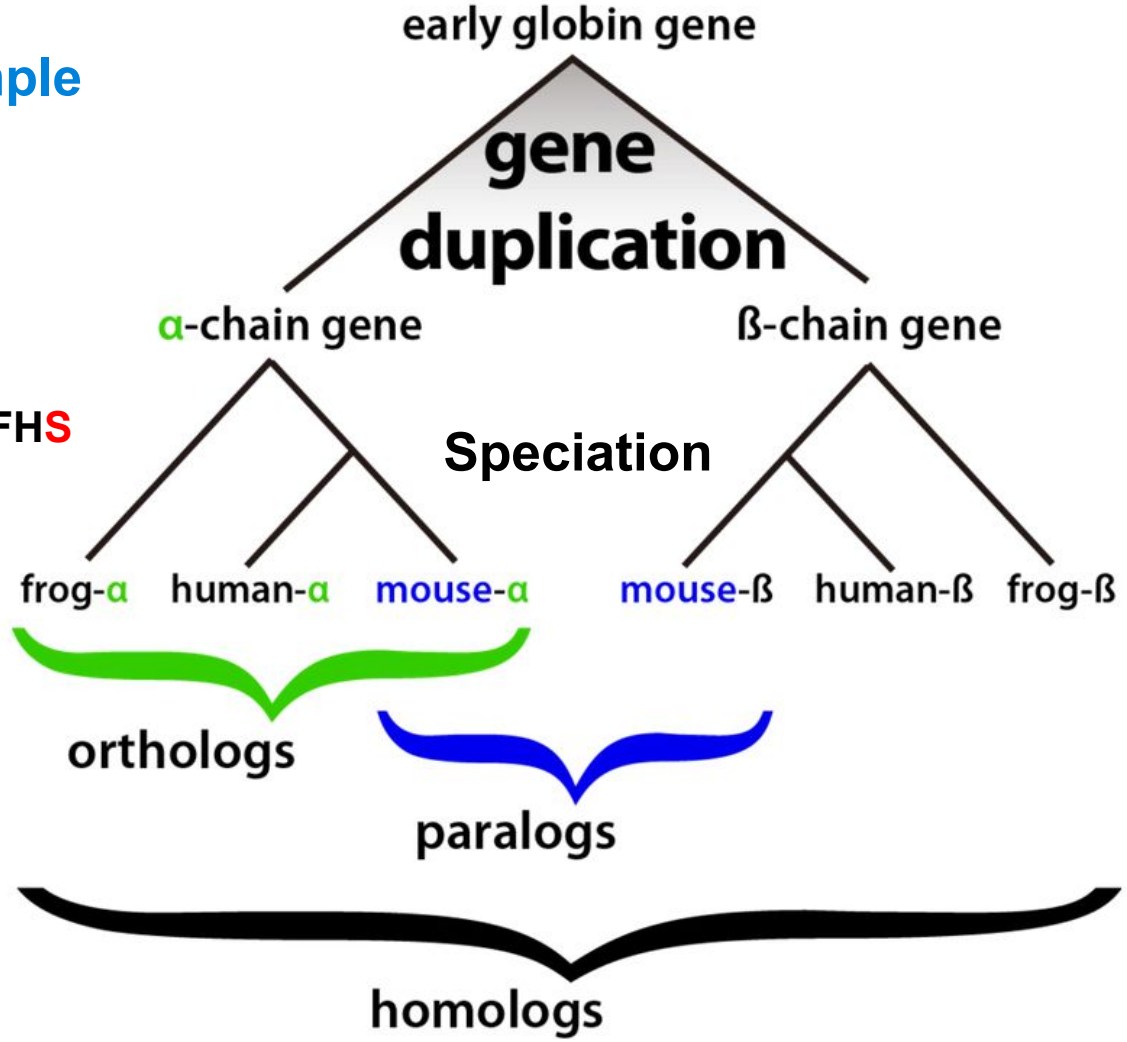


Tree structure retrieved from https://itol.embl.de/ (iTOL, Interactive Tree of Life), visualized with the *FigTree* software developed by Andrew Rambaut. Information of common ancestor of human and mouse is found via MGI.

Mutations in biological sequences make species in millions of years



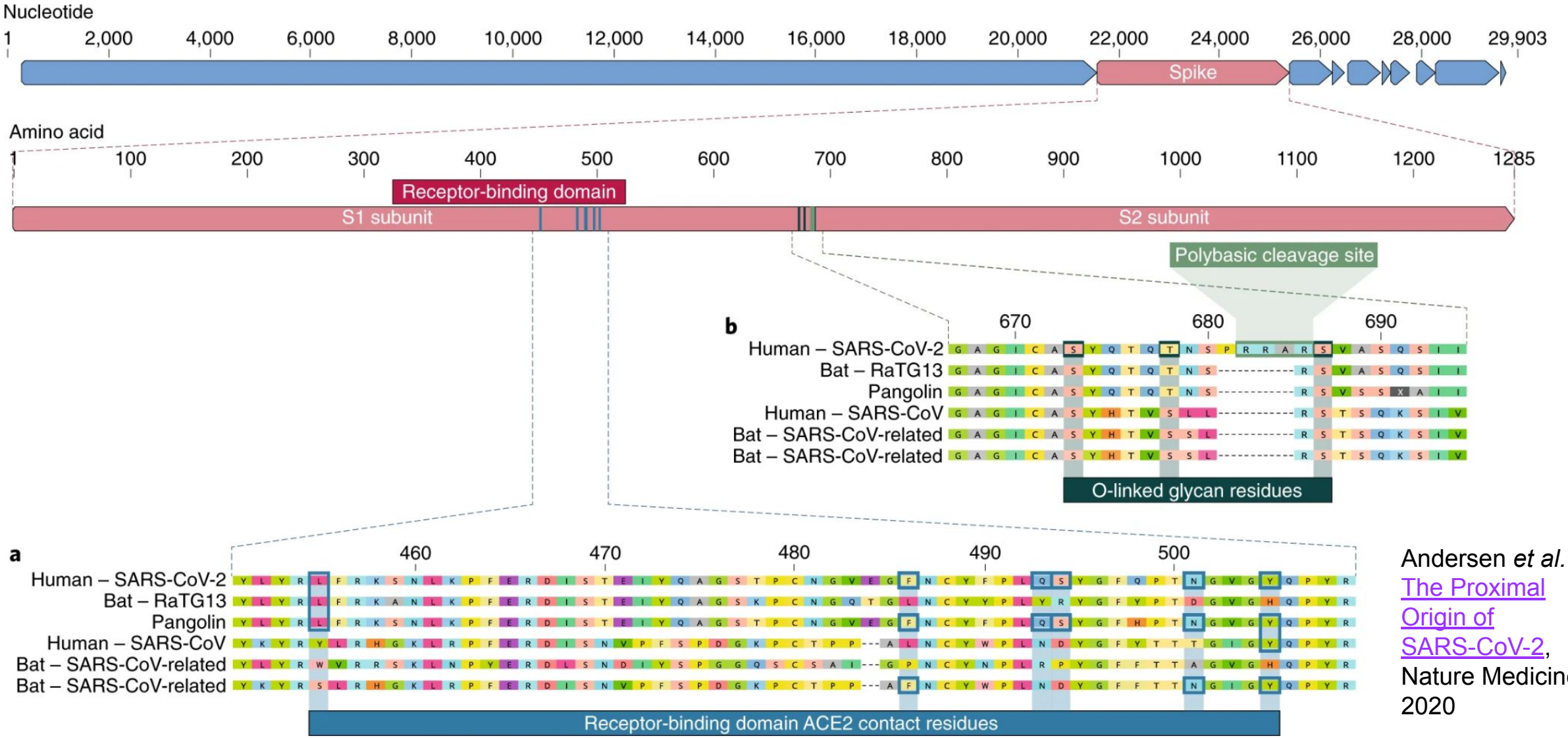
A toy example



- **Homologs**: two genes related by descent from a common ancestral gene.
- **Orthologs**: homologous genes in different species, which are evolved from a single ancestral gene by **speciation.**
- **Paralogs**: Two genes **in a genome** that are related by **duplication**.

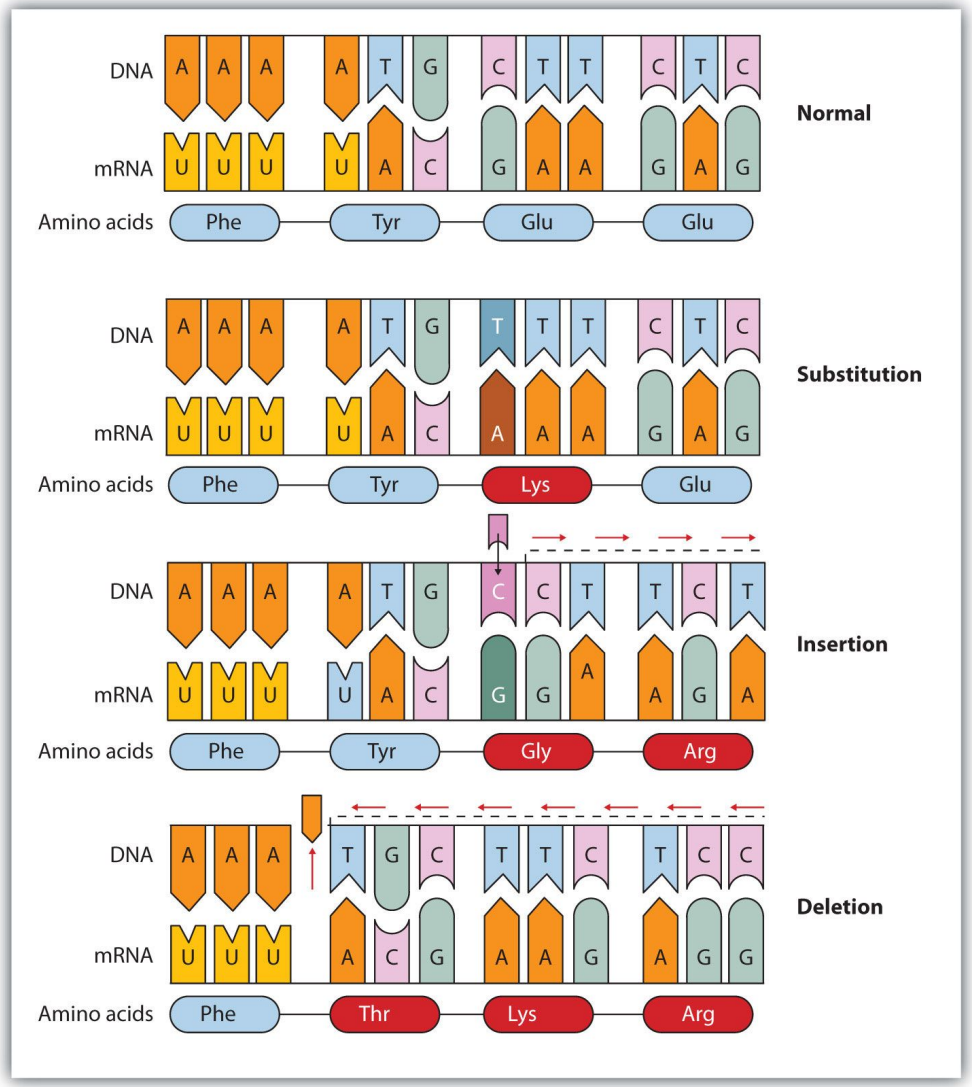
Adapted from Popo H. Liao, CC BY-SA 3.0, via Wikimedia Commons

Mutations in biological sequences are driving forces of evolution



Andersen *et al.*, The Proximal Origin of SARS-CoV-2, Nature Medicine, 2020

Mutations in biological sequences cause diseases in a lifetime



Disease	Responsible Protein or Enzyme
alkaptonuria	homogentisic acid oxidase
galactosemia	galactose 1-phosphate uridyl transferase, galactokinase, or UDP galactose epimerase
Gaucher disease	glucocerebrosidase
gout and Lesch-Nyhan syndrome	hypoxanthine-guanine phosphoribosyl transferase
hemophilia	antihemophilic factor (factor VIII) or Christmas factor (factor IX)
homocystinuria	cystathionine synthetase
maple syrup urine disease	branched chain α-keto acid dehydrogenase complex
McArdle syndrome	muscle phosphorylase
Niemann-Pick disease	sphingomyelinase
phenylketonuria (PKU)	phenylalanine hydroxylase
sickle cell anemia	hemoglobin
Tay-Sachs disease	hexosaminidase A
tyrosinemia	fumarylacetoacetate hydrolase or tyrosine aminotransferase
von Gierke disease	glucose 6-phosphatase
Wilson disease	Wilson disease protein

Three types of point mutations in DNA, The Basics of General, Organic, and Biological Chemistry

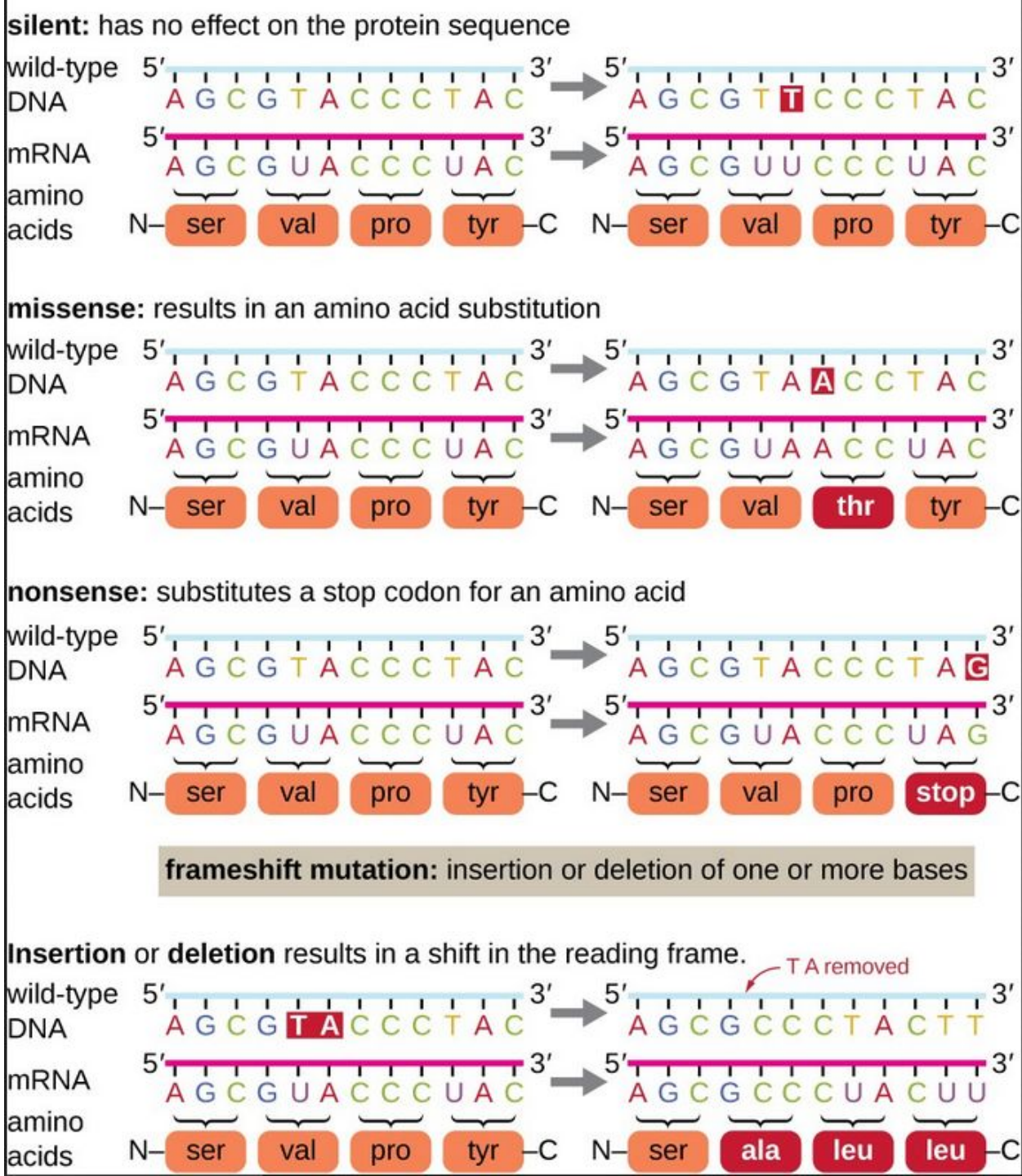
Loss-of-function (LoF) mutations

Nonsense, frameshift & splice site alterations

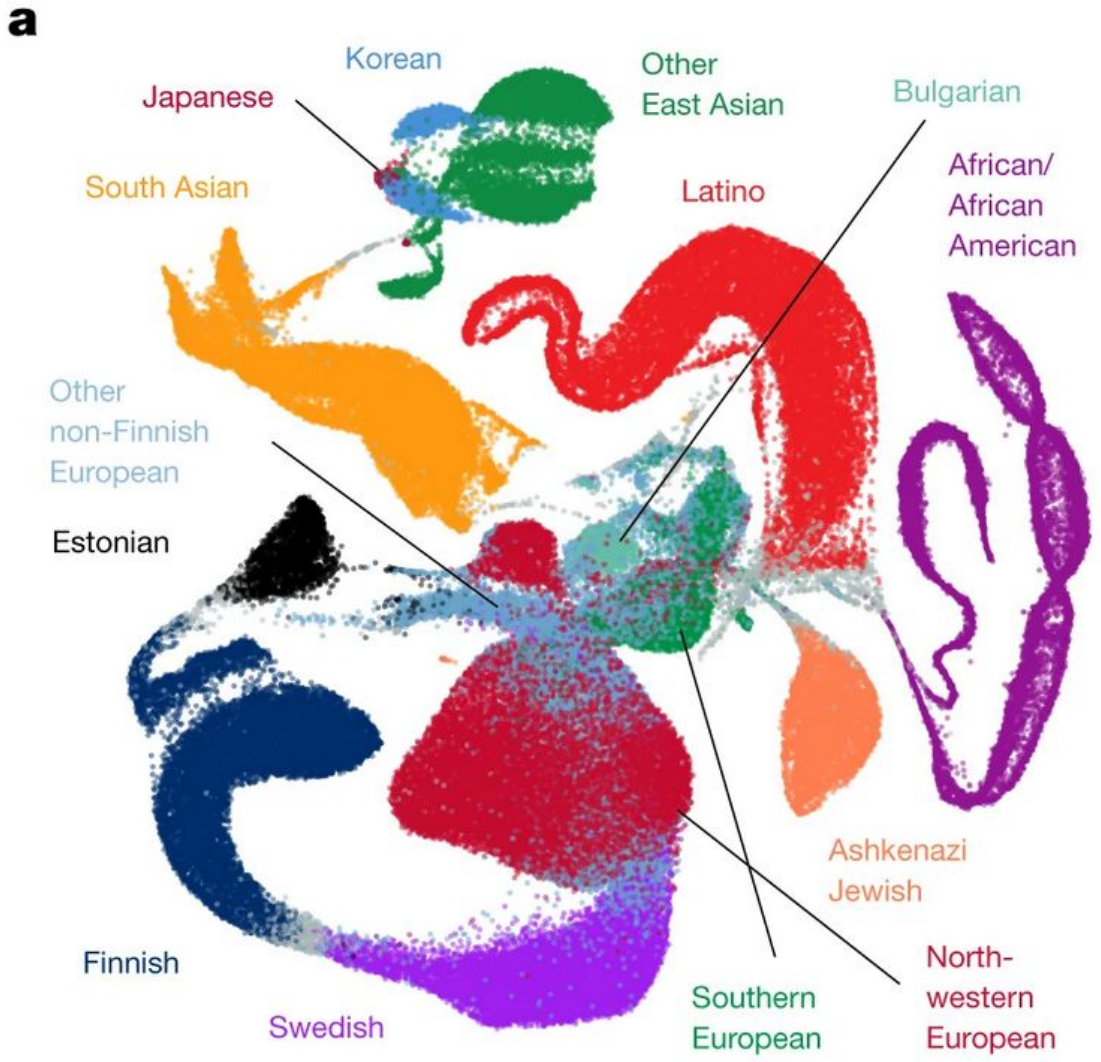
Many disease-causing mutations are Loss-of-Function (LoF) genetic variants. There are three types of LoF variants:

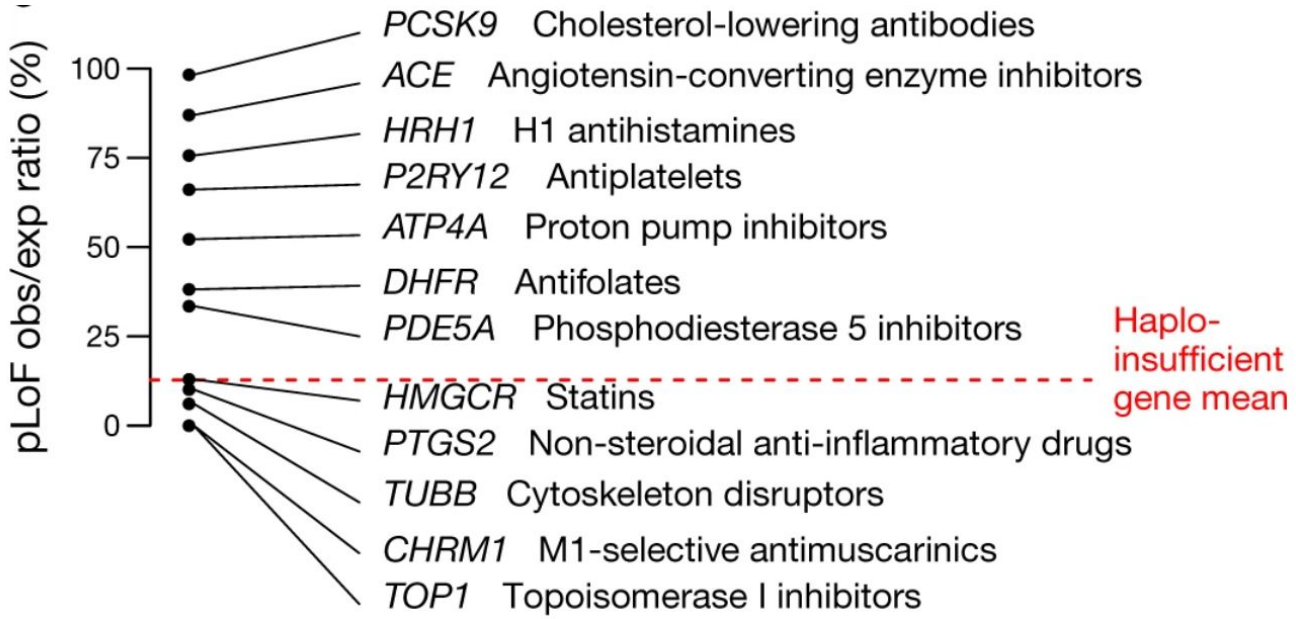
- Nonsense
- Frameshift
- Splice-site alterations (explained later)

Predicted Loss-of-Function (pLoF) genetic variants are pravelant (up to 800 per individual genome). Not all of them are truly pathogenic, because (1) sequencing and annotation error, (2) buffering effects of paralogs, and and (3) other, unknown reasons.



Mutations in biological sequences in human (since 200,000 years ago) can be used to identify *mutational constraints*





Left: Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. “The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans.” Nature 581, no. 7809 (May 2020): 434–43. https://doi.org/10.1038/s41586-020-2308-7.

Right: Minikel, Eric Vallabh, Konrad J. Karczewski, Hilary C. Martin, Beryl B. Cummings, Nicola Whiffin, Daniel Rhodes, Jessica Alföldi, et al. “Evaluating Drug Targets through Human Loss-of-Function Genetic Variation.” Nature 581, no. 7809 (May 2020): 459–64. https://doi.org/10.1038/s41586-020-2267-z.

Edit distance: a deterministic view of distance between two sequences

	Insertion	Deletion	Substitution	Transposition	Note
The Levenshtein distance	Allowed	Allowed	Allowed	Not allowed	
The longest common subsequence (LCS) distance	Allowed	Allowed	Not allowed	Not allowed	
The Hamming distance	Not allowed	Not allowed	Allowed	Not allowed	
The Damerau-Levenshtein distance	Allowed	Allowed	Allowed	Allowed (adjacent characters)	Not a distance metric, because triangle inequality is not satisfied
The Jaro-Winkler distance	Not allowed	Not allowed	Not allowed	Allowed	Not a distance metric

The Levenshtein distance best models changes in biological sequences

With *Levenshtein distance* we can compare any two pieces of DNA

Levenshtein distance: The minimum number of operations required to transform string a to string b with following operations:

- Insertion, *e.g.* **bat → bait**
- Deletion, *e.g.* **boat → bot**
- Substitution, *e.g.* **pig → big**

The Levenshtein distance between two strings a, b of length $|a|$ and $|b|$ respectively is given by $\mathrm{lev}_{a,b}(|a|, |b|)$ where

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

where $1_{(a_i \neq b_j)}$ is the indicator function equal to 0 when $a_i = b_j$ and equal to 1 otherwise, and $\mathrm{lev}_{a,b}(i,j)$ is the distance between the first i characters of a and the first j characters of b.

Offline activities

Required reading:

- Tsai, et al. “Discovery of a Selective Inhibitor of Oncogenic B-Raf Kinase with Potent Antimelanoma Activity.” PNAS (2008): 3041–46. https://doi.org/10.1073/pnas.0711741105.

Optional reading:

- Dolgin, Elie. “The Tangled History of MRNA Vaccines.” Nature 597, no. 7876 (September 14, 2021): 318–24. https://doi.org/10.1038/d41586-021-02483-w.

Offline activities (for Lecture 3 and Lecture 4):

- Exercises about (1) the genetic code, (2) the Levenshtein distance and dynamic programming, (3) about the BLAST program;
- Questions about the required readings
- Submit your replies via Google Form: https://forms.gle/VzGGbXBvy9ZhTSVK8

Backup slides

Calculate the Levenshtein distance with dynamic programming

What is the Levenshtein distance between `ATGC` and `AGC`?

		A	T	G	C
	0				
A					
G					
C					

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

where $1_{(a_i \neq b_j)}$ is the indicator function equal to 0 when $a_i = b_j$ and equal to 1 otherwise, and $\mathrm{lev}_{a,b}(i,j)$ is the distance between the first i characters of a and the first j characters of b.

Solution:

`ATGC`

`A-GC`

Dynamic programming breaks down a complex problem in sub-problems.
With a cost function (e.g. the Levenshtein distance), we enumerate all possible operations *at any position*, and record the operation that minimizes the cost. When we finish all positions, we find the solution(s) by working backwards the optimal path(s).

Calculate the Levenshtein distance with dynamic programming

What is the Levenshtein distance between `ACTGCTT` and `ACATT`?

		A	C	T	G	C	T	T
	0	1	2	3	4	5	6	7
A	1	0	1	2	3	4	5	6
C	2	1	0	1	2	3	4	5
A	3	2	1	1	2	3	4	5
T	4	3	2	1	2	3	3	3
T	5	4	3	2	2	3	3	3

```
ACTGCTT          ACTGCTT          ACTGCTT
AC--ATT          ACA--TT          AC-A-TT
```

The Needleman-Wunsch algorithm uses dynamic programming for *global alignment* of two sequences

Compared with the Levenshtein distance, the Needleman-Wunsch algorithm uses biologically meaningful parameters to score insertion or deletion (gap penalty *d*), and substitution or mutation events (a substitution matrix *M*). The dynamic programming technique is used in a similar way.

Task: align two sequences `ATCGAC` and `CATAC`.

Parameters: d=-4, $M = \begin{pmatrix} & A & C & T & G \\ A & 5 & -3 & -3 & -3 \\ C & -3 & 5 & -3 & -3 \\ T & -3 & -3 & 5 & -3 \\ G & -3 & -3 & -3 & 5 \end{pmatrix}$

Solution:

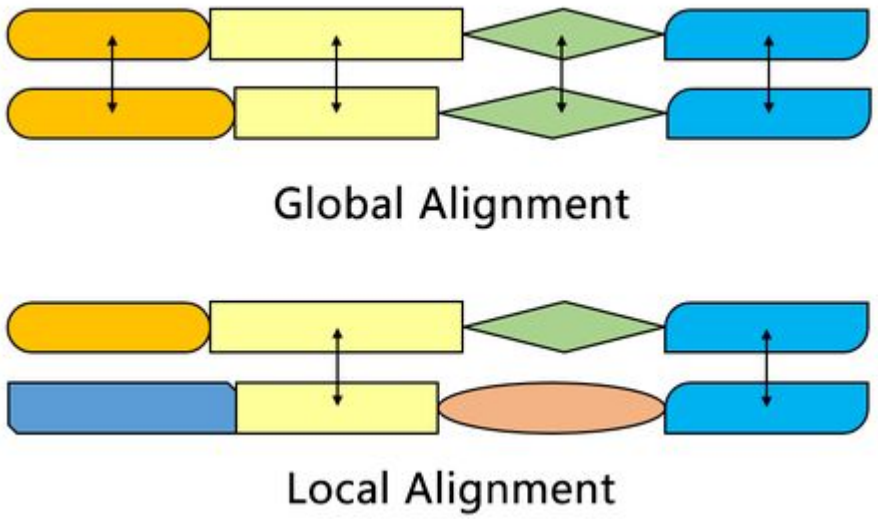
```
ATCGAC
||  ||
CAT--AC
```

-	-	A	T	C	G	A	C
-	0	-4	-8	-12	-16	-20	-24
C	-4	-3	-7	-3	-7	-11	-15
A	-8	1	-3	-7	-6	-2	-6
T	-12	-3	6	2	-2	-6	-5
A	-16	-7	2	3	-1	3	-1
C	-20	-11	-2	-1	0	-1	8

Source: https://commons.wikimedia.org/wiki/File:Needleman-wunsch.jpg

Check this video out if a step-by-step tutorial may help you: Link to YouTube (thanks to the contribution of Robert Do)

The Smith-Waterman algorithm uses dynamic programming for *local alignment* of two sequences



	Needleman-Wunsch	**Smith-Waterman**
Initialization	Gap penalty in first column and first row	0 in first column and first row
Scoring	Scores can be negative	Negative scores are set to 0
Traceback	Begin at the bottom right, and end at the top left cell.	Begin at the cell with the highest score, end when 0 is met.

Major differences between the Needleman-Wunsch and the Smith-Waterman algorithm. See this animation for an example.

Richard Bell on the origin of the name *Dynamic Programming*

I spent the Fall quarter (of 1950) at RAND. My first task was to find a name for multistage decision processes. An interesting question is, Where did the name, dynamic programming, come from?

The 1950s were not good years for mathematical research. We had a very interesting gentleman in Washington named Wilson. He was Secretary of Defense, and he actually had a pathological fear and hatred of the word, research. I'm not using the term lightly; I'm using it precisely. His face would suffuse, he would turn red, and he would get violent if people used the term, research, in his presence. You can imagine how he felt, then, about the term, mathematical. The RAND Corporation was employed by the Air Force, and the Air Force had Wilson as its boss, essentially. Hence, I felt I had to do something to shield Wilson and the Air Force from the fact that I was really doing mathematics inside the RAND Corporation. What title, what name, could I choose? In the first place I was interested in planning, in decision making, in thinking. But planning, is not a good word for various reasons. I decided therefore to use the word, "programming" I wanted to get across the idea that this was dynamic, this was multistage, this was time-varying I thought, lets kill two birds with one stone. Lets take a word that has an absolutely precise meaning, namely dynamic, in the classical physical sense. It also has a very interesting property as an adjective, and that is its impossible to use the word, dynamic, in a pejorative sense. Try thinking of some combination that will possibly give it a pejorative meaning. Its impossible. Thus, I thought dynamic programming was a good name. It was something not even a Congressman could object to. So I used it as an umbrella for my activities.

Dreyfus, Stuart. "Richard Bellman on the Birth of Dynamic Programming." Operations Research 50, no. 1 (February 2002): 48–51. https://doi.org/10.1287/opre.50.1.48.17791.

Sequence alignment is fundamental for many bioinformatics tasks and tools

Examples:

- BLAST (Basic Local Alignment Search Tool) is used for almost all biological sequence analysis tasks. At its core, a heuristic algorithm approximates the Smith-Waterman algorithm for local alignment.
- Software tools such as Bowtie/Bowtie2 (Langmead *et al.*, Genome Biology, 2009; source code on GitHub), STAR (Dobin *et al.*, Bioinformatics, 2013; source code on GitHub) and GSNAP (web link) use more sophisticated methods to map sequencing reads (usually ~30-200 nucleotides) to large genomes (*e.g.* ~10^9 base pairs of mouse or human).

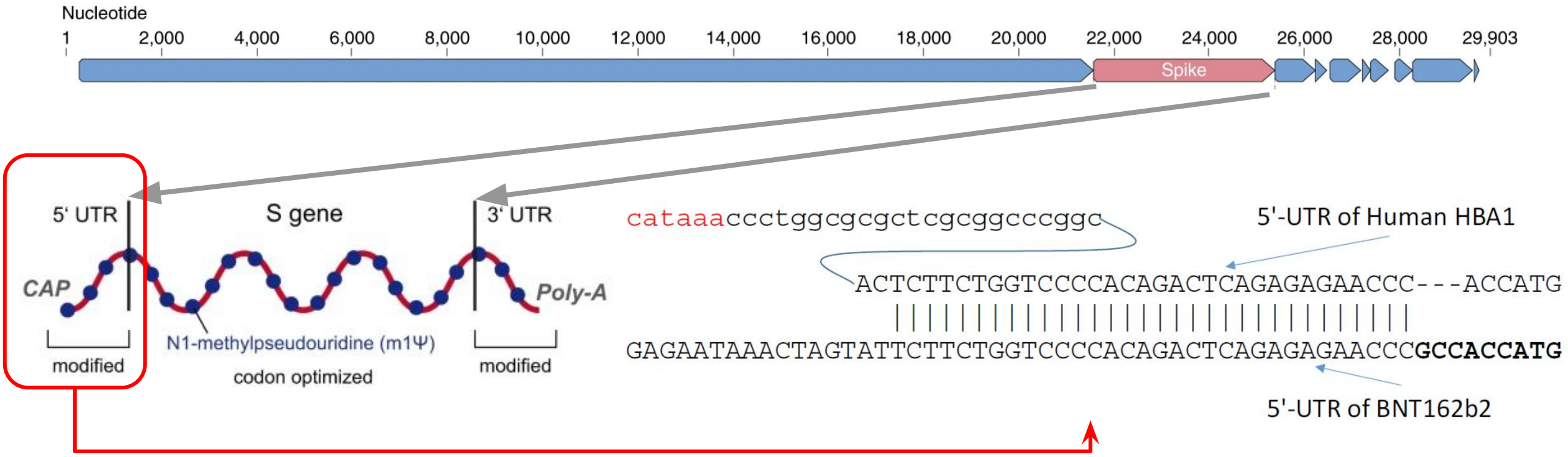
A typical case for Blast: we have a RNA sequence (see below). How can we know the original genome of the sequence, and ideally the gene encoding the sequences?

```
ATGTTTGTTTTTCTTGTTTTATTGCCACTAGTCT
CTAGTCAGTGTGTTAATCTTACAACCAGAACTCA
ATTACCCCCTGCATACACTAATTCTTTCACACGT
GGTGTTTATTACCCTGACAAAGTTTTCAGATCCT
CAGT
```

Tip: go to the NCBI BLAST tool (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome), copy and paste the sequence as the query sequence, and try your luck. The default parameter would do.

The Wiki page of BLAST is a good start to understand how it works

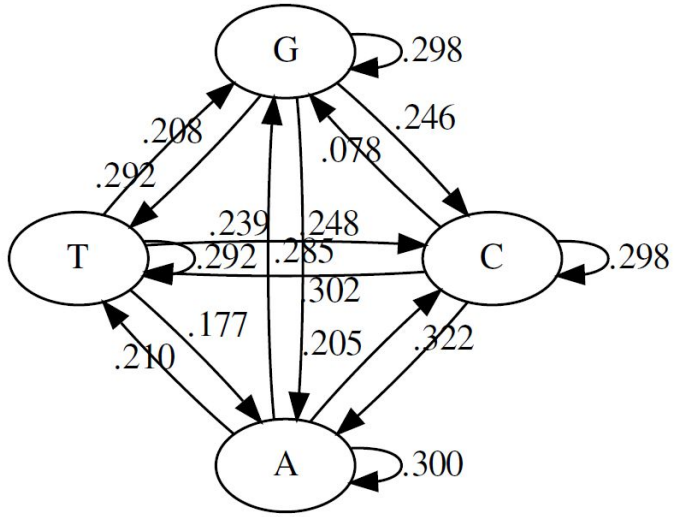
Further applications of biological sequence analysis in drug discovery: sequence-based drug design



References: Heinz, Franz X., and Karin Stiasny. "Distinguishing Features of Current COVID-19 Vaccines: Knowns and Unknowns of Antigen Presentation and Modes of Action." Npj Vaccines 6, no. 1 (August 16, 2021): 1–13. https://doi.org/10.1038/s41541-021-00369-6; Assemblies of putative SARS-CoV2-spike-encoding mRNA sequences for vaccines BNT-162b2 and mRNA-1273 (github.com/NAalytics); Xia, Xuhua. "Detailed Dissection and Critical Evaluation of the Pfizer/BioNTech and Moderna MRNA Vaccines." Vaccines 9, no. 7 (July 3, 2021): 734. https://doi.org/10.3390/vaccines9070734.

A probabilistic view of biological sequence analysis with Markov chains

- A **discrete-time** Markov chain is a sequence of random variables with the Markov property, namely that the probability of moving to the next state depends only on the present state and not on the previous states.
- A Markov chain is often represented by either **a directed graph** or **a transition matrix**.
- Two sides of application
  - Given a string, assuming that the Markov chain model is suitable, we can easily construct a Markov chain, for instance by counting transitions and normalize the count matrix (variants possible).
  - Given a Markov chain model and a string, we can calculate the probability that the string is generated by the specific Markov chain model with the **chain rule of conditional probability.**

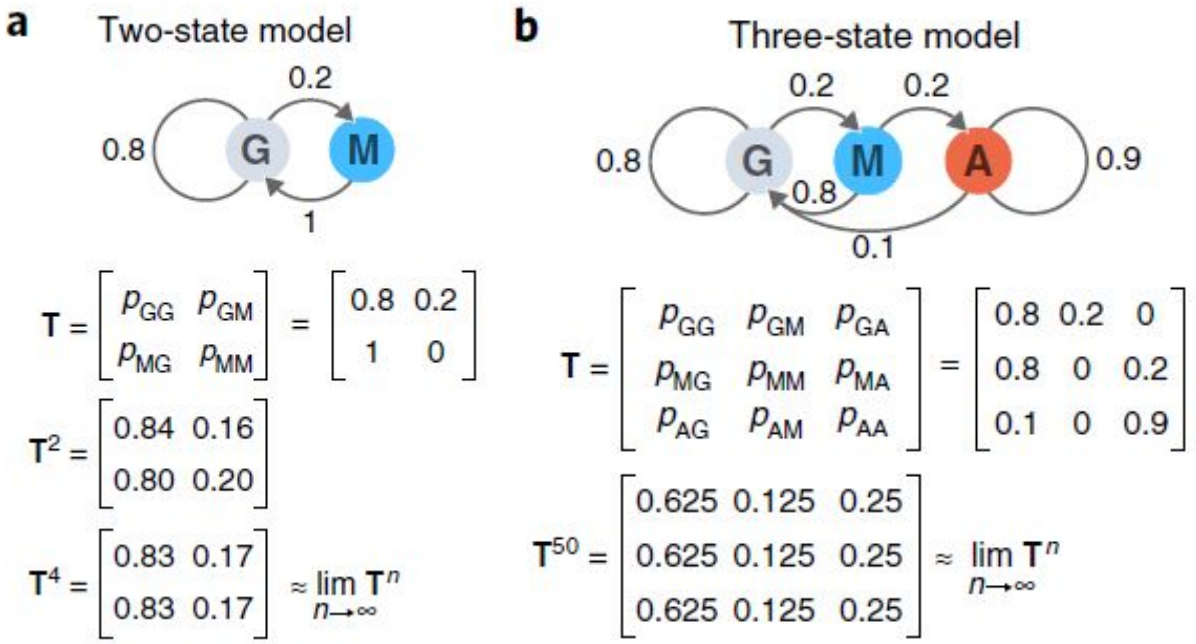


	A	C	G	T
A	.300	.205	.285	.210
C	.322	.298	.078	.302
G	.248	.246	.298	.208
T	.177	.239	.292	.292

Data: https://web.stanford.edu/class/stats366/exs/HMM1.html

Stationary distribution exist for ergodic (irreducible and aperiodic) Markov Chains

- A Markov Chain has stationary *n*-step transition probabilities, which are the *n*th power of the one-step transition probabilities. Namely, $P_n=P^n$.
- A stationary distribution π is a row vector whose entries are non-negative and sum to 1. It is unchanged by the operation matrix P on it, and is defined by πP=π.
  - In another word, it is the limit of the transition matrix multiplying itself.
  - Note that is has the form of the left eigenvector equation, $uA=\kappa u$, where κ is a scalar and u is a row vector. In fact, π is a normalised (sum to 1) multiple of a left eigenvector e of the transition matrix P with an eigenvalue of 1.
- Markov chains capture dependencies within a system and reveal interesting long-term behavior. They are subjects of the study of **stochastic processes.**

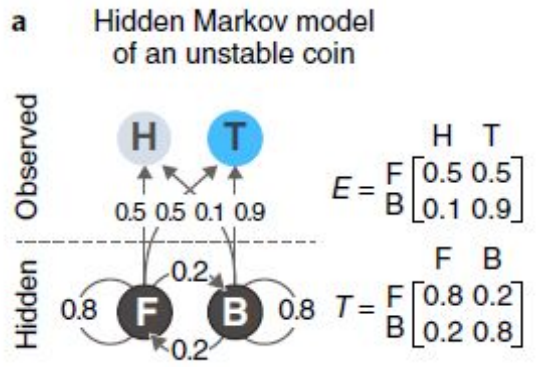


G=Growth, M=Mitosis, A=Arrest

Grewal, Jasleen K., Martin Krzywinski, and Naomi Altman. 2019. "Markov Models—Markov Chains." Nature Methods 16 (8): 663–64. https://doi.org/10.1038/s41592-019-0476-x.

Hidden Markov Chains

A Hidden Markov Model consists of two graphs (matrices): one of hidden states (corresponding to the transition matrix), and one of observed states (emitted by the hidden states according to the emission matrix). The Viterbi algorithm (based on dynamic programming), or the Baum-Welch algorithm (a special case of EM algorithms) is used to estimate its parameters.



A Hidden Markov Model of an unstable coin that has a 20% chance of switching between a fair state (F) and a biased state (B). Source: Grewal, Jasleen K., Martin Krzywinski, and Naomi Altman. 2019. "Markov Models — Hidden Markov Models." Nature Methods 16 (9): 795–96.

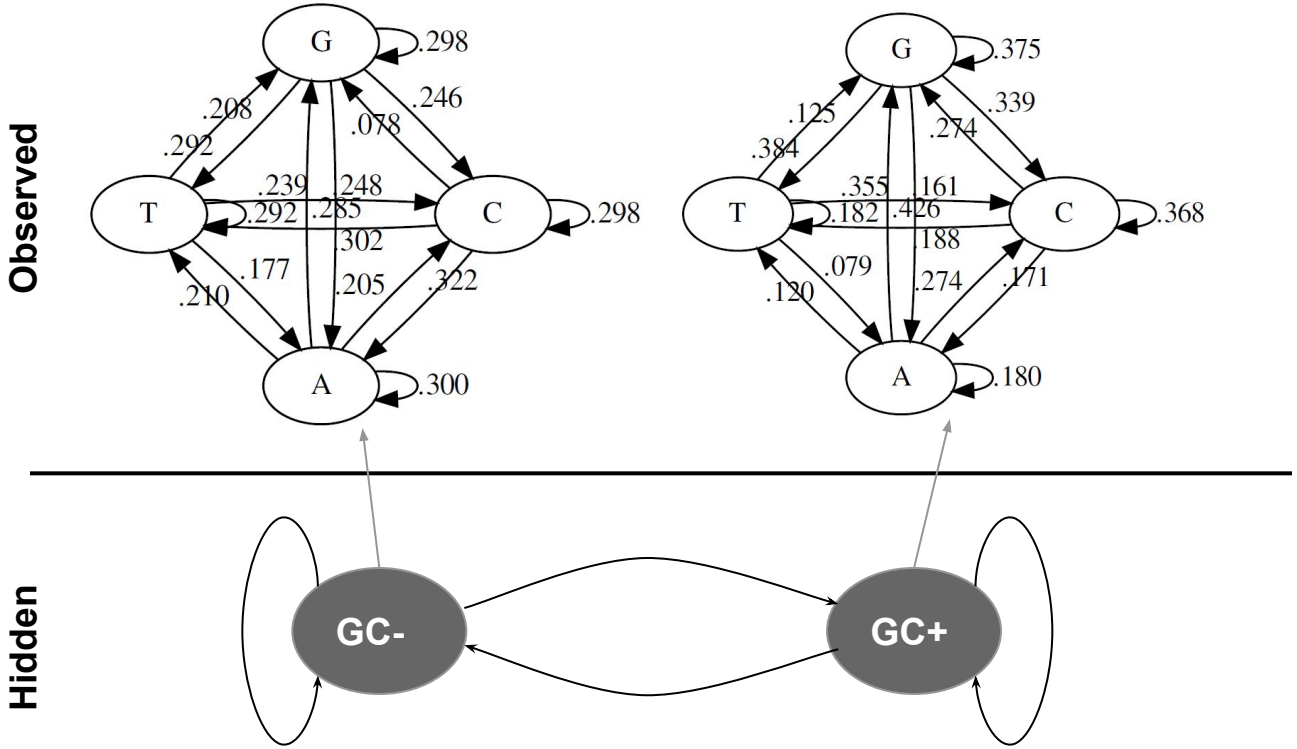
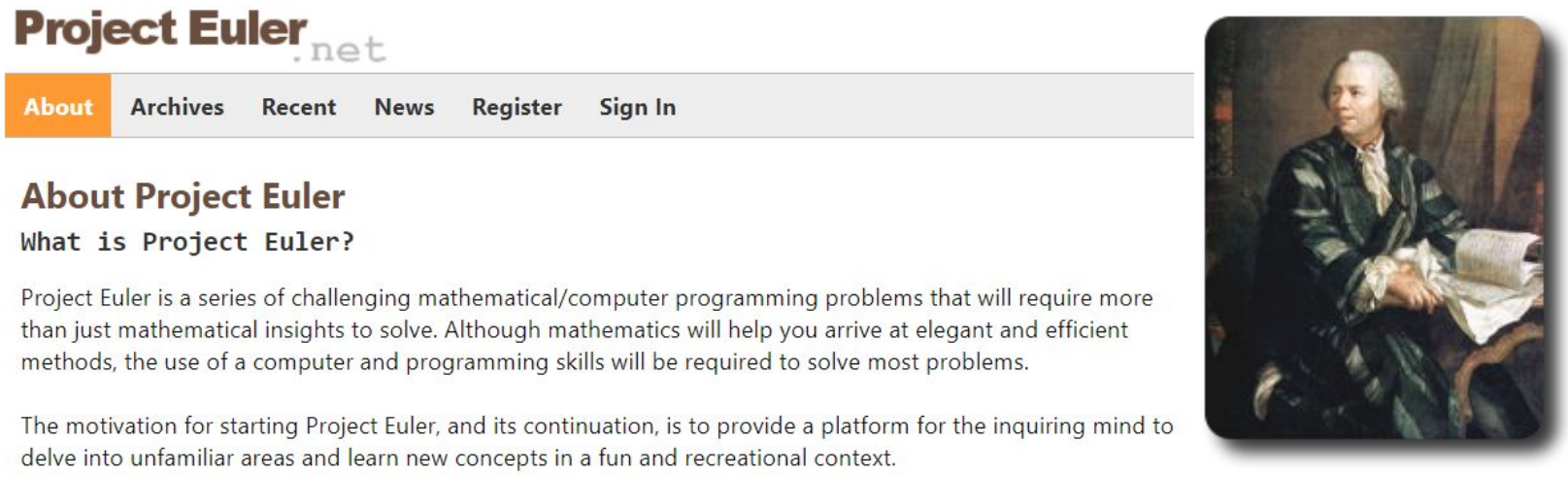


Illustration of a Hidden Markov Model predicting CpG islands in genomic sequences

Software tools

- **General biological sequence analysis**
  - EMBOSS software suite: http://emboss.sourceforge.net/, also available online at European Bioinformatics Institute (EBI): https://www.ebi.ac.uk/services
  - BLAST (=Basic Local Alignment Search Tool) can be run at many places, for instances from EBI and National Center for Biotechnology Information (NCBI): https://blast.ncbi.nlm.nih.gov/Blast.cgi
  - Programming access, for instance the Biopython project: https://biopython.org
- **RNA biology**
  - ViennaRNA package (https://www.tbi.univie.ac.at/RNA/)
  - RNA processing tools available at U Bielefeld, for instance RNAhybrid, which finds minimum free energy hybridization using dynamic programming (https://bibiserv.cebitec.uni-bielefeld.de/rnahybrid)
- **Profile Hidden Markov Models (HMMs)**
  - The HMMER package: http://hmmer.org/

The Euler Project



https://projecteuler.net/

- **Learning by problem-solving**
- **Free**
- **Math + CS**

Problem 1: Multiples of 3 and 5

If we list all the natural numbers below 10 that are multiples of 3 or 5, we get 3, 5, 6 and 9. The sum of these multiples is 23.

Find the sum of all the multiples of 3 or 5 below 1000.

Rosalind: a great scientist, and a platform for learning bioinformatics and programming through problem solving

ROSALIND http://rosalind.info/problems/locations/

Rosalind Elsie Franklin

1920-1958

A Rapid Introduction to Molecular Biology click to expand

Problem

A **string** is simply an ordered collection of symbols selected from some **alphabet** and formed into a word; the **length** of a string is the number of symbols that it contains.

An example of a length 21 **DNA string** (whose alphabet contains the symbols 'A', 'C', 'G', and 'T') is "ATGCTTCAGAAAGGTCTTACG."

Given: A DNA string s of length at most 1000 nt.

Return: Four integers (separated by spaces) counting the respective number of times that the symbols 'A', 'C', 'G', and 'T' occur in s.

Sample Dataset

```
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
```

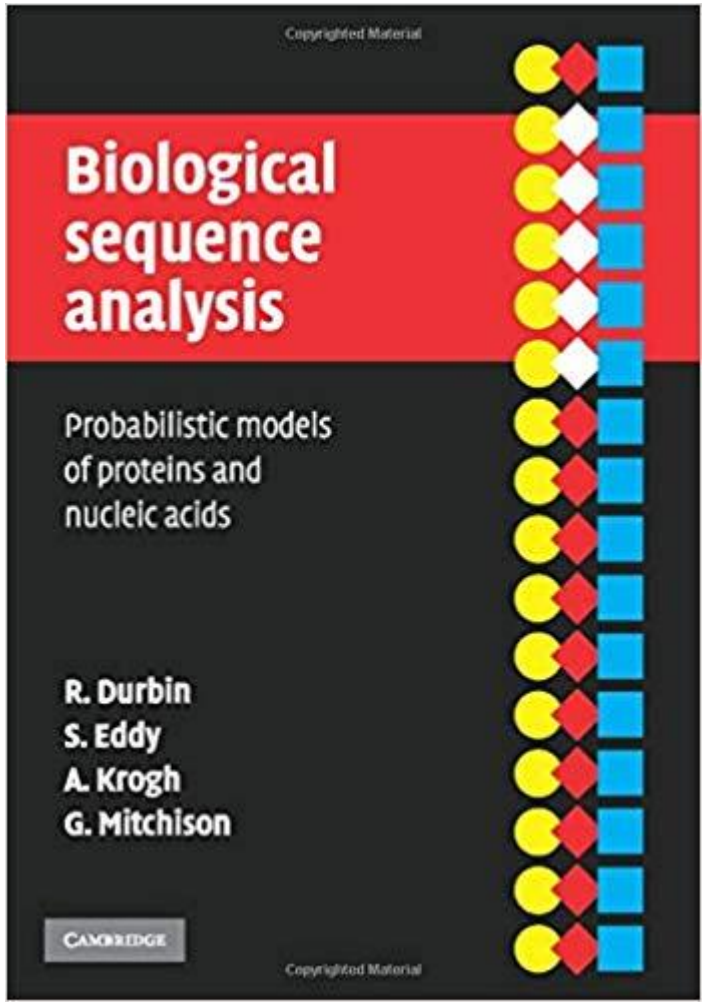
Sample Output

```
20 12 17 21
```

Please login to solve this problem.

Further resources

***Biological Sequence Analysis* by Durbin, Eddy, Krogh, and Mitchison**



Teaching RNA algorithms by the Backofen Lab at U Freiburg, with source codes available on GitHub.

The website hosts among others an interactive tool to visualize how dynamic programming (DP) helps to predict RNA secondary structure.

For a gentle introduction, see also *How Do RNA Folding Algorithms Work?* by Eddy, Sean R, *Nature Biotechnology* 22, Nr. 11 (November 2004): 1457–58. https://doi.org/10.1038/nbt1104-1457.

An Introduction to Applied Bioinformatics by Greg Caporaso (NAU)

The tutorial is written in Python using Jupyter. It introduces concepts in (a) pairwise sequence alignment, (b) sequence homology searching, (c) generalized dynamic programming for multiple sequence alignment, (d) phylogenetic reconstruction, (e) sequence mapping and clustering, as well as (f) machine learning in bioinformatics. Applications and exercises are also available.

Summary and Q&A

- Biological sequence analysis is used throughout the drug discovery process, and is essential for molecular modelling in the multiscale-modelling view of drug discovery.
- We explored both deterministic views of sequence analysis, with the example of Levenshtein distance, and probabilistic views, with the example of Markov chains and hidden Markov chains.
- Mathematical techniques such as dynamic programming, when implemented as algorithms software tools, are important for many tasks. We discussed in particular the Needleman-Wunsch algorithm, the Smith-Waterman algorithm, the BLAST software, and sequencing read alignment tools such as Bowtie2, STAR, and GSNAP.
- We provided further resources for further study and exploration.

Continuous-time Markov Chains

- Continuous-time Markov Chains are used for **phylogenetic analysis**, for instance of orthologous genes and of bacterial/viral genomes. They satisfy the Markovian property: $P(t+\tau)=P(t)P(\tau)$.
- The process makes a transition from the current state *i* after an amount of time modelled by an *exponential random variable* E_i, known as the **holding time**. Random variables of each state is independent.
- When a transition is made, the process moves according to the **jump chain**, a discrete-time Markov chain with a transition matrix.
- If there are *n* states, then at the time of transition, there are *n-1* competing exponentials. Since the distribution of the minimum of exponential random variables is also exponential, the continuous-time Markov chain changes its state from *i* according to a parameter $E_{i,j} \sim \text{Exp}(q_{i,j})$ ($i \neq j$). The parameters are known as the Q-matrix, or the **rate matrix**. The transition rate *E* is the product of holding time and the transition probability.
- Whereas the row sums of a transition matrix are always 1, the row sums of a rate matrix are always zero.

Given the transition matrix

$$P(t) = \begin{pmatrix} p_{AA}(t) & p_{AG}(t) & p_{AC}(t) & p_{AT}(t) \\ p_{GA}(t) & p_{GG}(t) & p_{GC}(t) & p_{GT}(t) \\ p_{CA}(t) & p_{CG}(t) & p_{CC}(t) & p_{CT}(t) \\ p_{TA}(t) & p_{TG}(t) & p_{TC}(t) & p_{TT}(t) \end{pmatrix}$$

We model the probability of seeing the same alphabet after a small increment of time as the sum of the starting probability, minus its loss, and plus its gain

$$\mu_x = \sum_{y \neq x} \mu_{xy}$$

$$p_A(t + \Delta t) = p_A(t) - p_A(t)\mu_A \Delta t + \sum_{x \neq A} p_x(t)\mu_{xA}\Delta t\,.$$

$$\mathbf{p}(t + \Delta t) = \mathbf{p}(t) + \mathbf{p}(t)Q\Delta t\,,$$

where

$$Q = \begin{pmatrix} -\mu_A & \mu_{AG} & \mu_{AC} & \mu_{AT} \\ \mu_{GA} & -\mu_G & \mu_{GC} & \mu_{GT} \\ \mu_{CA} & \mu_{CG} & -\mu_C & \mu_{CT} \\ \mu_{TA} & \mu_{TG} & \mu_{TC} & -\mu_T \end{pmatrix}$$

Source: https://en.wikipedia.org/wiki/Models_of_DNA_evolution

Interaction of drug and target: an example with HIV-1 Protease Inhibitor

Protein atoms: ball and stick, in blue and green

The small-molecule drug: ball and stick with traditional atomic coloration

Water: small red spheres