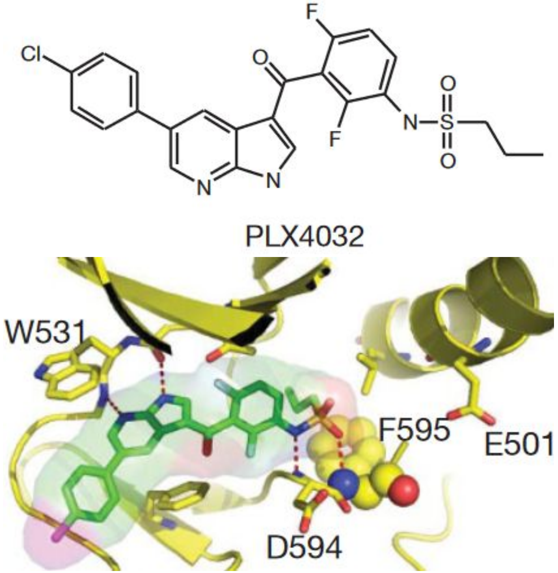
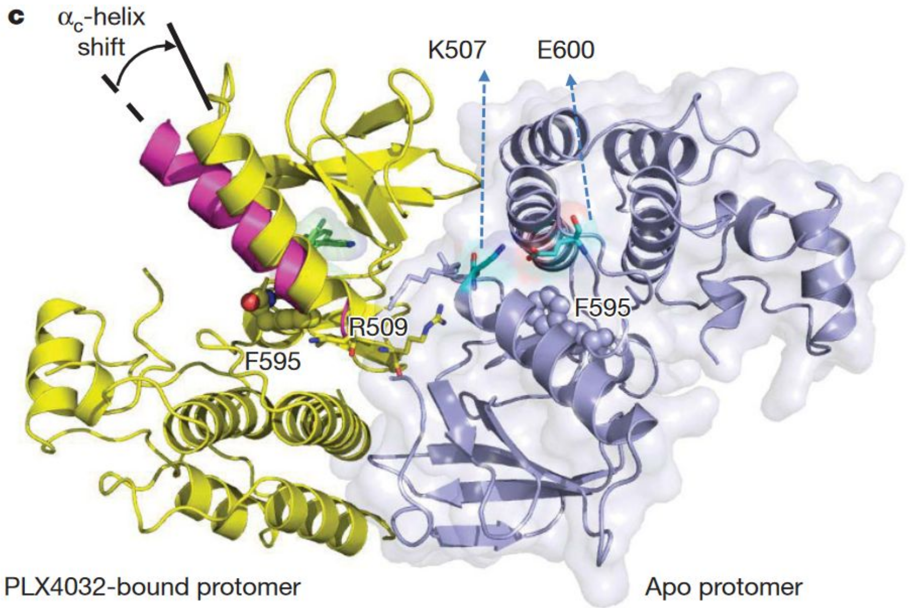
# AMIDD 2023 Lecture 7: Biological Sequence Analysis





*Dr. Jitao David Zhang, Computational Biologist*

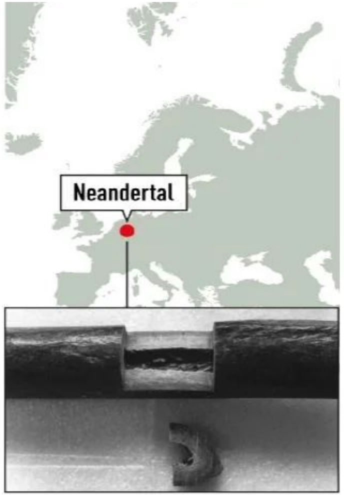
*[1] Pharmaceutical Sciences, Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche*

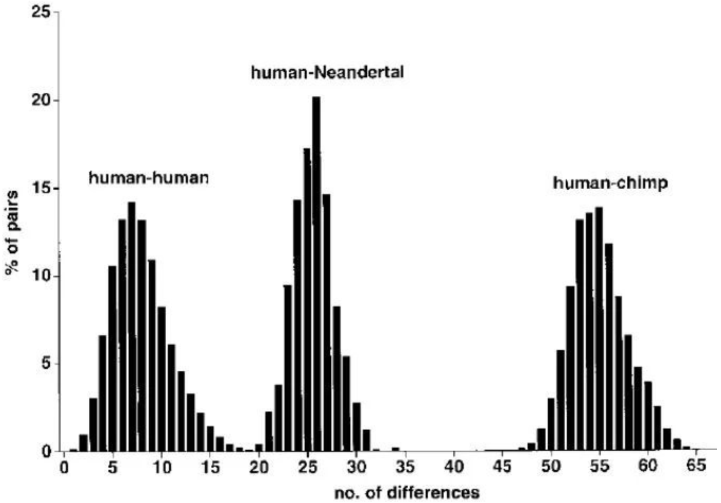
*[2] Department of Mathematics and Informatics, University of Basel*

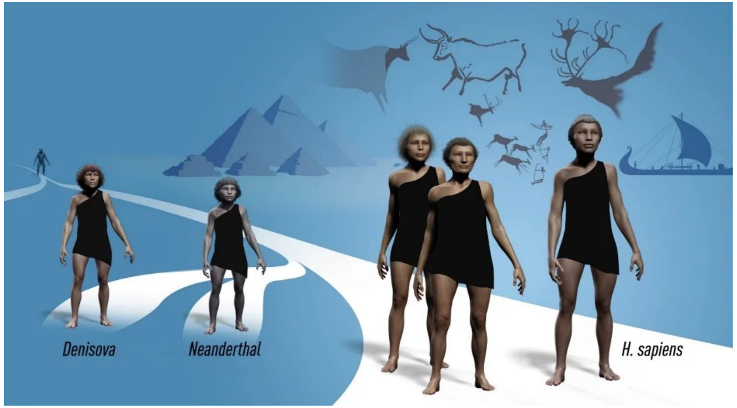
# Today’s goals

- Learning drug-discovery concepts with the help of Bollag *et al.*
- Biological sequence analysis
  - Levenshtein distance and dynamic programming
  - Markov chain and hidden markov chains
  - Convolutional neural networks
- Selected applications of biological sequence analysis for drug discovery

# Nobel Prize in Physiology or Medicine 2022 was awarded to Svante Pääbo “for his discoveries concerning the genomes of extinct hominins and human evolution”







**Left**: Sample removed from the right humerus bone of the Neanderthal-type specimen.

**Middle**: Pairwise differences of mitochondria sequences among Humans, the Neandertal, and Chimpanzees

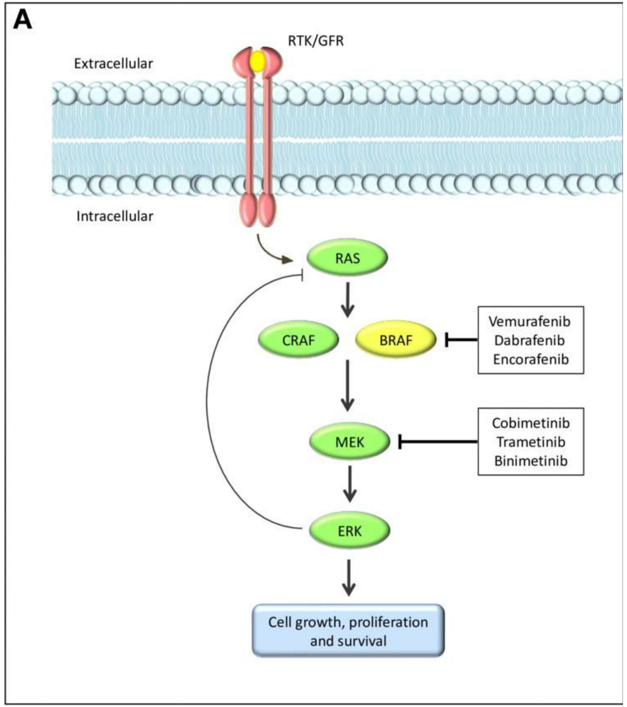
**Right**: A Schematic Phylogenetic Tree of Denisova, Neanderthal, and H. *sapiens.*

Krings, Matthias, Anne Stone, Ralf W. Schmitz, Heike Krainitzki, Mark Stoneking, and Svante Pääbo. “Neandertal DNA Sequences and the Origin of Modern Humans.” Cell 90, no. 1 (July 11, 1997): 19–30; https://www.nobelprize.org/prizes/medicine/

# 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

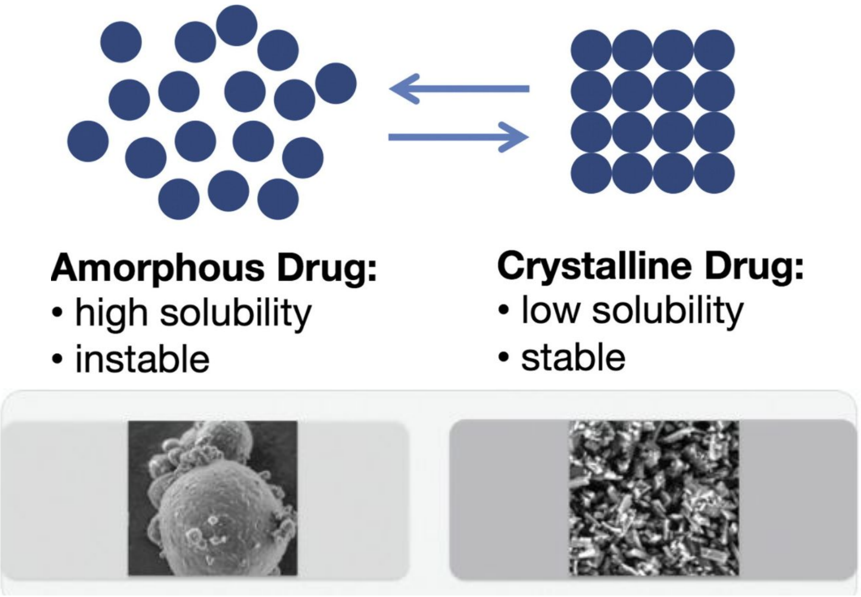


RTK=receptor tyrosine kinase;
GFR=growth factor receptor
RAS/RAF/MEK/ERK=MAP kinases

**(A)** normal pathway; **(B)** the most common resistance mechanisms, among others BRAF amplification and alternative splicing. 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*).



**Amorphous Drug:**
- high solubility
- instable

**Crystalline Drug:**
- low solubility
- stable

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.

# 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.

# Selected comments and questions

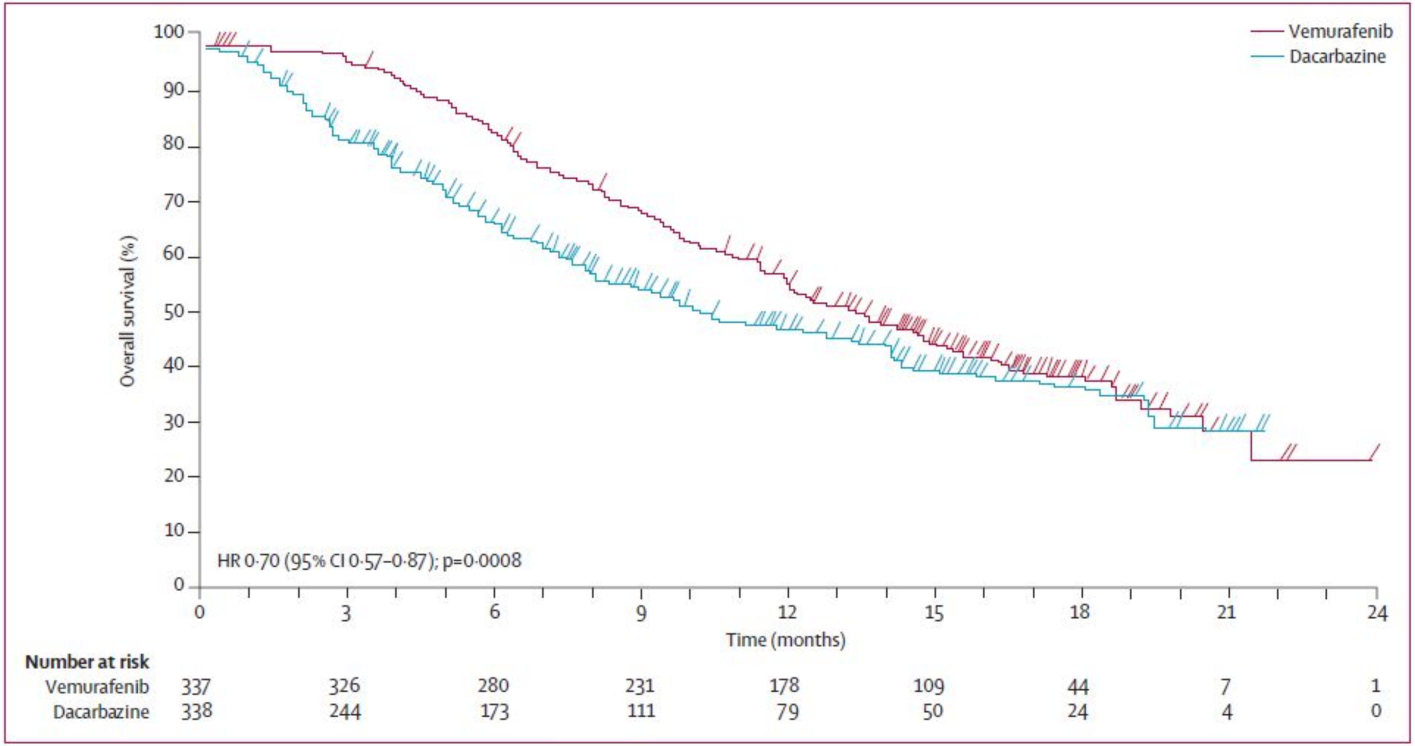
- *I'm no animalist, but why test on dogs and not other animals that aren't as affectionate to humans? Certainly the psychological trauma of being treated badly by us must be greater than for other animals.*
- *As a non biologist, it was quite a wall of text and weird words.*
- *I learned that abstract should be contain only the essential information, like a trailer in a film, which entice people to read. It is also useful because we can directly understand if this paper could be useful for our research.*
- *It shows that even though a drug is in a clinical trial, the outcome can be different one expected. This means, one was to go back to the drawing board and find out, how to improve it.*

# 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**

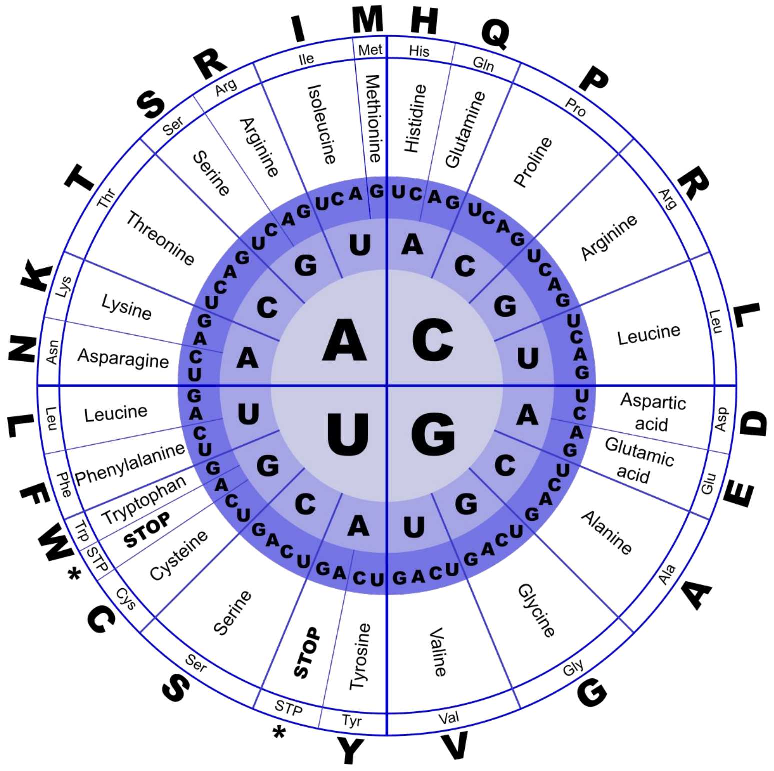
*The Lancet Oncology* 15, Nr. 3 (1. März 2014): 323–32. https://doi.org/10.1016/S1470-2045(14)70012-9.

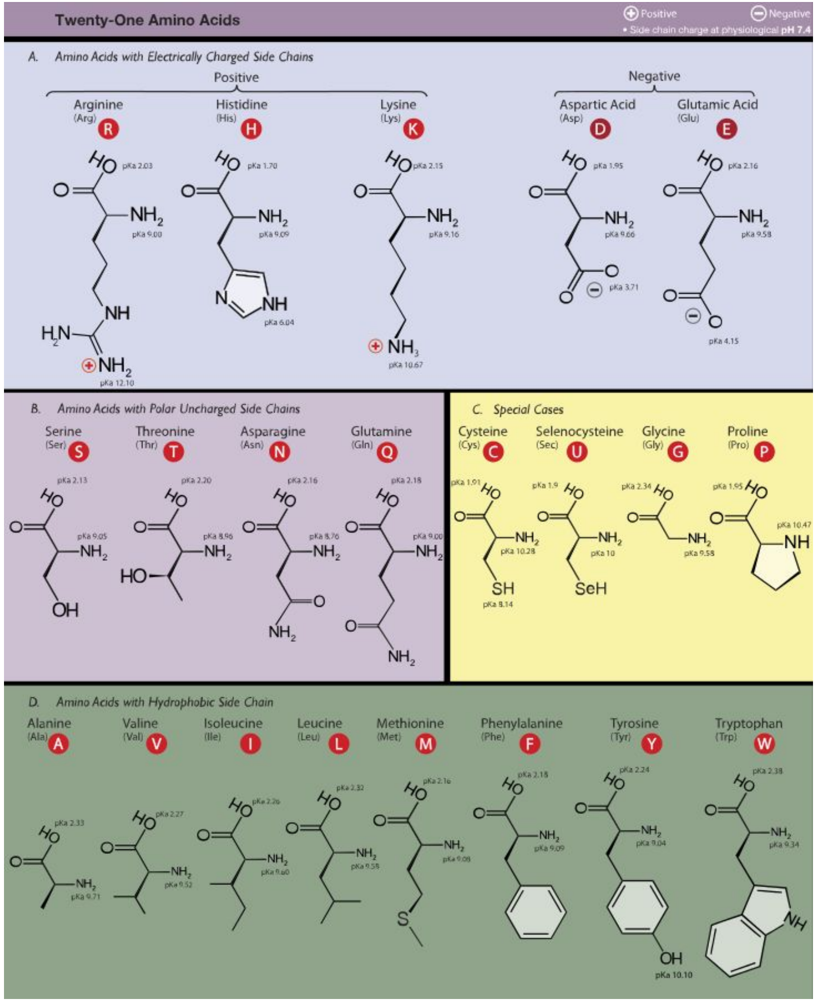


| Number at risk | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Vemurafenib | 337 | 326 | 280 | 231 | 178 | 109 | 44 | 7 | 1 |
| Dacarbazine | 338 | 244 | 173 | 111 | 79 | 50 | 24 | 4 | 0 |

*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



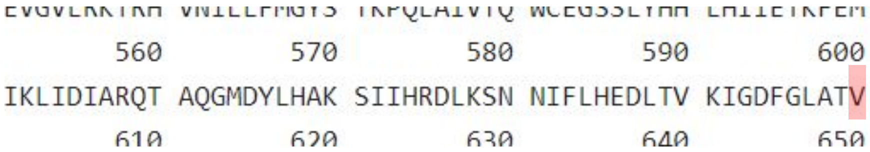


Left: Genetic codes (OpenChart, public domain); Right: Dancojocari, CC BY-SA 3.0, via Wikimedia Commons

# Vemurafenib (Zelboraf, PLX4032)
# V600E mutated BRAF inhibition

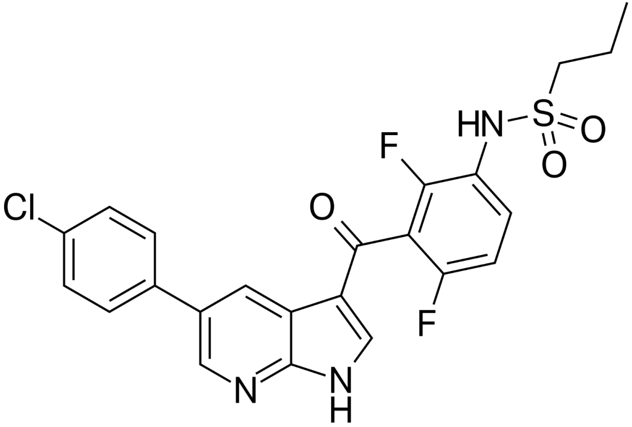
- V600E: Valine (V) on the amino-acid position 600 is substituted by glutamic acid (E).
- **Question: what mutations in DNA would cause the change in the amino-acid sequence?**

- 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.



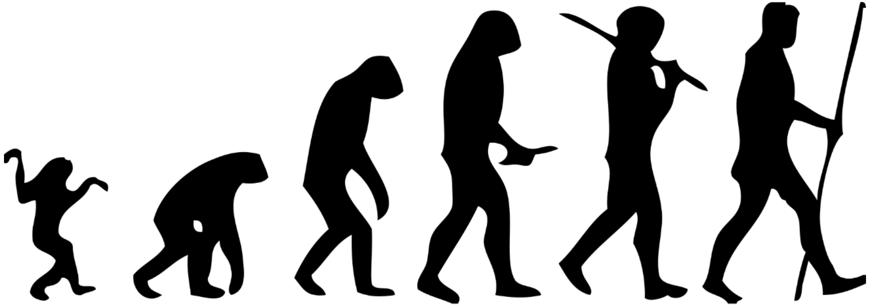
```
EVGVLRKTRH VNILLFMGYS TKPQLAIVTQ WCEGSSLYHH LHIIETKFEM
       560        570        580        590        600
IKLIDIARQT AQGMDYLHAK SIIHRDLKSN NIFLHEDLTV KIGDFGLATV
       610        620        630        640        650
```

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

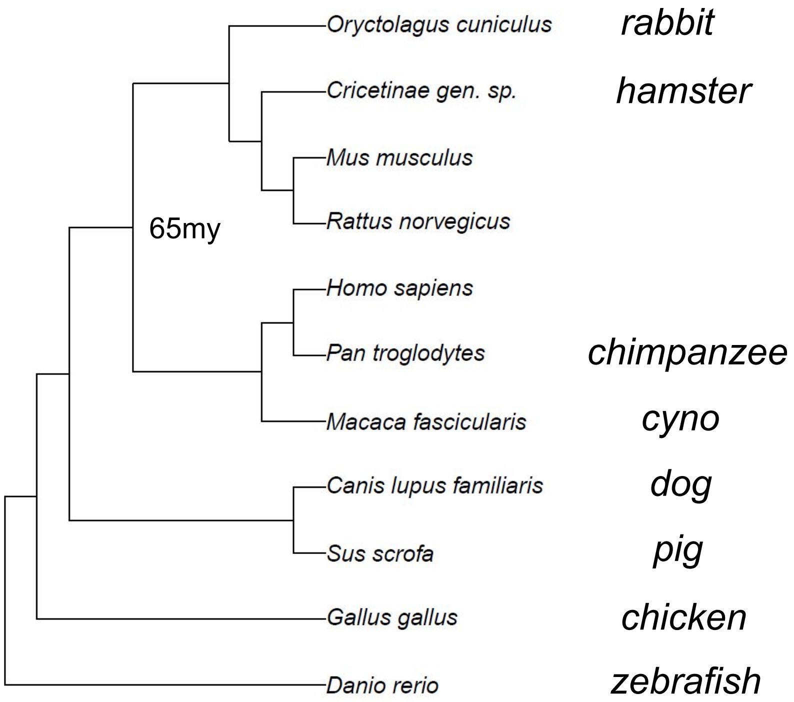


Source: https://commons.wikimedia.org/wiki/File:Vemurafenib_structure.svg

# Evolution: what is wrong with this figure?

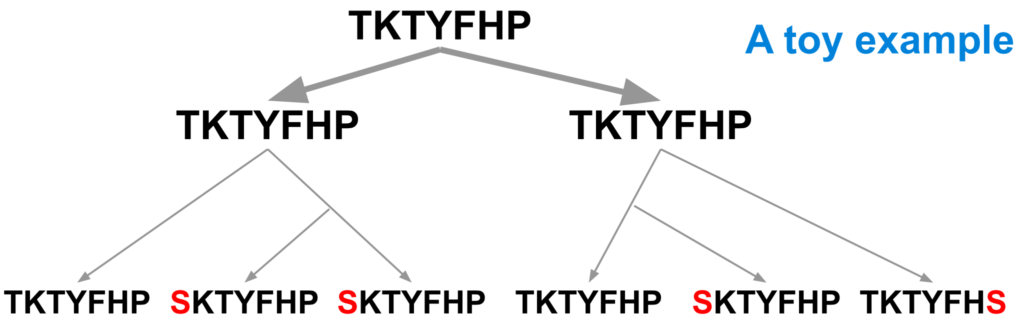


# Phylogeny of commonly used species for animal studies



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.

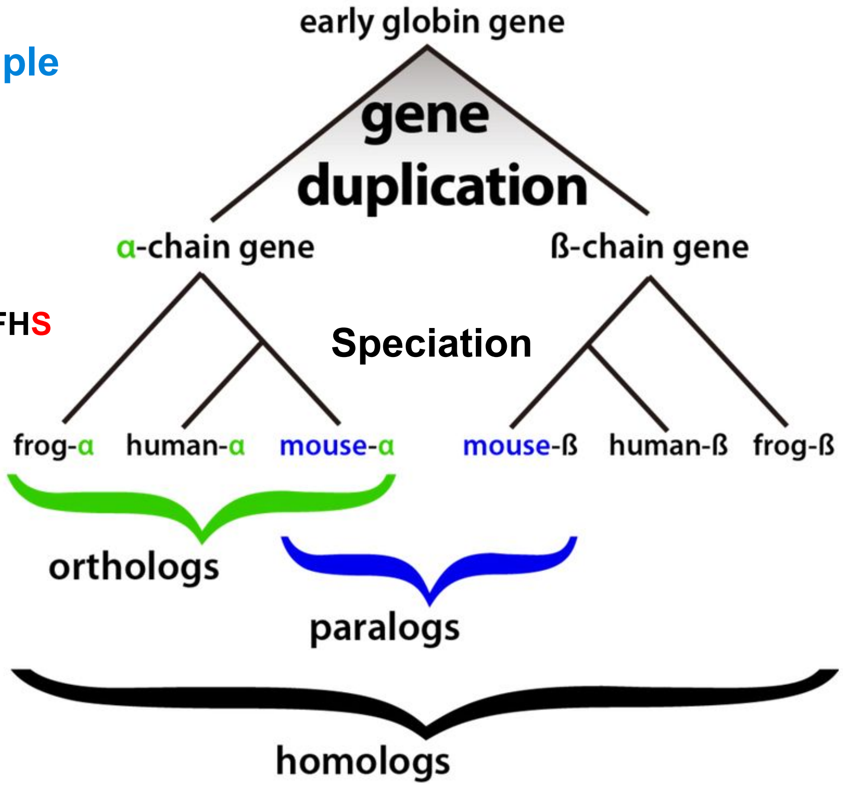
# Mutations in biological sequences make species in millions of years



- **Homologs**: two genes related by descent from a common ancestral gene.

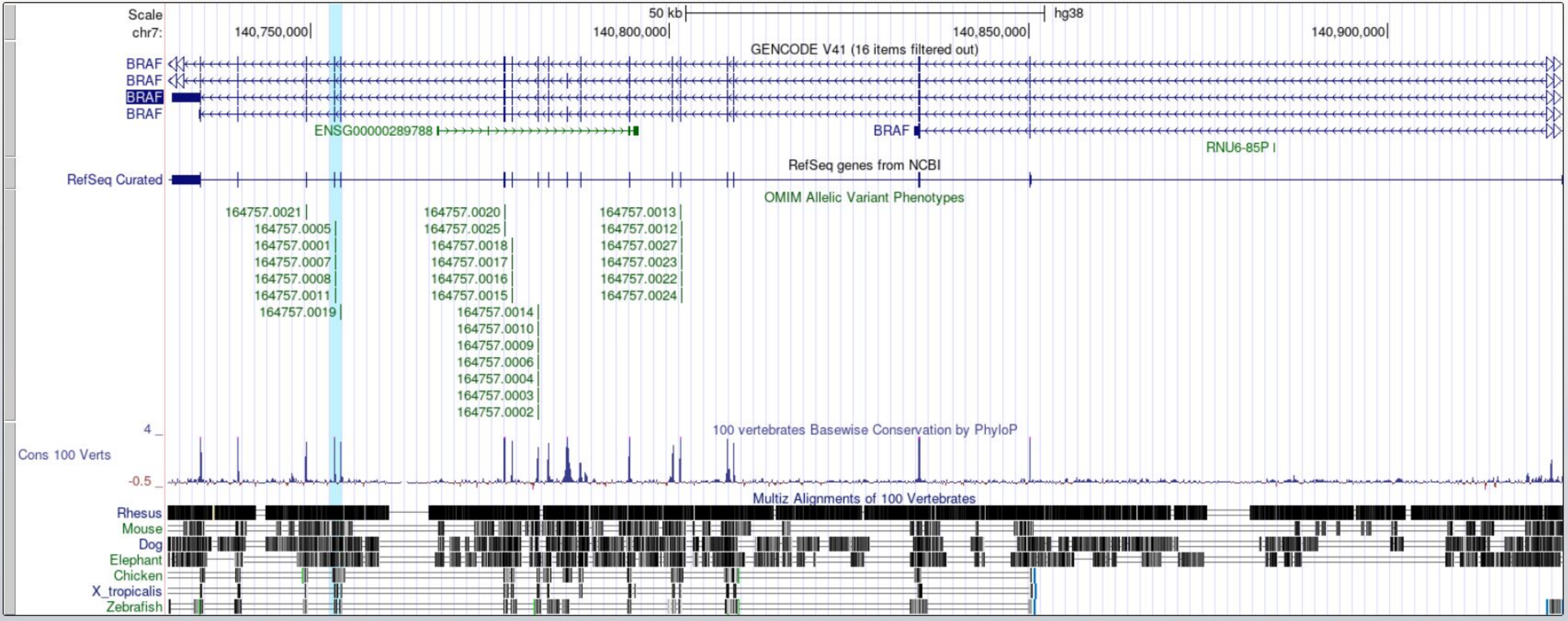
- **Orthologs**: homologous genes in different species, which are evolved from a single ancestral gene by **speciation.**

- **Paralogs**: Two genes **in a genome** that are related by **duplication**.



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

# The BRAF gene on Chromosome 7 in UCSC Genome Viewer



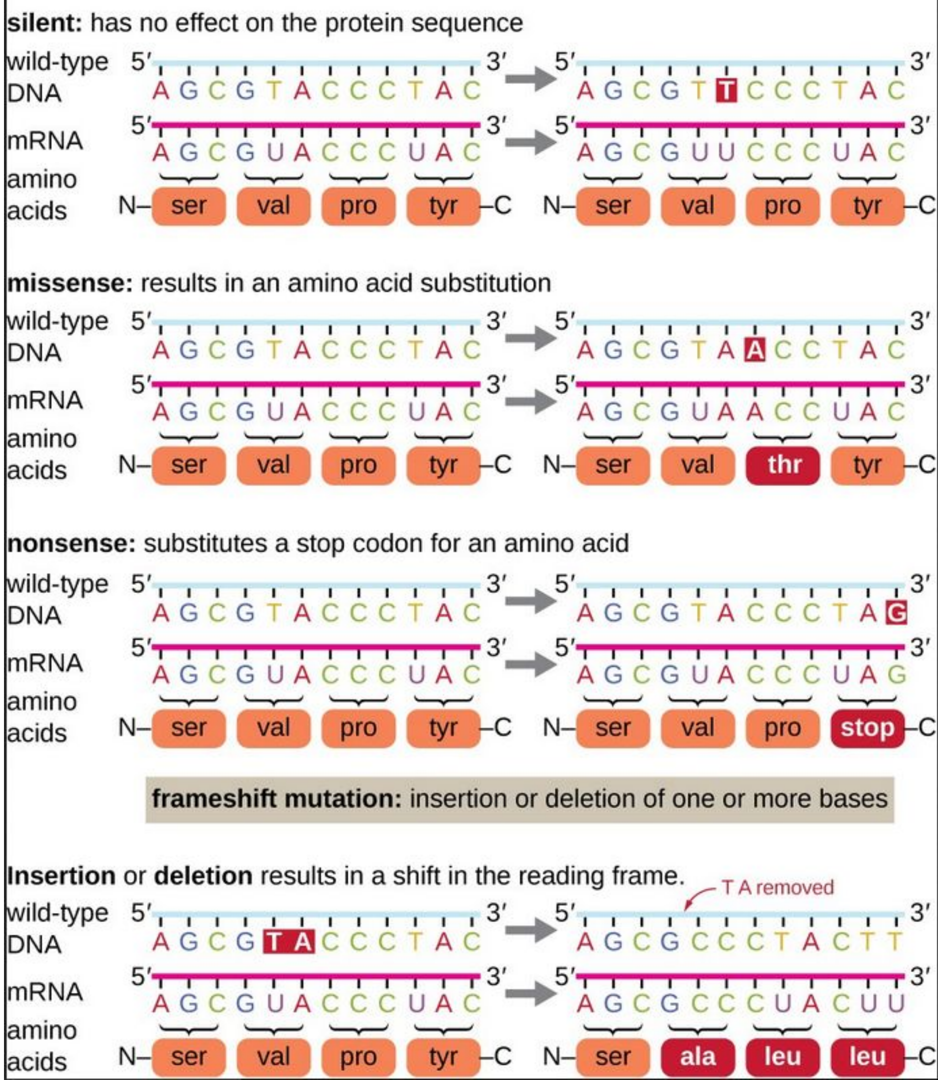
# Loss-of-function (LoF) mutations

***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.



silent: has no effect on the protein sequence
wild-type DNA 5′ A G C G T A C C C T A C 3′ → 5′ A G C G T T C C C T A C 3′
mRNA 5′ A G C G U A C C C U A C 3′ → 5′ A G C G U U C C C U A C 3′
amino acids N– ser val pro tyr –C → N– ser val pro tyr –C
missense: results in an amino acid substitution
wild-type DNA 5′ A G C G T A C C C T A C 3′ → 5′ A G C G T A A C C T A C 3′
mRNA 5′ A G C G U A C C C U A C 3′ → 5′ A G C G U A A C C U A C 3′
amino acids N– ser val pro tyr –C → N– ser val thr tyr –C
nonsense: substitutes a stop codon for an amino acid
wild-type DNA 5′ A G C G T A C C C T A C 3′ → 5′ A G C G T A C C C T A G 3′
mRNA 5′ A G C G U A C C C U A C 3′ → 5′ A G C G U A C C C U A G 3′
amino acids N– ser val pro tyr –C → N– ser val pro stop –C
frameshift mutation: insertion or deletion of one or more bases
Insertion or deletion results in a shift in the reading frame.
T A removed
wild-type DNA 5′ A G C G T A C C C T A C 3′ → 5′ A G C G C C C T A C T T 3′
mRNA 5′ A G C G U A C C C U A C 3′ → 5′ A G C G C C C U A C U U 3′
amino acids N– ser val pro tyr –C → N– ser ala leu leu –C

# 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 strings

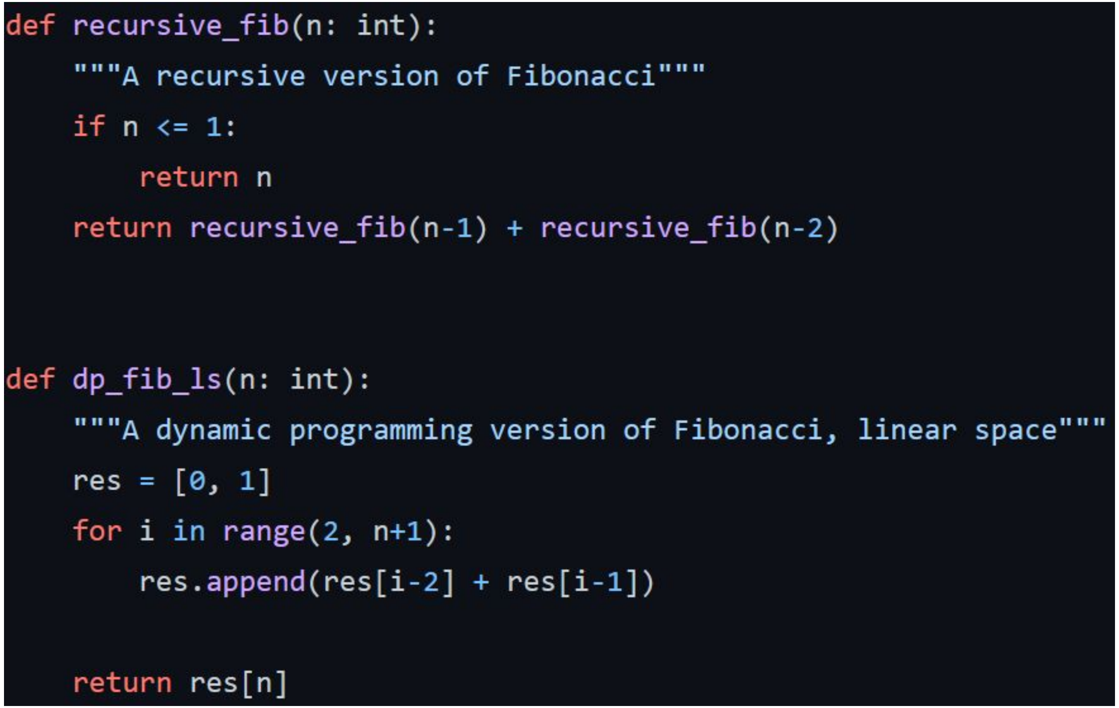
**Levenshtein distance:**
The minimum number of operations required to transform string *a* to string *b* with following operations:

- Insertion (**bat** → **bait**)
- Deletion (**boat** → **bot**)
- Substitution (**pig** → **big**)

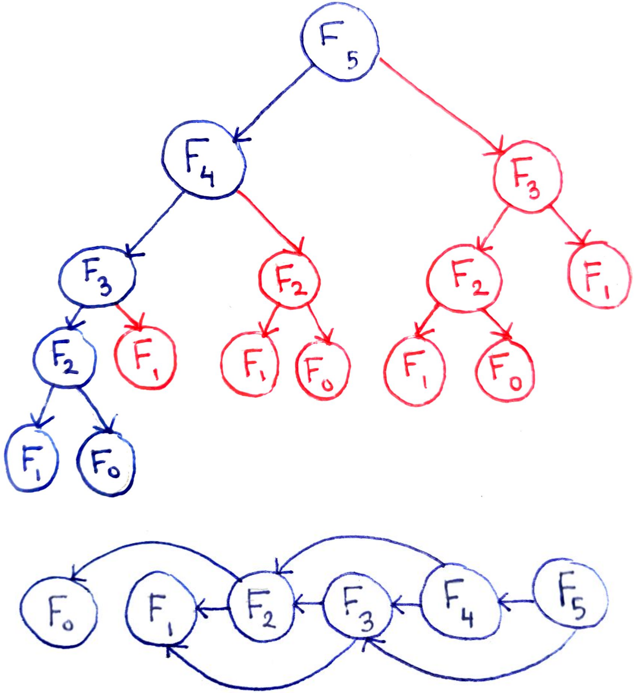
**Questions:**

1. What is the Levenshtein distance between `ATGC` and `AGC`?
2. What is the Levenshtein distance between `ACTGCTT` and `ACATT`?
3. How would you calculate the Levenshtein distance between `ATTAATGCC` and `ATATTTCGCC`?

# Dynamic programming turns a exponential problem into a linear one



```python
def recursive_fib(n: int):
    """A recursive version of Fibonacci"""
    if n <= 1:
        return n
    return recursive_fib(n-1) + recursive_fib(n-2)


def dp_fib_ls(n: int):
    """A dynamic programming version of Fibonacci, linear space"""
    res = [0, 1]
    for i in range(2, n+1):
        res.append(res[i-2] + res[i-1])

    return res[n]
```

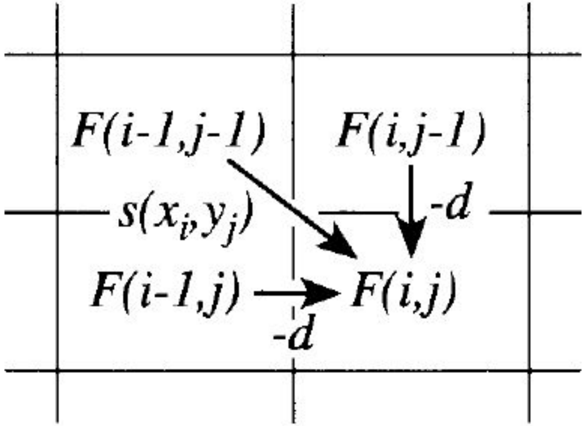


For fib(25), I observed a difference in the order of $10^4$ in run time. Check out the source code for the constant-space version.

Directed Acyclic Graphs (DAGs) by Avik Das

# Dynamic programming for biological sequence analysis

- The calculation of Levenshtein distance is *repeated* to fill a matrix of distances between two sequences.
- At each step, we keep a pointer in each cell back to the cell from which its value is derived.
- The value in the final (bottom right) cell of the matrix is the best score of an alignment.
- The best alignment is done by *traceback*, i.e. building the alignment in reverse, starting from the final cell, and following the pointers of the optimal path.

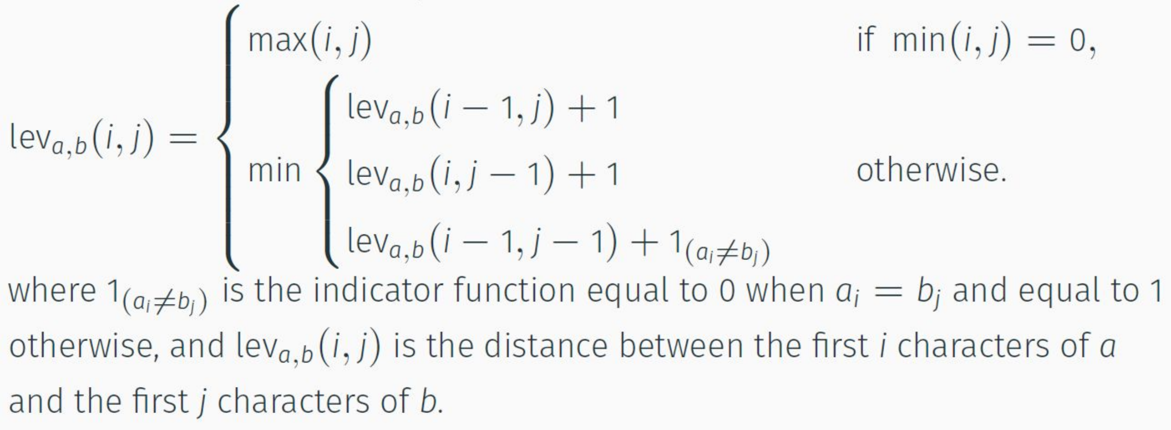


Chapter 2, Durbin, Richard, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge: Cambridge University Press, 1998. https://doi.org/10.1017/CBO9780511790492.

# Calculate the Levenshtein distance with dynamic programming

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

| | | A | T | G | C |
|---|---|---|---|---|---|
| | 0 | | | | |
| A | | | | | |
| G | | | | | |
| C | | | | | |



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

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 $\operatorname{lev}_{a,b}(i,j)$ 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).

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

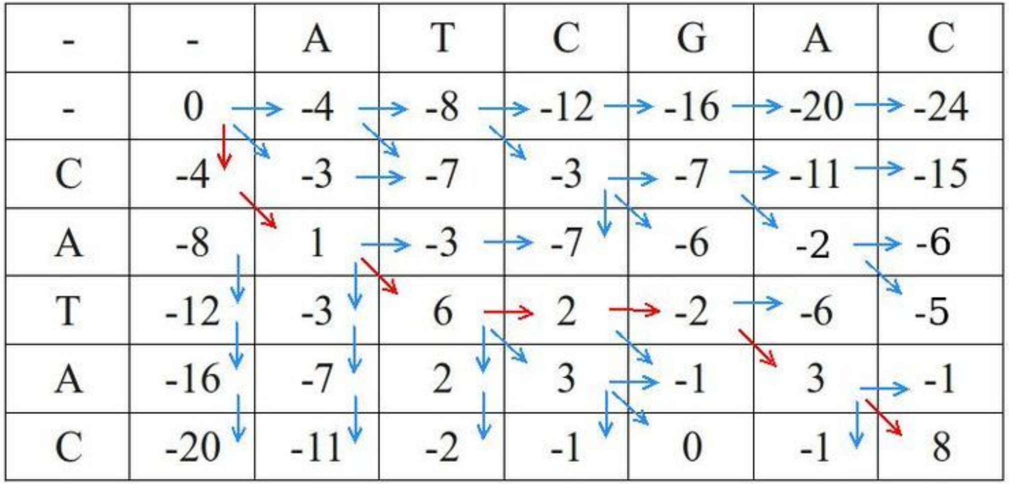
Compared with the Levenshtein distance, the Needleman-Wunsch algorithm uses biologically meaningful parameters to score insertion or deletion (gap penalty *d*), and substitution or mutation events (a substitution matrix *M*). The dynamic programming technique is used in a similar way.

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

**Parameters**: d=-4, $M = \left(\begin{array}{c|cccc} & A & C & T & G \\ A & 5 & -3 & -3 & -3 \\ C & -3 & 5 & -3 & -3 \\ T & -3 & -3 & 5 & -3 \\ G & -3 & -3 & -3 & 5 \end{array}\right)$

**Solution**:

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

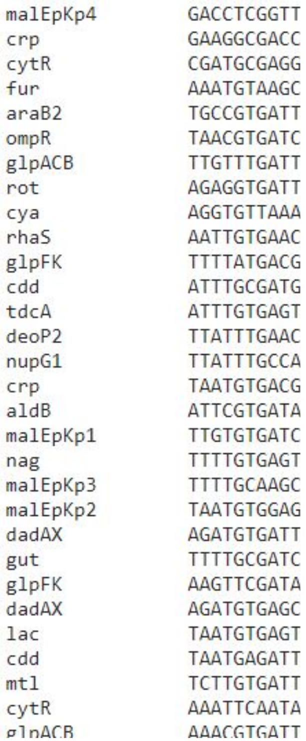


| - | - | A | T | C | G | A | C |
|---|---|---|---|---|---|---|---|
| - | 0 | -4 | -8 | -12 | -16 | -20 | -24 |
| C | -4 | -3 | -7 | -3 | -7 | -11 | -15 |
| A | -8 | 1 | -3 | -7 | -6 | -2 | -6 |
| T | -12 | -3 | 6 | 2 | -2 | -6 | -5 |
| A | -16 | -7 | 2 | 3 | -1 | 3 | -1 |
| C | -20 | -11 | -2 | -1 |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

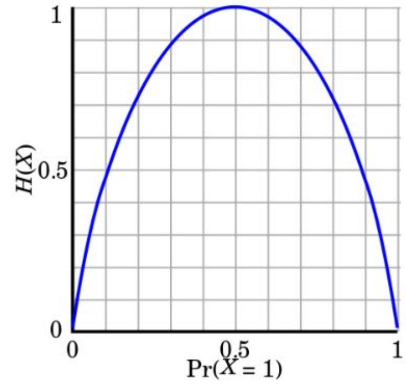
# *WebLogo*: a transition from the deterministic view to a probabilistic view



| | |
|---|---|
| malEpKp4 | GACCTCGGTT |
| crp | GAAGGCGACC |
| cytR | CGATGCGAGG |
| fur | AAATGTAAGC |
| araB2 | TGCCGTGATT |
| ompR | TAACGTGATC |
| glpACB | TTGTTTGATT |
| rot | AGAGGTGATT |
| cya | AGGTGTTAAA |
| rhaS | AATTGTGAAC |
| glpFK | TTTTATGACG |
| cdd | ATTTGCGATG |
| tdcA | ATTTGTGAGT |
| deoP2 | TTATTTGAAC |
| nupG1 | TTATTTGCCA |
| crp | TAATGTGACG |
| aldB | ATTCGTGATA |
| malEpKp1 | TTGTGTGATC |
| nag | TTTTGTGAGT |
| malEpKp3 | TTTTGCAAGC |
| malEpKp2 | TAATGTGGAG |
| dadAX | AGATGTGATT |
| gut | TTTTGCGATC |
| glpFK | AAGTTCGATA |
| dadAX | AGATGTGAGC |
| lac | TAATGTGAGT |
| cdd | TAATGAGATT |
| mtl | TCTTGTGATT |
| cytR | AAATTCAATA |
| glpACB | AAACGTGATT |

**Aligned sequences**

Information entropy of a Bernoulli trial

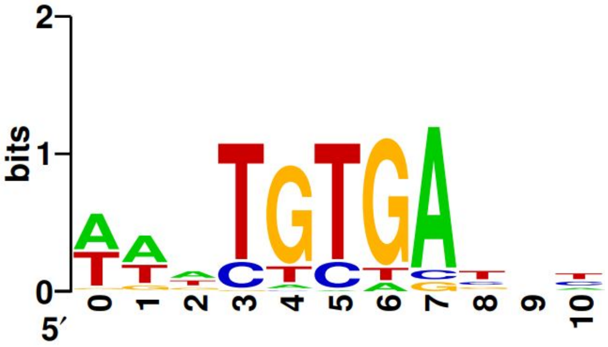


Brona, Wikimedia, shared with CC BY-SA 3.0

$$R_{seq} = S_{max} - S_{obs} = \log_2 N - \left( - \sum_{n=1}^{N} p_n \log_2 p_n \right)$$

**Conservation per site defined as difference between maximal and observed information**

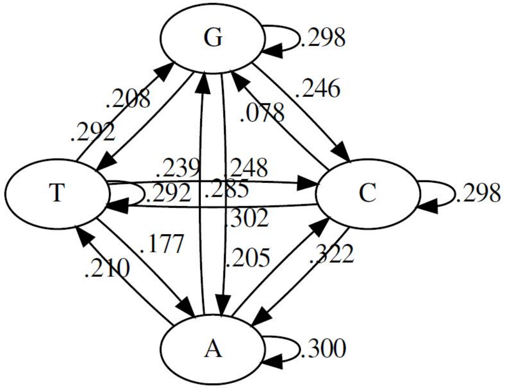
1. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18, 6097–6100 (1990).
2. 1. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: A Sequence Logo Generator. Genome Res. 14, 1188–1190 (2004).



**WebLogo**

# 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**.



## Applications

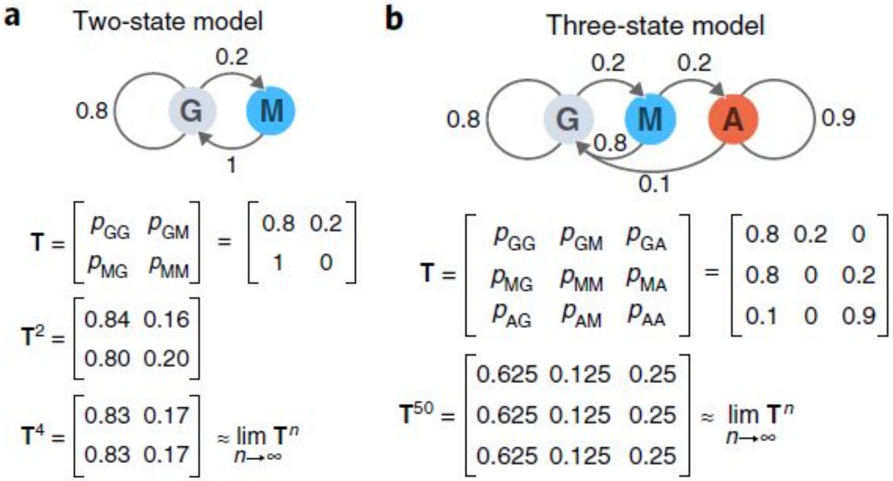
- Given a string, assuming that the Markov chain model is suitable, we can construct a Markov chain, for instance by counting transitions and normalize the count matrix.
- Given a Markov chain model and a string, we can calculate the probability that the string is generated by the specific model with 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.
  - Note that is has the form of the left eigenvector equation, $uA=\kappa u$, where $\kappa$ 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.



$$\mathbf{T} = \begin{bmatrix} p_{GG} & p_{GM} \\ p_{MG} & p_{MM} \end{bmatrix} = \begin{bmatrix} 0.8 & 0.2 \\ 1 & 0 \end{bmatrix}$$

$$\mathbf{T}^2 = \begin{bmatrix} 0.84 & 0.16 \\ 0.80 & 0.20 \end{bmatrix}$$

$$\mathbf{T}^4 = \begin{bmatrix} 0.83 & 0.17 \\ 0.83 & 0.17 \end{bmatrix} \approx \lim_{n\to\infty} \mathbf{T}^n$$

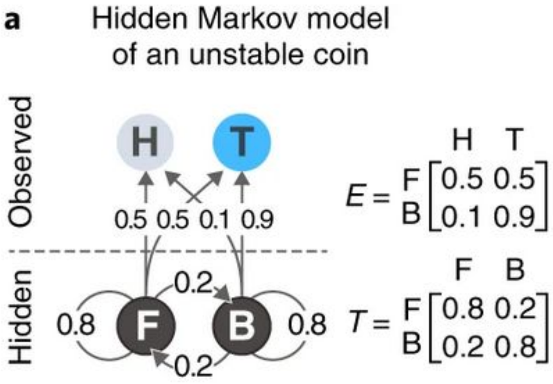
$$\mathbf{T} = \begin{bmatrix} p_{GG} & p_{GM} & p_{GA} \\ p_{MG} & p_{MM} & p_{MA} \\ p_{AG} & p_{AM} & p_{AA} \end{bmatrix} = \begin{bmatrix} 0.8 & 0.2 & 0 \\ 0.8 & 0 & 0.2 \\ 0.1 & 0 & 0.9 \end{bmatrix}$$

$$\mathbf{T}^{50} = \begin{bmatrix} 0.625 & 0.125 & 0.25 \\ 0.625 & 0.125 & 0.25 \\ 0.625 & 0.125 & 0.25 \end{bmatrix} \approx \lim_{n\to\infty} \mathbf{T}^n$$

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.

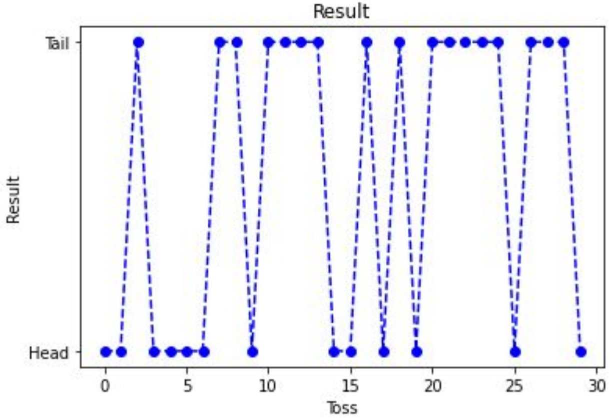
# Hidden Markov Models model hidden states based on observations



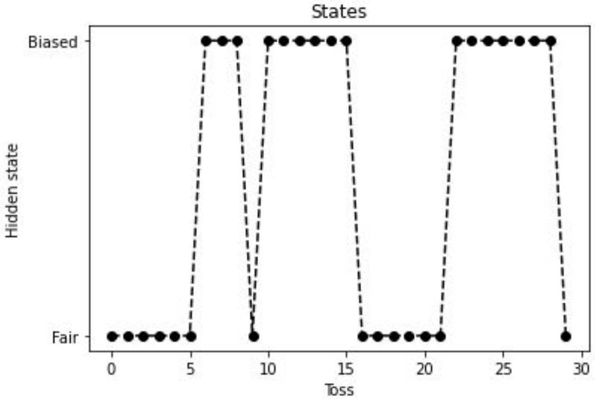
$$E = \begin{array}{c|cc} & H & T \\ \hline F & 0.5 & 0.5 \\ B & 0.1 & 0.9 \end{array}$$

$$T = \begin{array}{c|cc} & F & B \\ \hline F & 0.8 & 0.2 \\ B & 0.2 & 0.8 \end{array}$$

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.



[0 0 1 0 0 0 0 1 1 0
1 1 1 1 0 0 1 0 1 0
1 1 1 1 1 0 1 1 1 0]
0=Head, T=Tail



[0 0 0 0 0 0 1 1 1 0 1 1
1 1 1 1 0 0 0 0 0 0 1 1 1
1 1 1 1 0]
0=Fair, 1=Biased

# The Viterbi algorithm estimates transmission and emission matrices

A Hidden Markov Model consists of two graphs (matrices): one of **hidden states**, which corresponds to the **transmission matrix**, and one of **observed states**, which corresponds to the **emission matrix.**

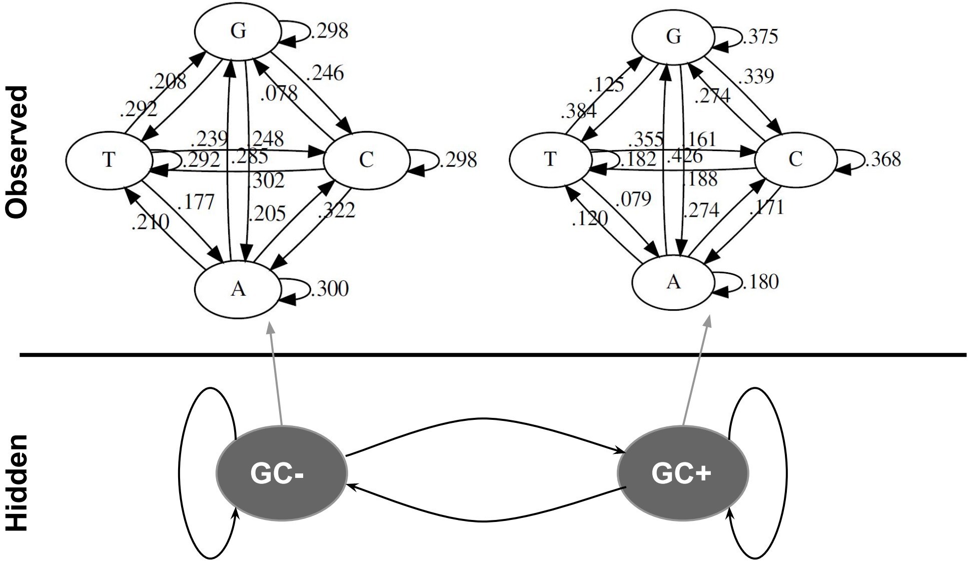
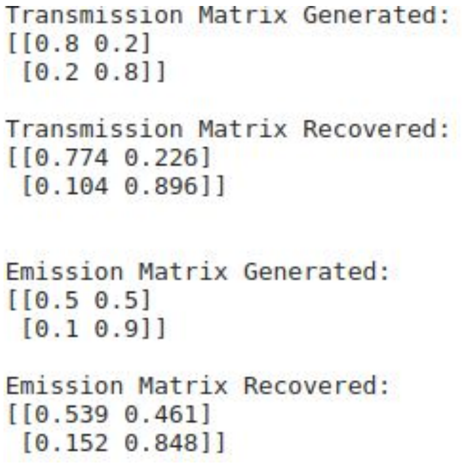


Illustration of a Hidden Markov Model predicting CpG islands in genomic sequences

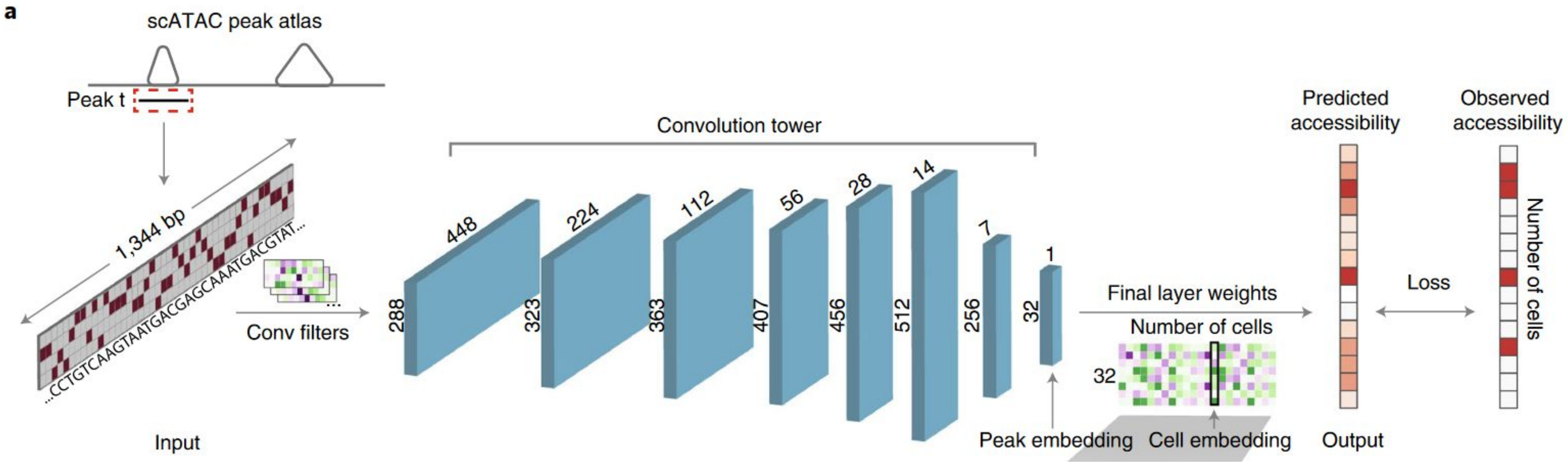
The **Viterbi algorithm** (based on dynamic programming), or the **Baum-Welch algorithm** (a special case of EM algorithms) is used to estimate its parameters.



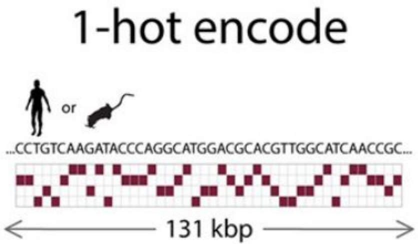
```
Transmission Matrix Generated:
[[0.8 0.2]
 [0.2 0.8]]

Transmission Matrix Recovered:
[[0.774 0.226]
 [0.104 0.896]]


Emission Matrix Generated:
[[0.5 0.5]
 [0.1 0.9]]

Emission Matrix Recovered:
[[0.539 0.461]
 [0.152 0.848]]
```

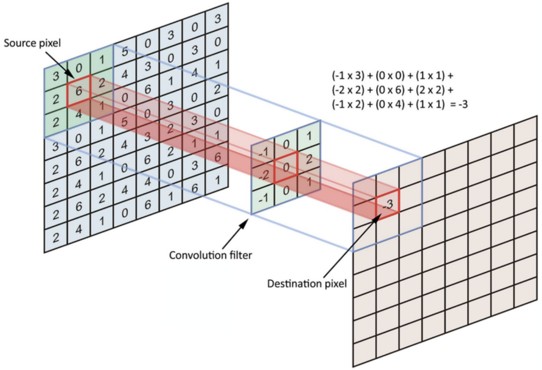
The transmission and emission matrices estimated by the Viterbi algorithm from 1000 observations generated by the HMM model in the last slide. Source code

# Sequence analysis with convolutional neural networks



1. Kelley, D. R. & Reshef, Y. A. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. bioRxiv 161851 (2017) doi:10.1101/161851.
2. Kelley, D. R. et al. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. 28, 739–750 (2018).
3. Yuan, H. & Kelley, D. R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. Nat Methods 19, 1088–1096 (2022).



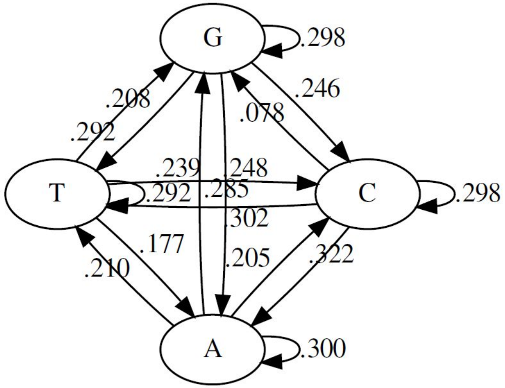

Source pixel
(-1 x 3) + (0 x 0) + (1 x 1) +
(-2 x 2) + (0 x 6) + (2 x 2) +
(-1 x 2) + (0 x 4) + (1 x 1) = -3
Convolution filter
Destination pixel
-3

# 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.
- Mathematical techniques such as dynamic programming, when implemented as algorithms software tools, are important for many tasks.
- We explored both deterministic views of sequence analysis, with the example of Levenshtein distance, and probabilistic views, with the example of Markov chains, hidden Markov chains, and convolutional neural networks.

# Offline exercises

1. Check out the examples of dynamic programming and constant-space solution of the Fibonacci problem, and make sure that you can implement them yourself.
2. Calculate the Levenshtein distance between `ATTAATGCC` and `ATATTTCGCC`.
3. Compare *p(*`ACGTGGT`|M*)* and *p(*`ACCTGGT`|M*)*, where M stands for the model on the right side.
4. We have got a RNA sequence by sequencing sputum from a patient (see below). How can we know the original genome of the sequence, and ideally the gene encoding the sequences? Tips: 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. Default parameters are okay.



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

```
ATGTTTGTTTTTCTTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACAACCAGAACTCAATTACCC
CCTGCATACACTAATTCTTTCACACGTGGTGTTTATTACCCTGACAAAGTTTTCAGATCCTCAGT
```

# Backup slides

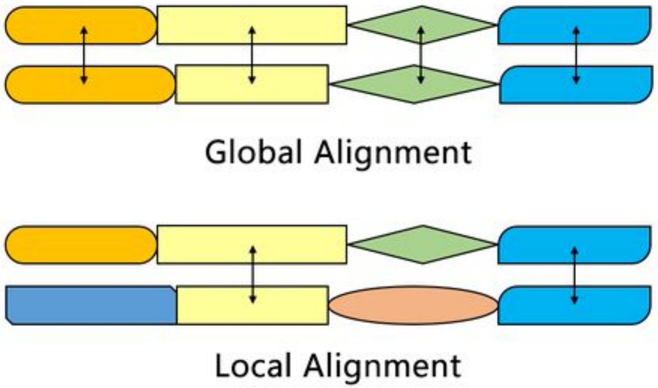
# Richard Bell on the origin of the name *Dynamic Programming*

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.

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

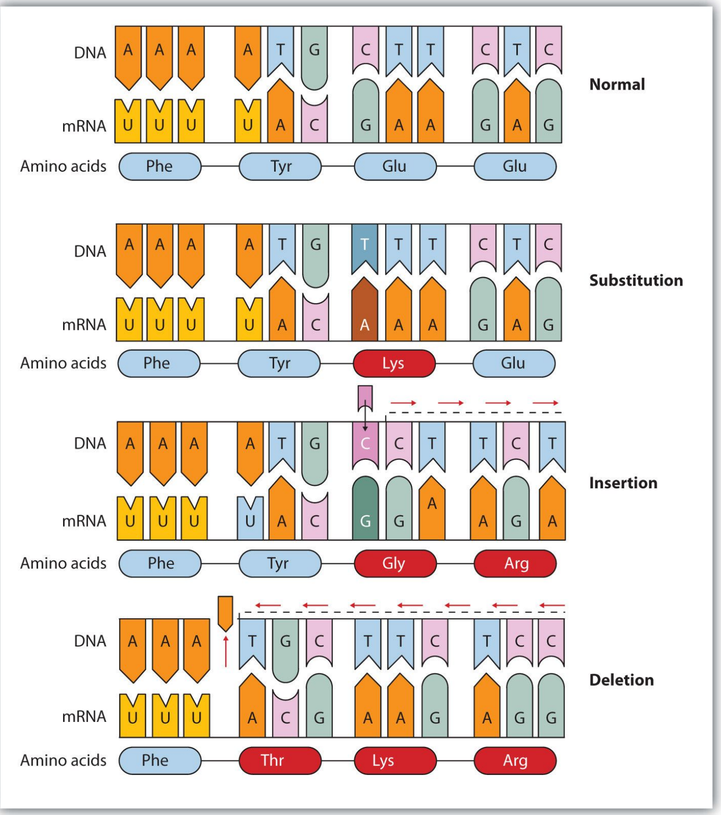


Yz CS5160, CC-BY SA 4.0

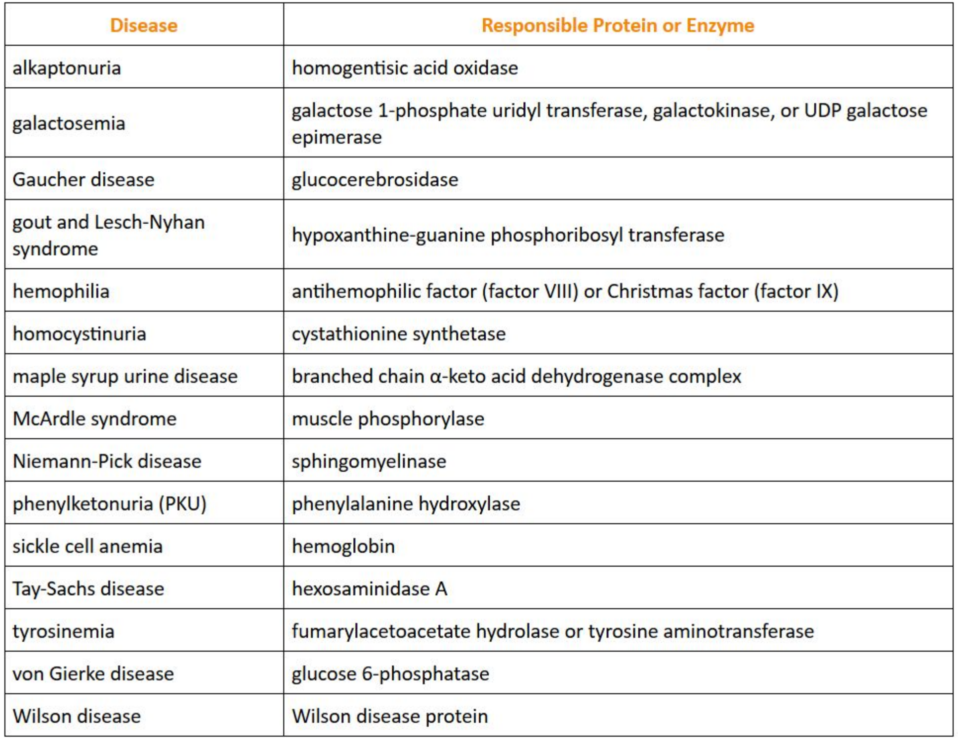
| | Needleman-Wunsch | Smith-Waterman |
|---|---|---|
| **Initialization** | Gap penalty in first column and first row | 0 in first column and first row |
| **Scoring** | Scores can be negative | Negative scores are set to 0 |
| **Traceback** | Begin at the bottom right, and end at the top left cell.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

# Mutations in biological sequences cause diseases in a lifetime



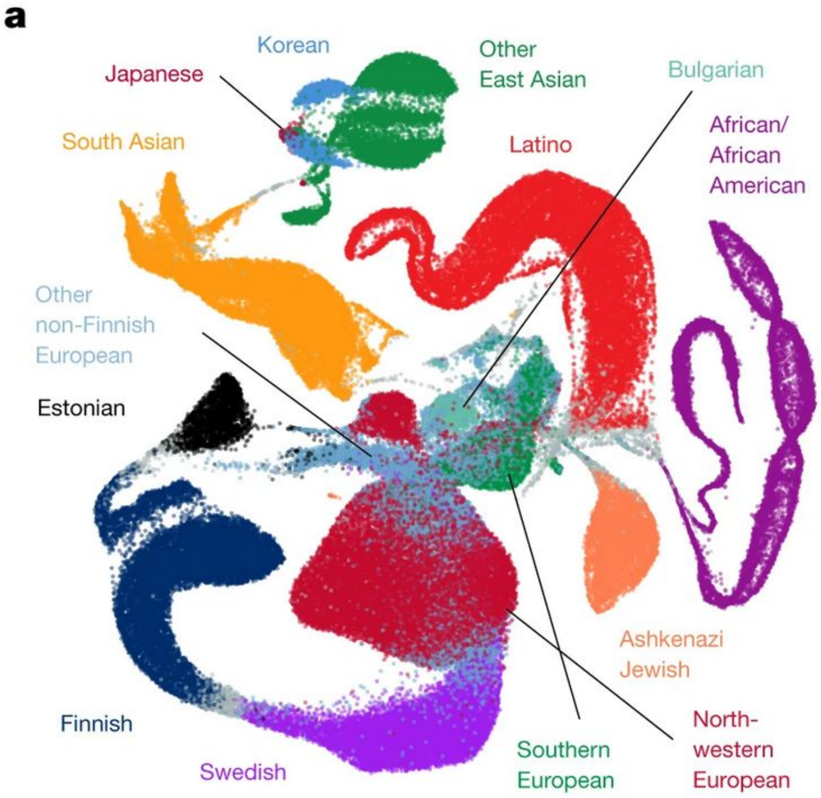
Normal
DNA A A A A T G C T T C T C
mRNA U U U U A C G A A G A G
Amino acids Phe Tyr Glu Glu
Substitution
DNA A A A A T G T T T C T C
mRNA U U U U A C A A A G A G
Amino acids Phe Tyr Lys Glu
Insertion
DNA A A A A T G C C T T C T
mRNA U U U U A C G G A A G A
Amino acids Phe Tyr Gly Arg
Deletion
DNA A A A T G C T T C T C C
mRNA U U U A C G A A G A G G
Amino acids Phe Thr Lys Arg

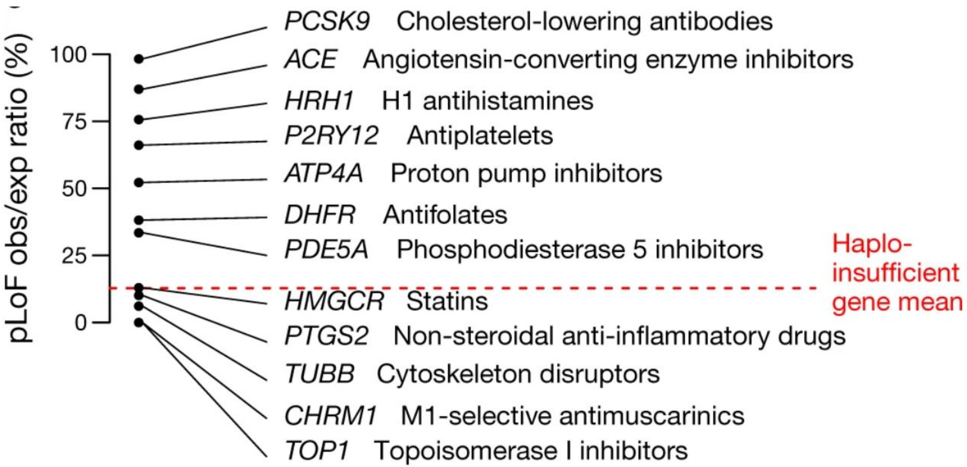


| Disease | Responsible Protein or Enzyme |
|---|---|
| alkaptonuria | homogentisic acid oxidase |
| galactosemia | galactose 1-phosphate uridyl transferase, galactokinase, or UDP galactose epimerase |
| Gaucher disease | glucocerebrosidase |
| gout and Lesch-Nyhan syndrome | hypoxanthine-guanine phosphoribosyl 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

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





Left: Karczewski, et al. “The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans.” Nature 581, no. 7809 (May 2020): 434–43..

Right: Minikel, et al. “Evaluating Drug Targets through Human Loss-of-Function Genetic Variation.” Nature 581, no. 7809 (May 2020): 459–64..

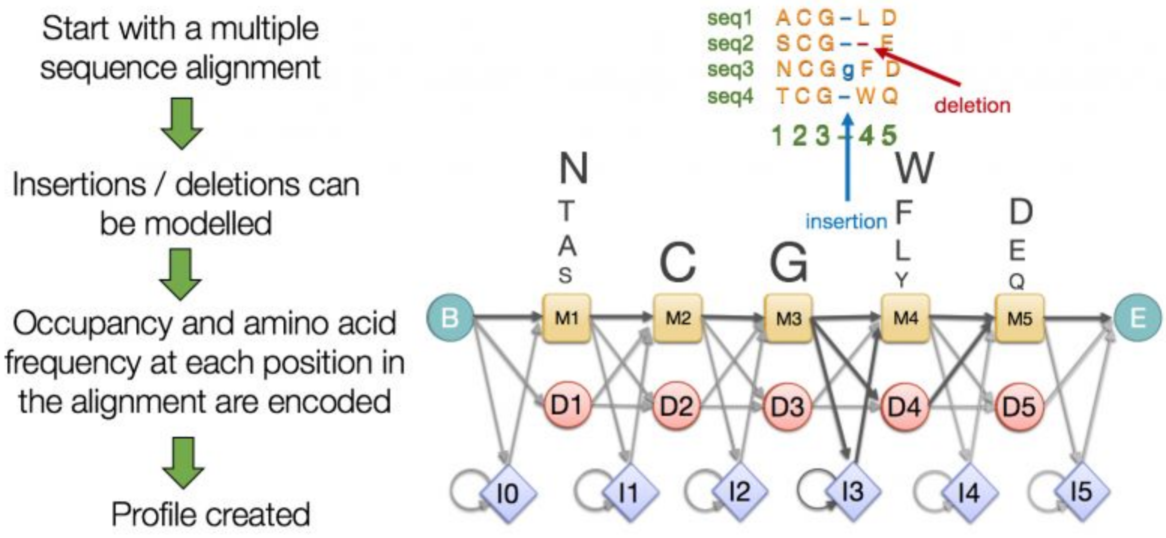
# Profile Hidden Markov Models capture evolutionary changes in homologs

**M:** match states. In the match state, the probability distribution is the frequency of the amino acids in that position.

**I**: insert states, which model highly variable regions in the alignment

**D**: delete states, which allows gaps and deletion.

Profile HMMs belongs to ***generative models***.



Start with a multiple sequence alignment

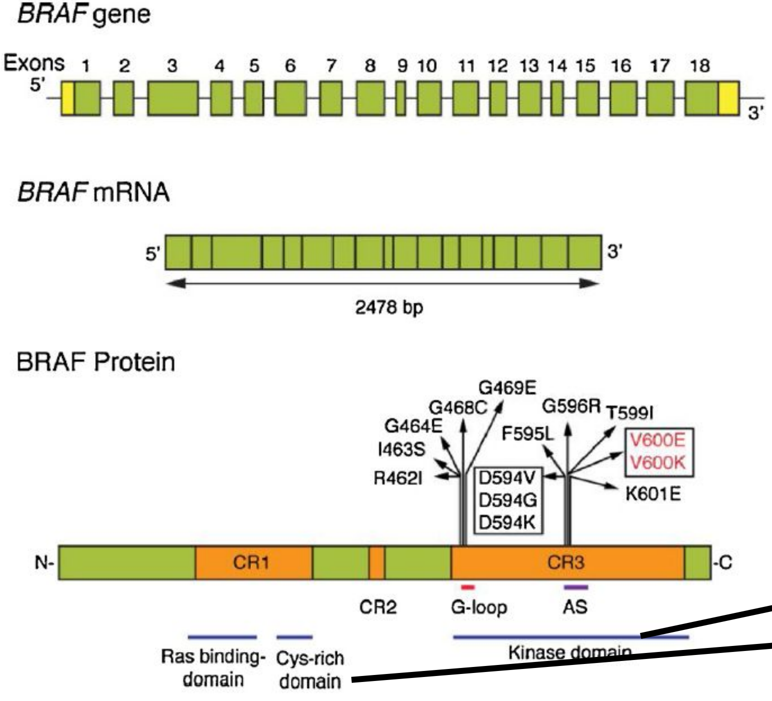
Insertions / deletions can be modelled

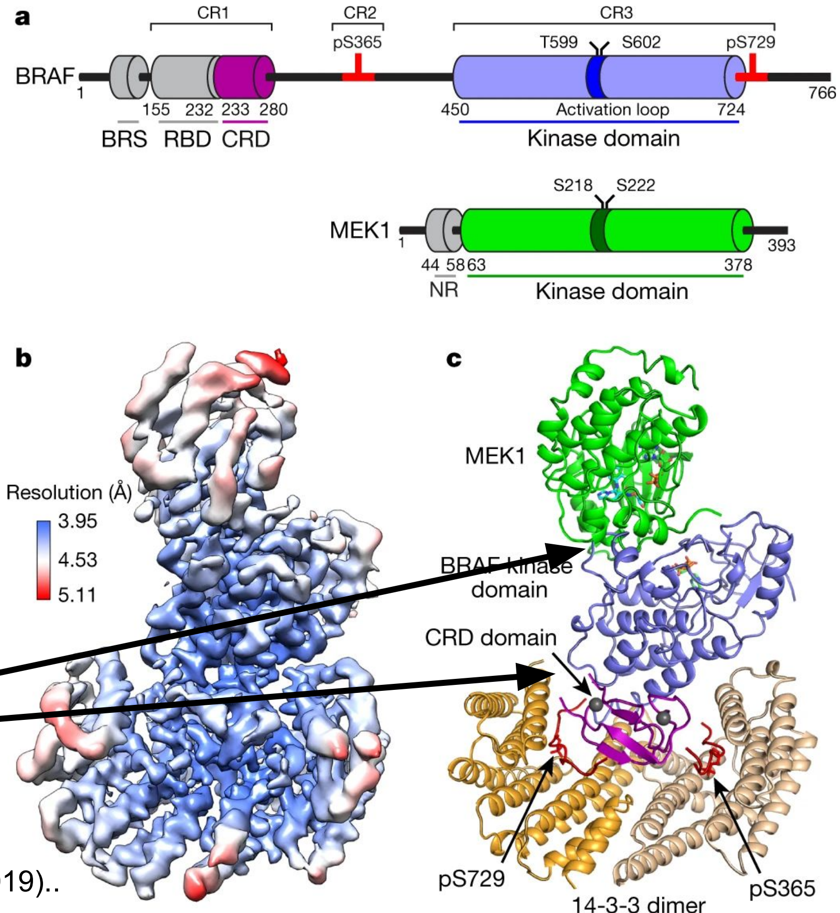
Occupancy and amino acid frequency at each position in the alignment are encoded

Profile created

Figure from Pfam

# Protein domains: self-stabilizing and folding independently from the rest





Left: Frisone, *et al.*, A BRAF New World, Critical Reviews in Oncology/Hematology (2020); Right: Park, *et al.*, Architecture of Autoinhibited and Active BRAF–MEK1–14-3-3 Complexes, Nature (2019)..

# 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?

```
ATGTTTGTTTTTCTTGTTTTATTGCCACTAGT
CTCTAGTCAGTGTGTTAATCTTACAACCAGAA
CTCAATTACCCCCTGCATACACTAATTCTTTC
ACACGTGGTGTTTATTACCCTGACAAAGTTTT
CAGATCCTCAGT
```

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

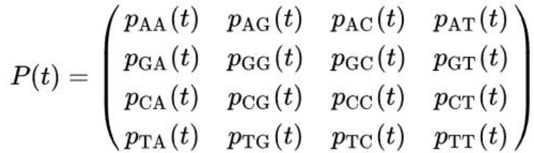
# More software tools for sequence analysis

- **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/

# Continuous-time Markov Chains

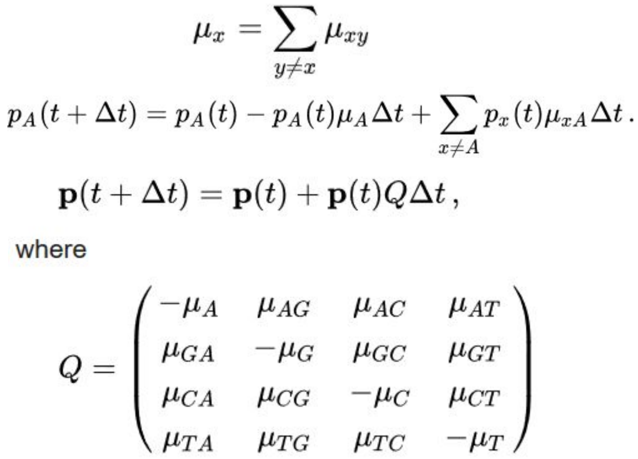
- Continuous-time Markov Chains are used for **phylogenetic analysis**, for instance of orthologous genes and of bacterial/viral genomes. They satisfy the Markovian property: P(t+τ)=P(t)P(τ).
- The process makes a transition from the current state *i* after an amount of time modelled by an *exponential 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 $E_{i,j}$~Exp($q_{i,j}$) (i ≠ 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



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

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

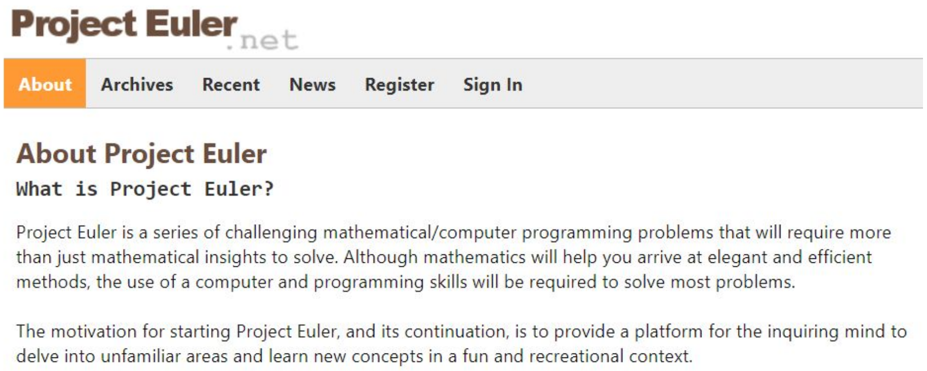
$$\mathbf{p}(t+\Delta t)=\mathbf{p}(t)+\mathbf{p}(t)Q\Delta t\,,$$

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

# The Euler Project



About Project Euler

What is Project Euler?

Project Euler is a series of challenging mathematical/computer programming problems that will require more than just mathematical insights to solve. Although mathematics will help you arrive at elegant and efficient methods, the use of a computer and programming skills will be required to solve most problems.

The motivation for starting Project Euler, and its continuation, is to provide a platform for the inquiring mind to delve into unfamiliar areas and learn new concepts in a fun and recreational context.

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.

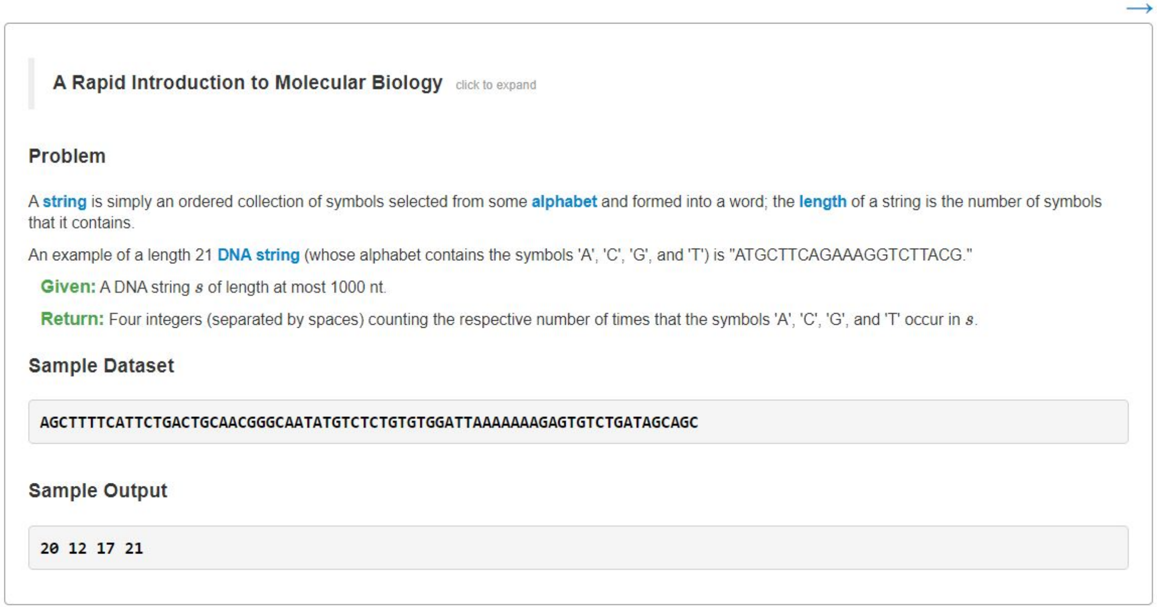
Find the sum of all the multiples of 3 or 5 below 1000.

# 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
```

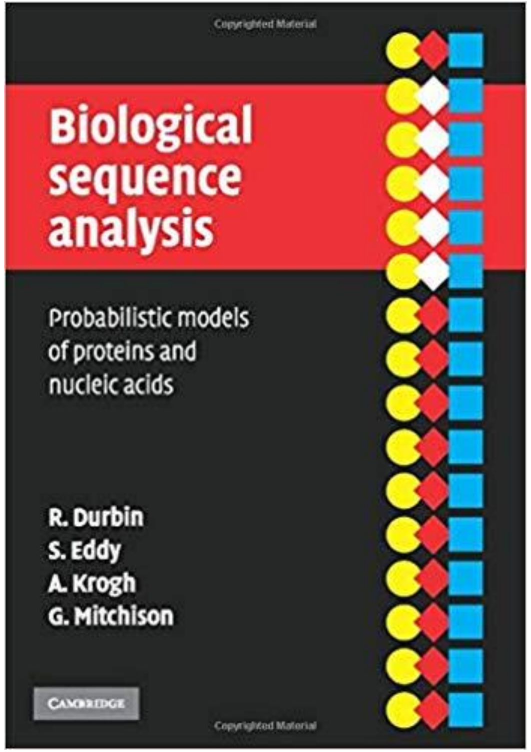
### Sample Output

```
20 12 17 21
```

Please login to solve this problem.

# Further resources

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


Copyrighted Material

Biological sequence analysis

Probabilistic models of proteins and nucleic acids

R. Durbin
S. Eddy
A. Krogh
G. Mitchison

CAMBRIDGE

Copyrighted Material

**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.