# Crash course on Computational Biology for Computer Scientists

Bartek Wilczyński
bartek@mimuw.edu.pl
http://regulomics.mimuw.edu.pl

Phd Open lecture series

17-19 XI 2016

# Topics for the course

- Sequences in Biology – what do we study?
- Sequence comparison and searching – how to quickly find relatives in large sequence banks
- Tree-of-life and its construction(s)
- DNA sequencing – puzzles for experts
- Short sequence mapping – where did this word come from
- Sequence segmentation – finding modules by flipping coins
- Data storage and compression – from DNA to bits and back again
- Structures in Biology – small and smaller

# How to make it efficient

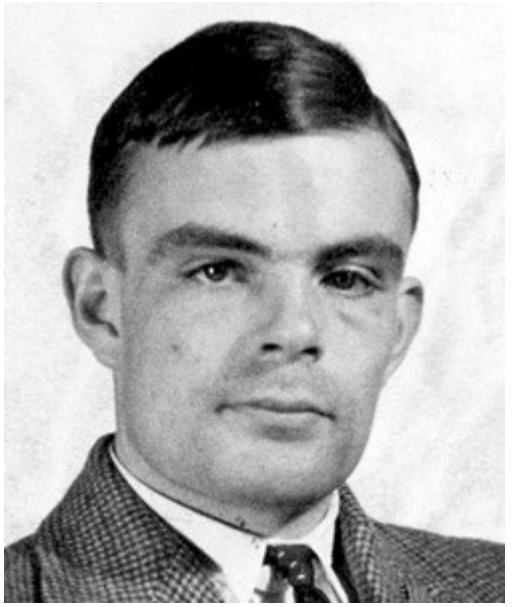
- Diverse audience, I don’t know what you know
- Please **do** interrupt me if you have a question!
- I will not go very deeply into biological details, so if you want more, please ask me later for links to more materials
- I will not go deeply into proofs or derivations, so if you want more, please ask me later for links to more materials
- If you need to ask later: bartek@mimuw.edu.pl

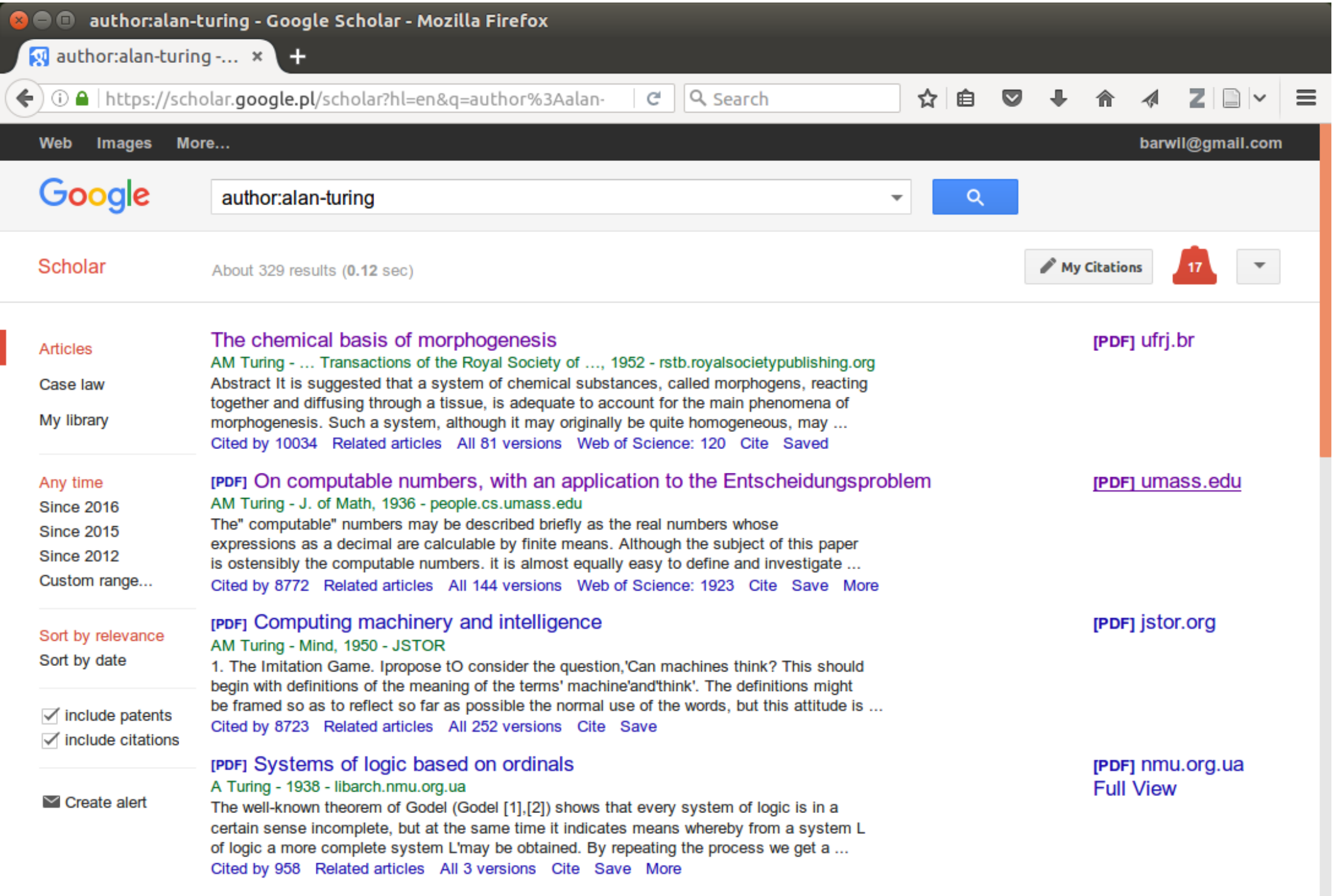
# Homework

- I will post a few (>= 5) questions at the end, depending how far we will get in the lectures
- The nature of them will be diverse: derivation, proofs, computation, data analysis.
- If you want to pass the course and get credit, I'd ask you to solve N-1 questions to get grade N
- You e-mail solutions to me at bartek@mimuw.edu.pl

# Alan Turing (1912 - 1954)

- Very influential mathematician
- Turing machine
- Turing test
- Enigma cracking
- Why is he here?




author:alan-turing - Google Scholar - Mozilla Firefox
author:alan-turing -...
https://scholar.google.pl/scholar?hl=en&q=author%3Aalan-
Search
Web
Images
More...
barwil@gmail.com
Google
author:alan-turing
Scholar
About 329 results (0.12 sec)
My Citations
17
Articles
Case law
My library
Any time
Since 2016
Since 2015
Since 2012
Custom range...
Sort by relevance
Sort by date
include patents
include citations
Create alert
The chemical basis of morphogenesis
AM Turing - ... Transactions of the Royal Society of ..., 1952 - rstb.royalsocietypublishing.org
Abstract It is suggested that a system of chemical substances, called morphogens, reacting together and diffusing through a tissue, is adequate to account for the main phenomena of morphogenesis. Such a system, although it may originally be quite homogeneous, may ...
Cited by 10034 Related articles All 81 versions Web of Science: 120 Cite Saved
[PDF] ufrj.br
[PDF] On computable numbers, with an application to the Entscheidungsproblem
AM Turing - J. of Math, 1936 - people.cs.umass.edu
The" computable" numbers may be described briefly as the real numbers whose expressions as a decimal are calculable by finite means. Although the subject of this paper is ostensibly the computable numbers. it is almost equally easy to define and investigate ...
Cited by 8772 Related articles All 144 versions Web of Science: 1923 Cite Save More
[PDF] umass.edu
[PDF] Computing machinery and intelligence
AM Turing - Mind, 1950 - JSTOR
1. The Imitation Game. Ipropose tO consider the question,'Can machines think? This should begin with definitions of the meaning of the terms' machine'and'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is ...
Cited by 8723 Related articles All 252 versions Cite Save
[PDF] jstor.org
[PDF] Systems of logic based on ordinals
A Turing - 1938 - libarch.nmu.org.ua
The well-known theorem of Godel (Godel [1],[2]) shows that every system of logic is in a certain sense incomplete, but at the same time it indicates means whereby from a system L of logic a more complete system L'may be obtained. By repeating the process we get a ...
Cited by 958 Related articles All 3 versions Cite Save More
[PDF] nmu.org.ua
Full View

# THE CHEMICAL BASIS OF MORPHOGENESIS

By A. M. TURING, F.R.S. *University of Manchester*

(*Received* 9 *November* 1951—*Revised* 15 *March* 1952)

It is suggested that a system of chemical substances, called morphogens, reacting together and diffusing through a tissue, is adequate to account for the main phenomena of morphogenesis. Such a system, although it may originally be quite homogeneous, may later develop a pattern or structure due to an instability of the homogeneous equilibrium, which is triggered off by random disturbances. Such reaction-diffusion systems are considered in some detail in the case of an isolated ring of cells, a mathematically convenient, though biologically unusual system. The investigation is chiefly concerned with the onset of instability. It is found that there are six essentially different forms which this may take. In the most interesting form stationary waves appear on the ring. It is suggested that this might account, for instance, for the tentacle patterns on *Hydra* and for whorled leaves. A system of reactions and diffusion on a sphere is also considered. Such a system appears to account for gastrulation. Another reaction system in two dimensions gives rise to patterns reminiscent of dappling. It is also suggested that stationary waves in two dimensions could account for the phenomena of phyllotaxis.

The purpose of this paper is to discuss a possible mechanism by which the genes of a zygote may determine the anatomical structure of the resulting organism. The theory does not make any new hypotheses; it merely suggests that certain well-known physical laws are sufficient to account for many of the facts. The full understanding of the paper requires a good knowledge of mathematics, some biology, and some elementary chemistry. Since readers cannot be expected to be experts in all of these subjects, a number of elementary facts are explained, which can be found in text-books, but whose omission would make the paper difficult reading.

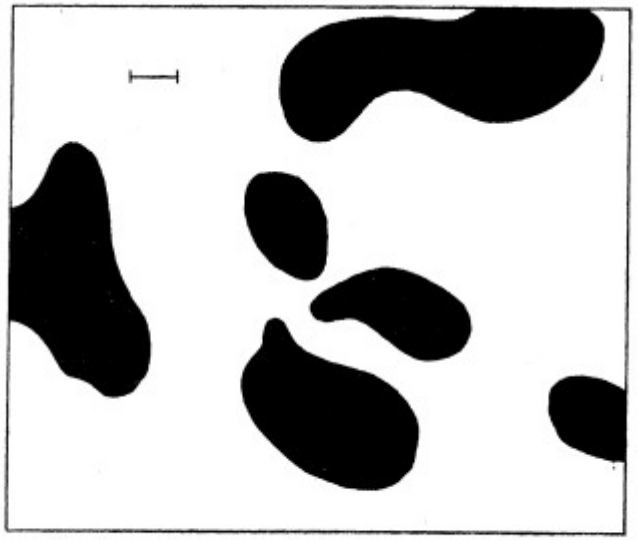
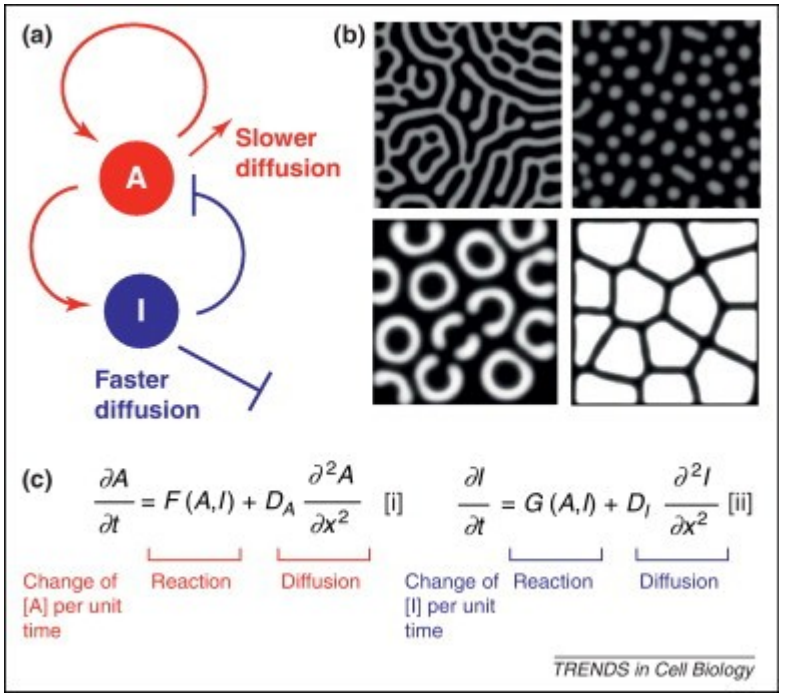
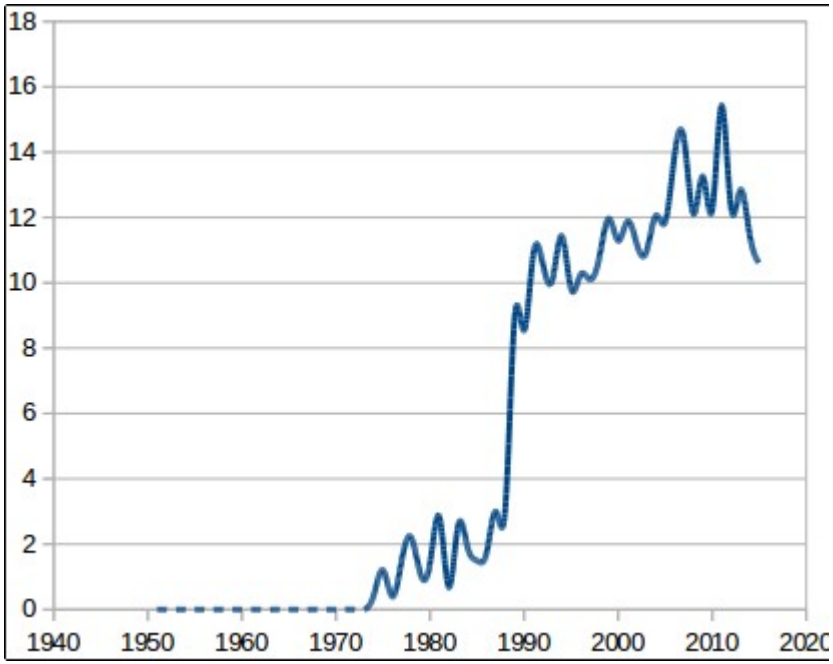


FIGURE 2. An example of a 'dappled' pattern as resulting from a type (*a*) morphogen system. A marker of unit length is shown. See text, §9, 11.

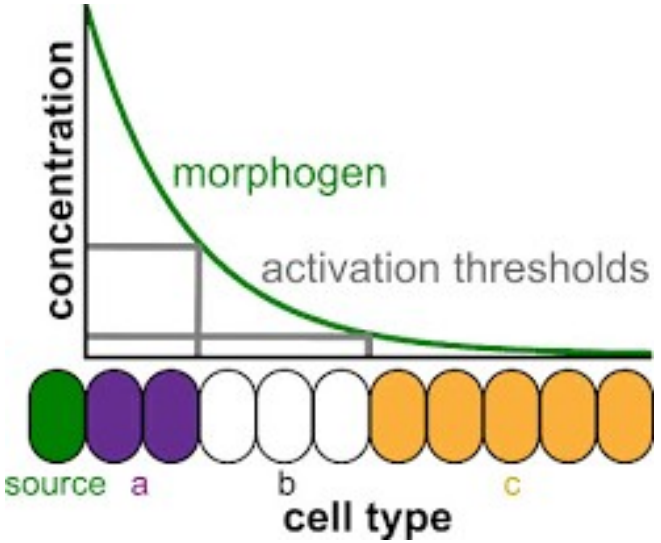


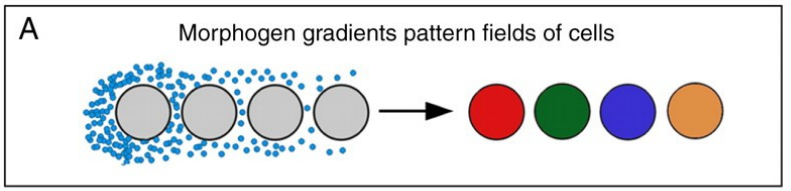

"morphogen" in publications

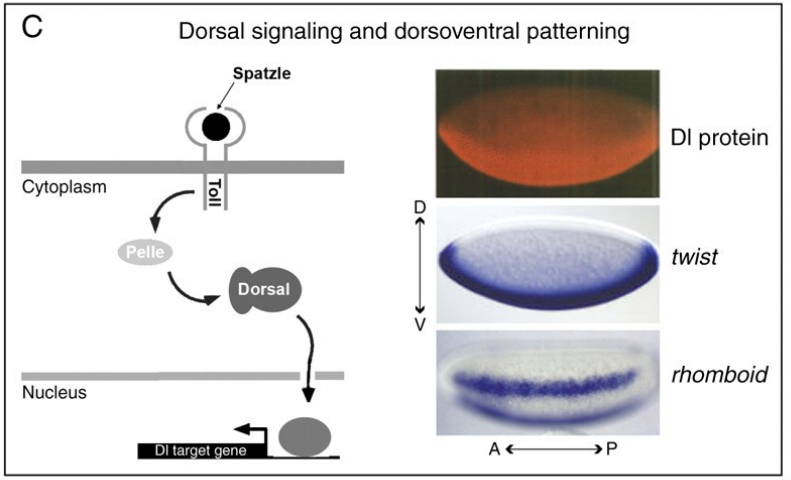
# Molecular morphogens

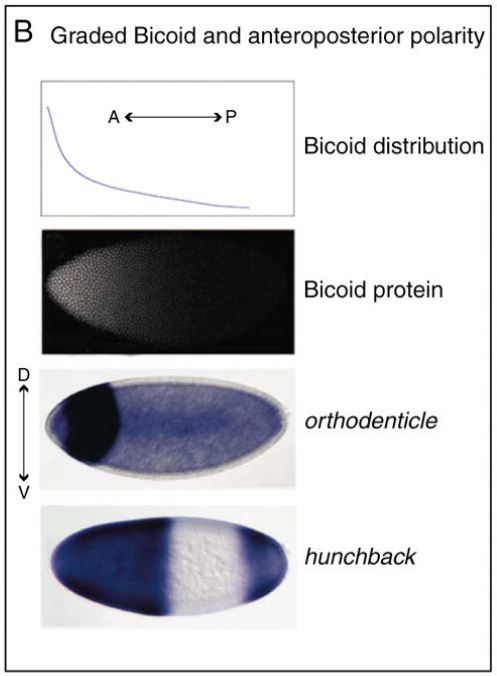
Skin pattern

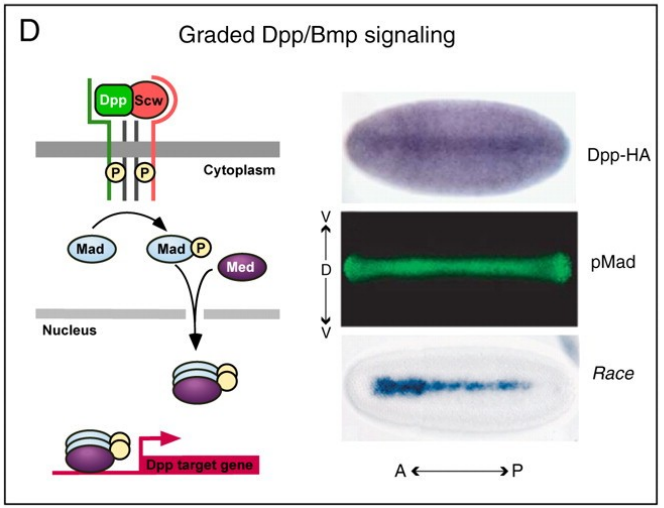


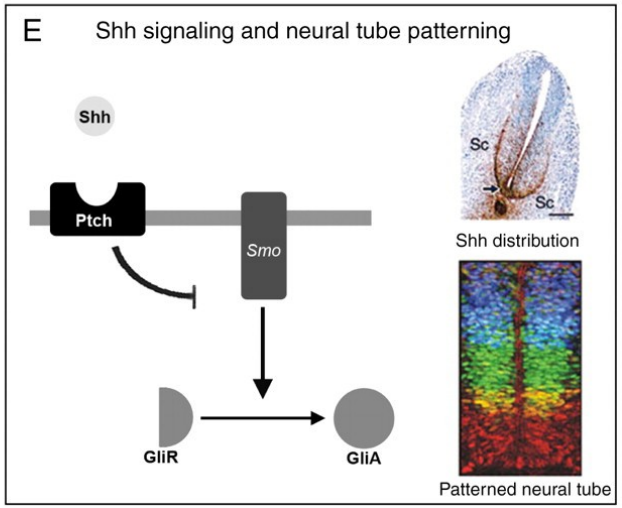

Molecular level



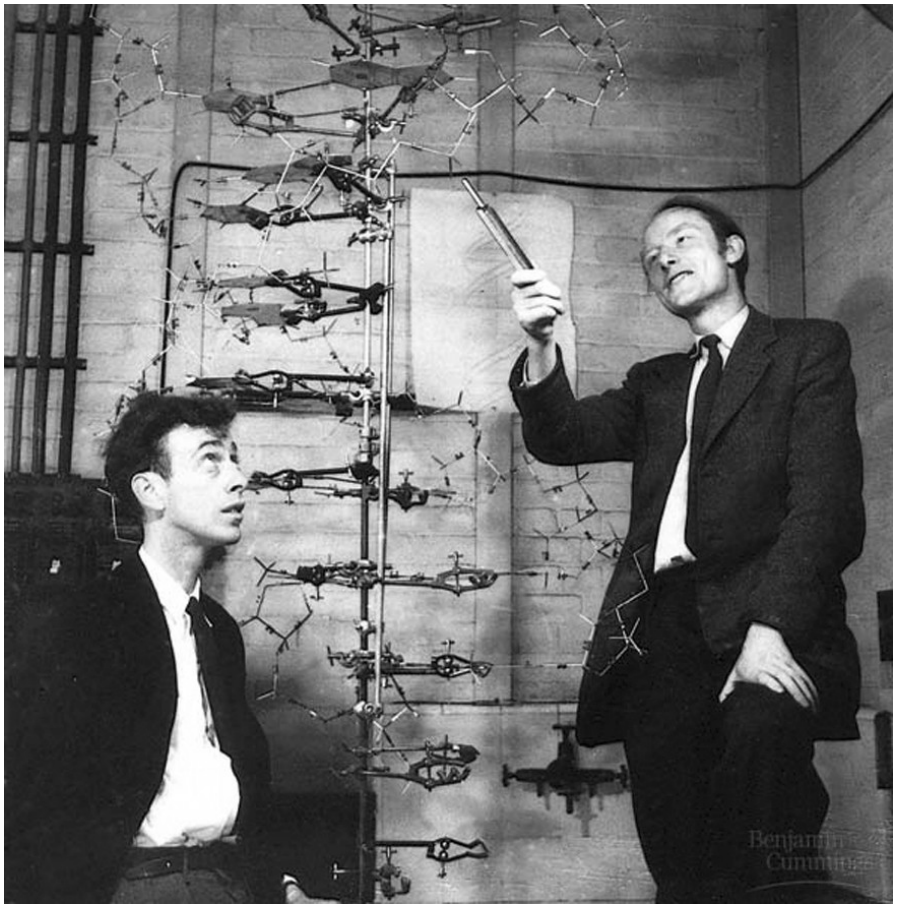



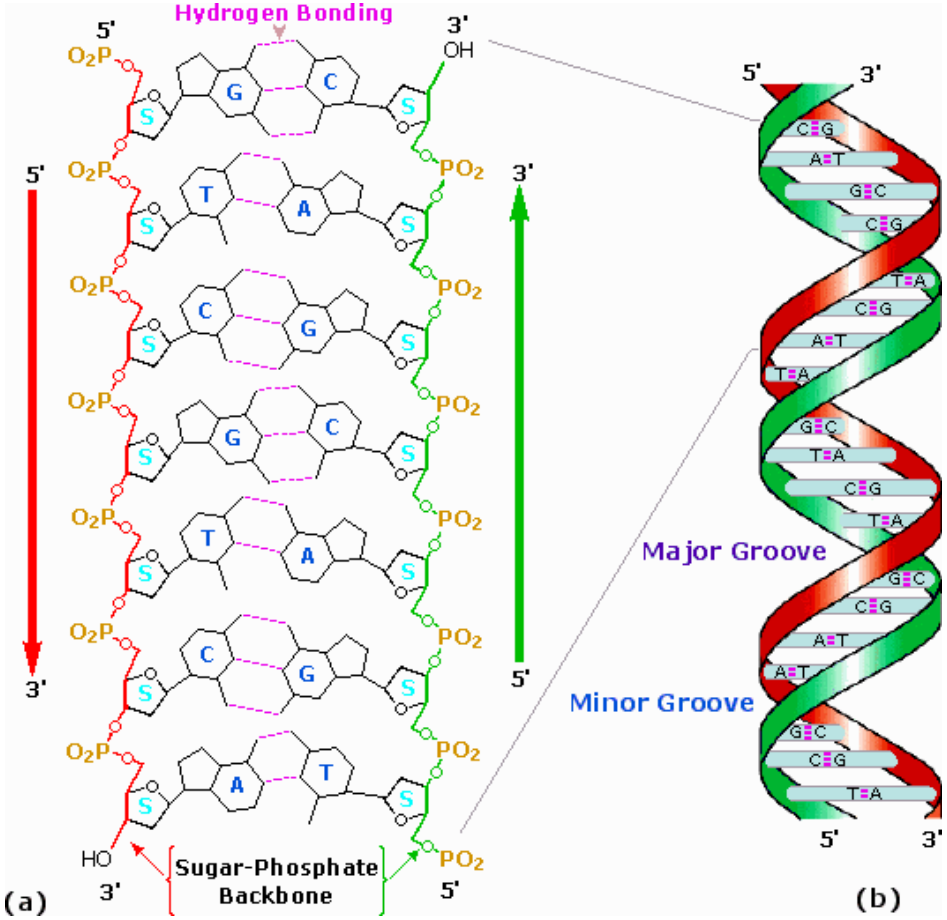



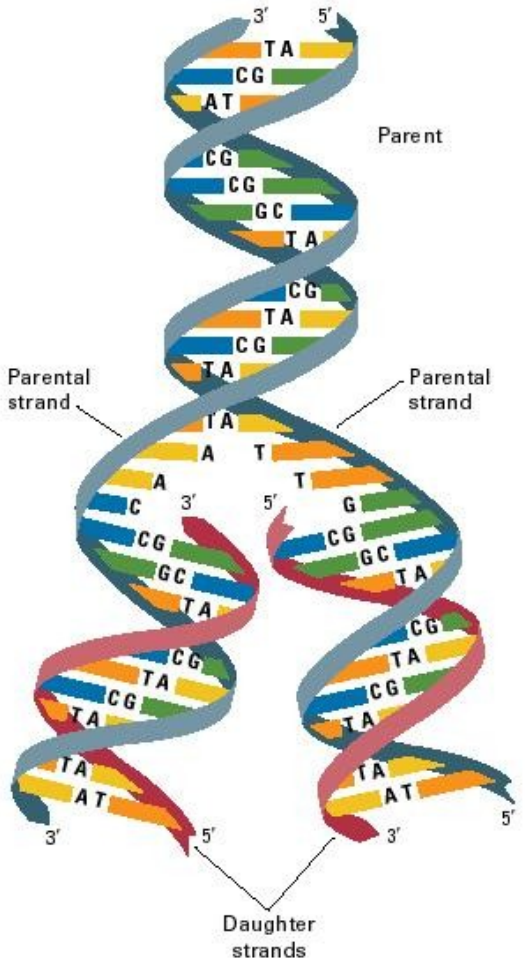
# The foundation of molecular biology



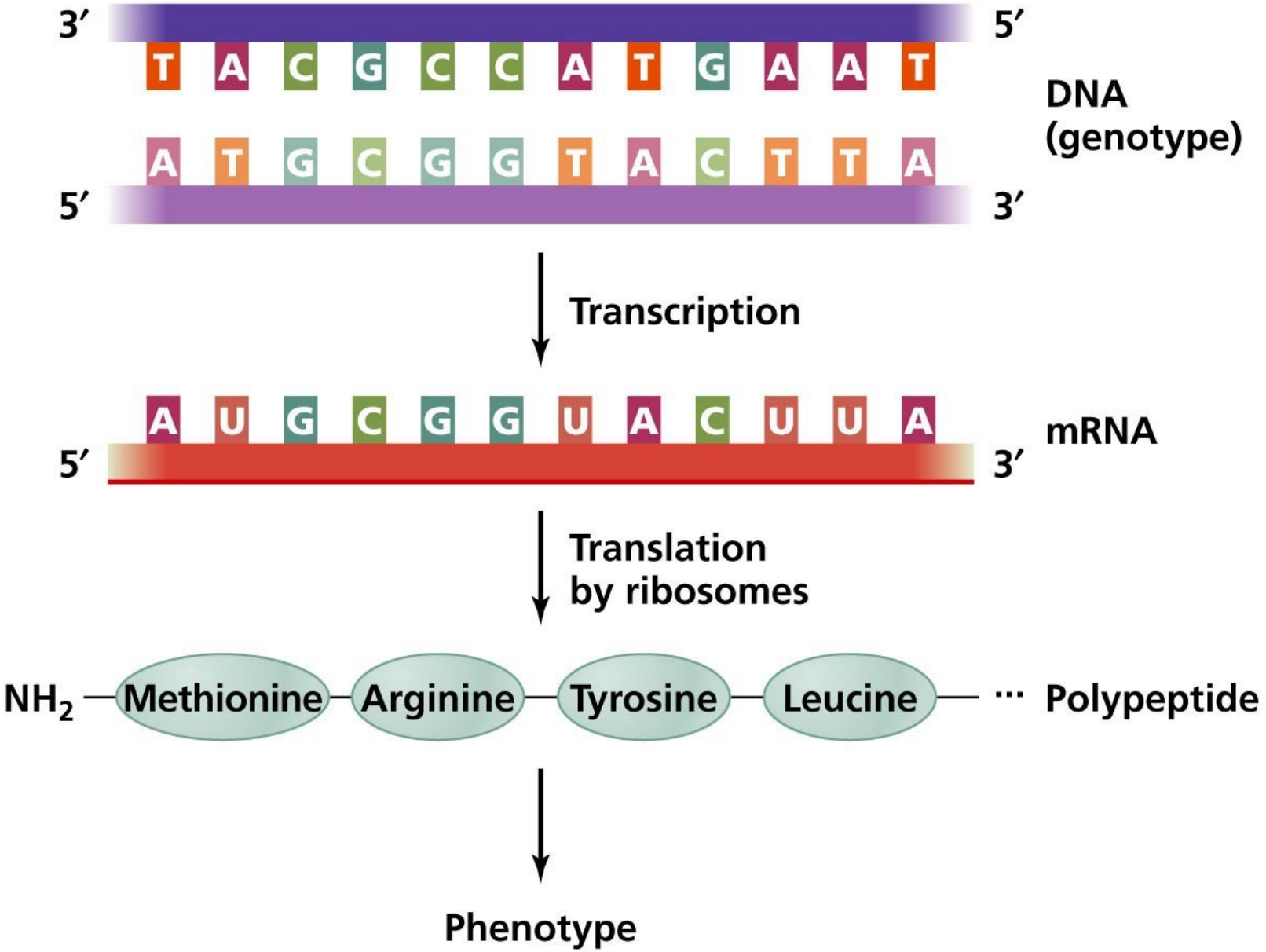
- Watson and Crick publish DNA structure in 1953 (using data from Franklin and Wilkins)
- That leads to understanding of the nature of information storage in DNA
- Now it is possible to have a vastly simplified model of DNA sequence just as a sequence of letters over DNA alphabet, that captures most of the heritable information

# DNA structure

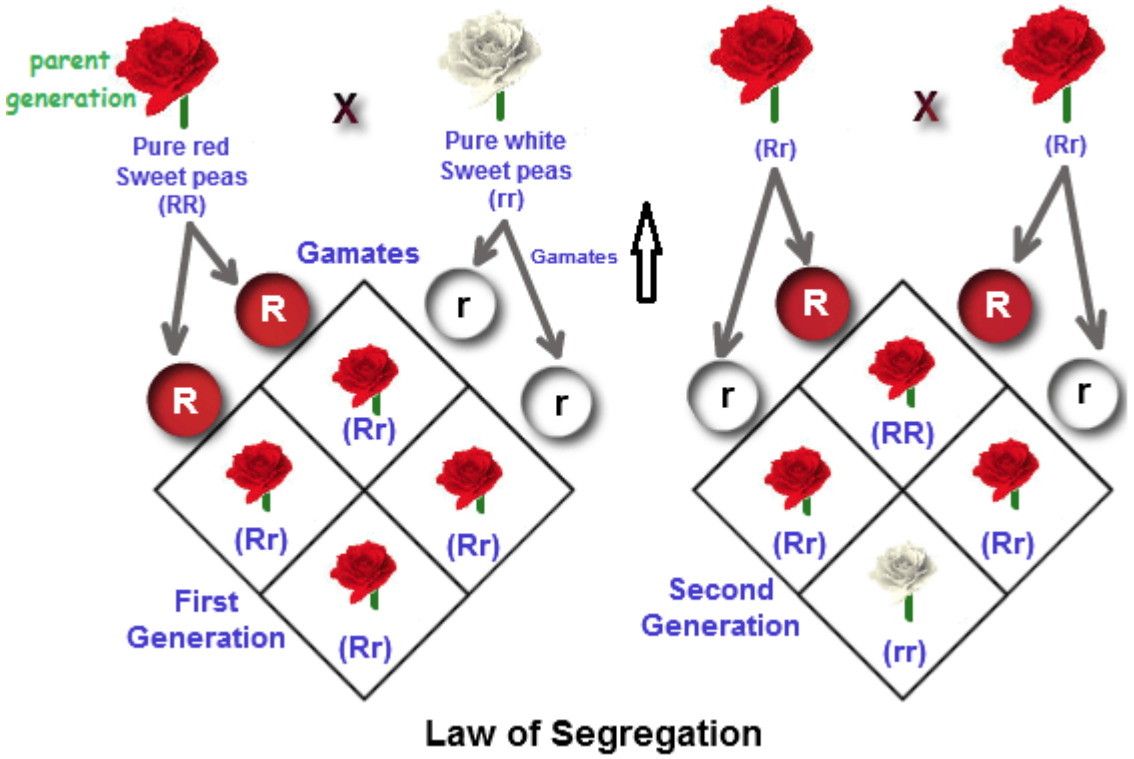


# The DNA is not the only sequence



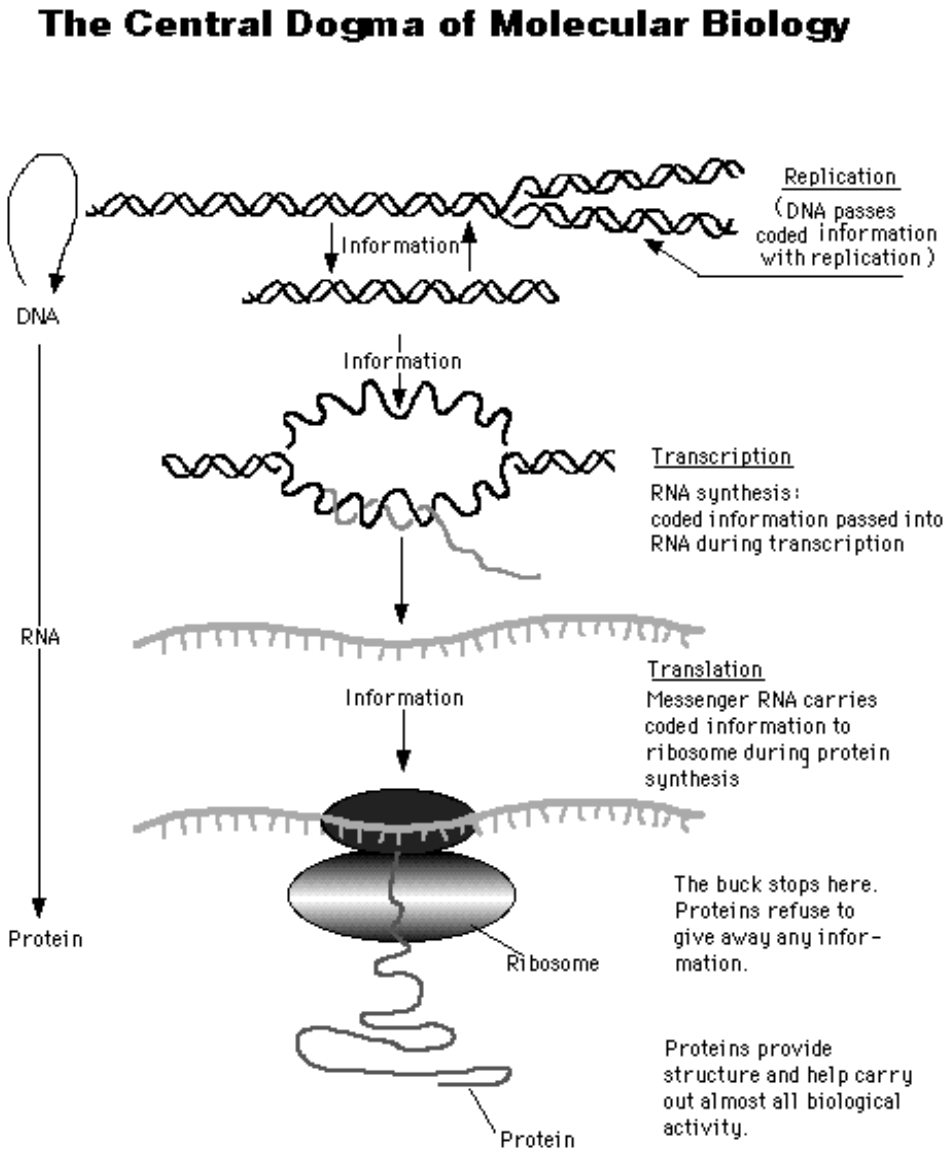
# Another idea ahead of its time

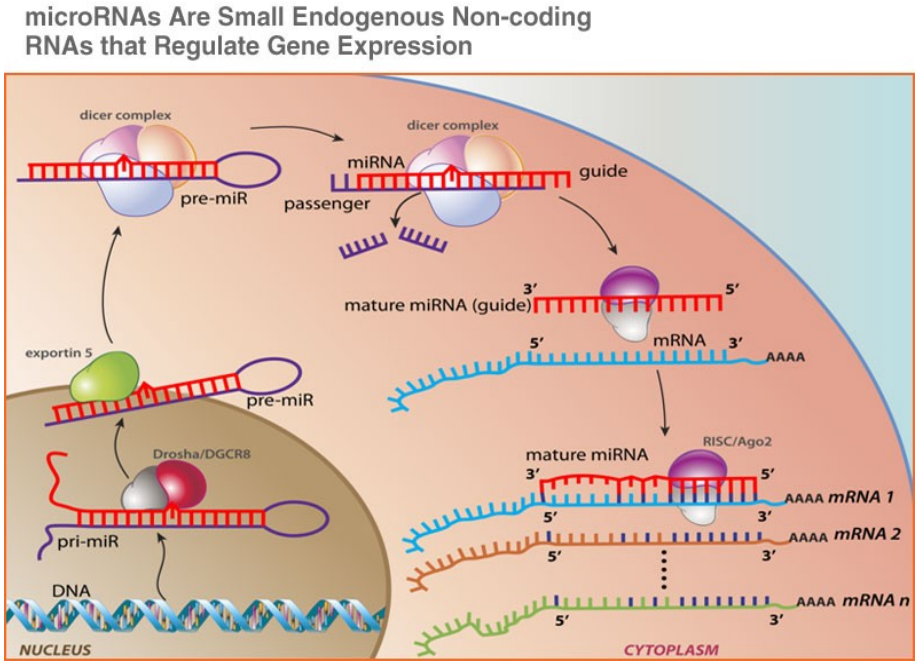


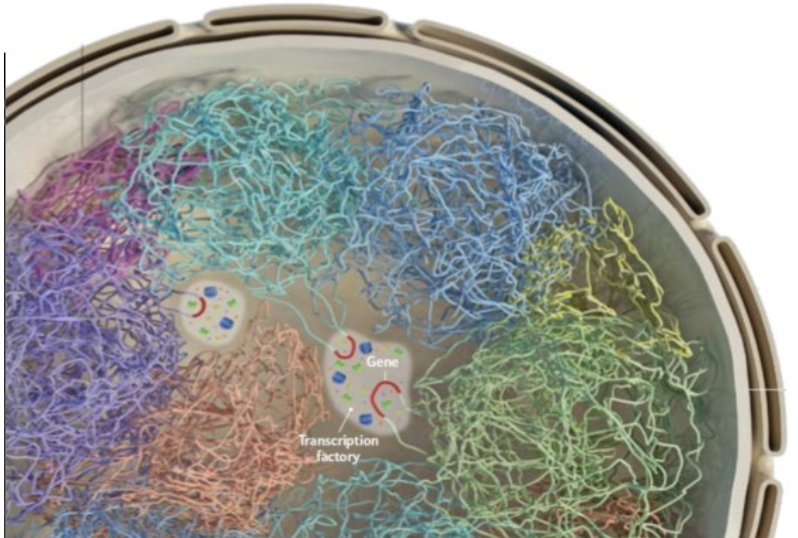

Law of Segregation

- Gregor Mendel (1822 -1884)
- Introduced the idea of “factors” that we now (since early XX century) call genes
- Smallest units of heritable information
- Now we know they reside in DNA

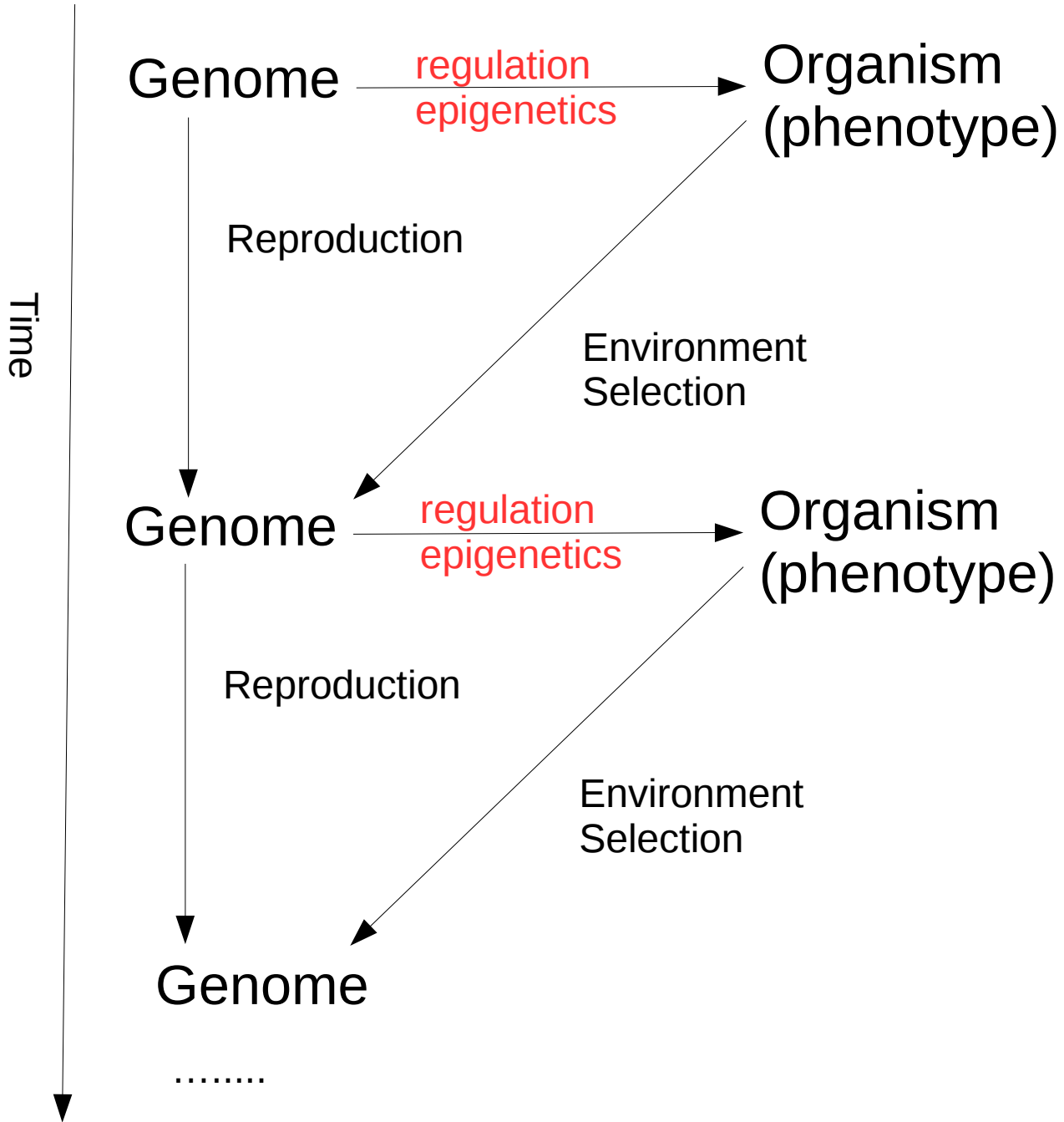
# Where are the genes?



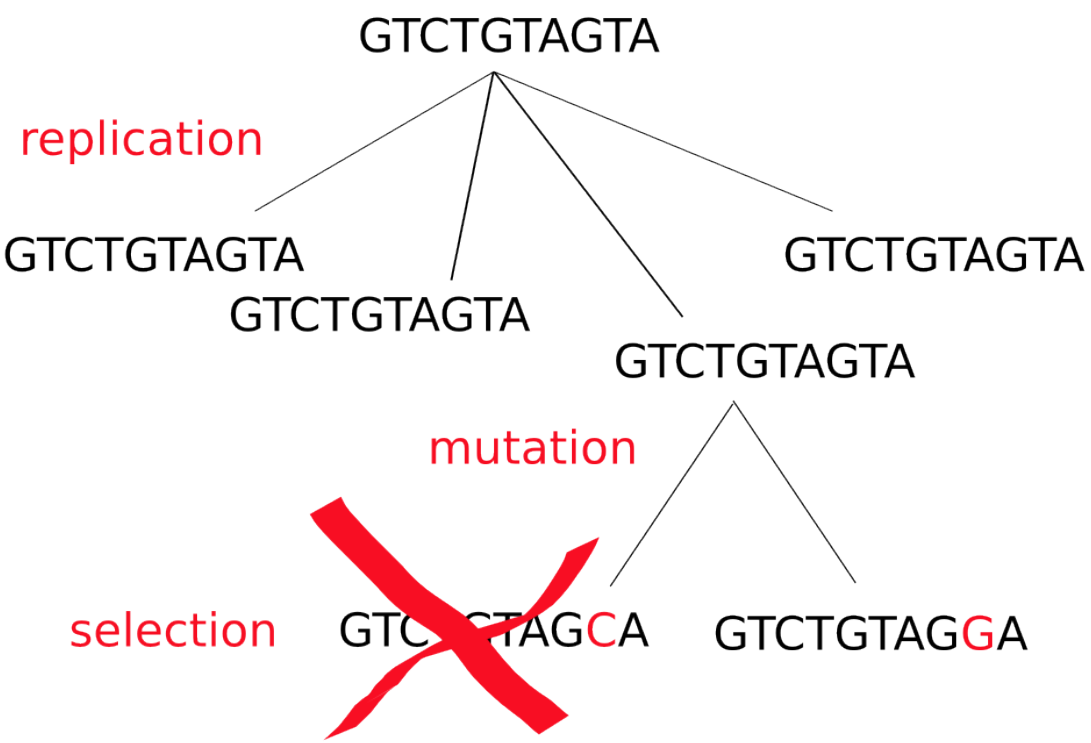



# The really big picture - evolution



# Sequence evolution




- Conceptually simple model, reproduction with mutation
- Mutation rate very small, but given genome sizes and cell number, considerable
- Mutation on the DNA level, selection on the protein level

# Fundamental problem

- How far in evolution are sequences we can observe in different living species?
- More formally: Can we define a measure of sequence similarity

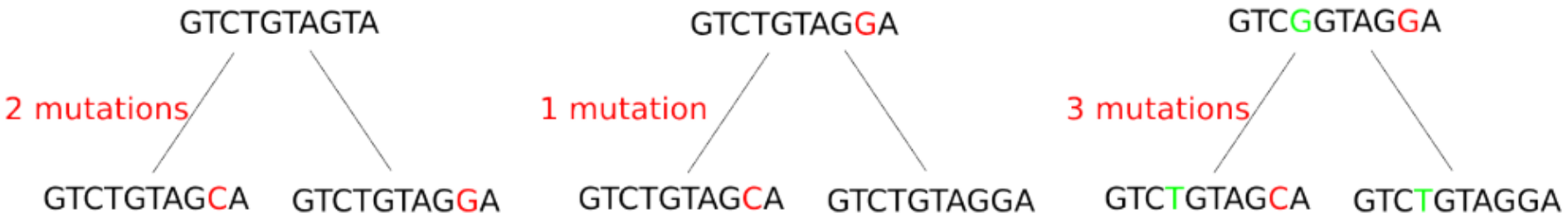
$$d : \Sigma^* \times \Sigma^* \rightarrow \mathcal{R}^+$$

approximating the true evolutionary distance?

- Hint: We should count the number of mutations leading to the observed divergence.

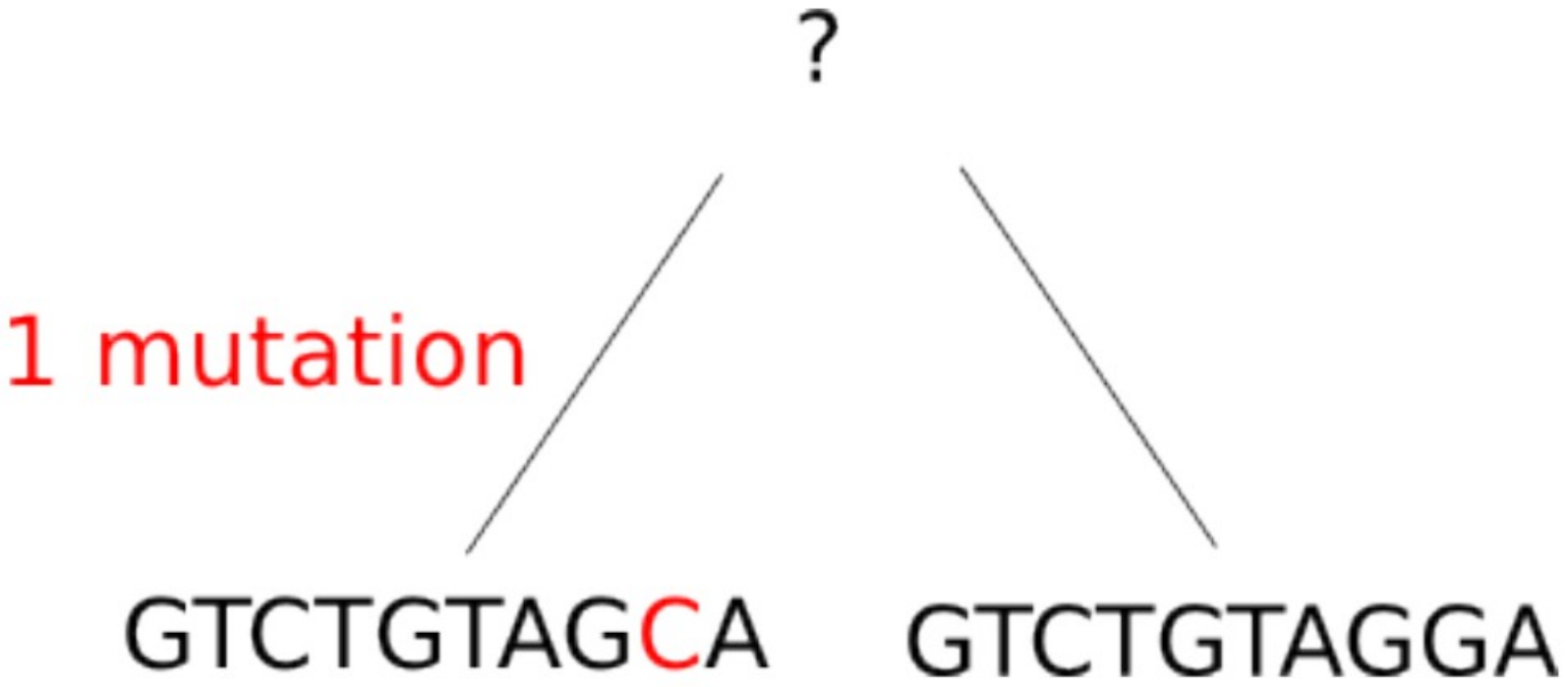
# Lack of data on ancestral DNA

We can observe only the current situation. What about ancestral sequences?




**Solution:** *Parsimony* – In case of lack of evidence for a more complex situation, take the simplest possible explanation.

# Time reversibility




Technically, in order to estimate the ancestral sequence, we need to assume that the process is "time-reversible", i.e. The chances of mutating the sequence $s_1$ into $s_2$ are the same as $s_2$ into $s_1$. This is a reasonable simplification for "short" evolutionary time-scales.

# Naive approach

- Time-reversible Markov Chain (symmetric transition matrix)
- Sequences from $\Sigma^k$ are states (How many of them?)
- Transition probabilities assume independent base substitution
- We need to define a symmetric base *substitution matrix*
- (*) In fact, we should consider a continuous-time Markov chain, to avoid problems with exact generation times...

# More reasonable model – Jukes-Cantor JC-69

Only one parameter: $\mu$

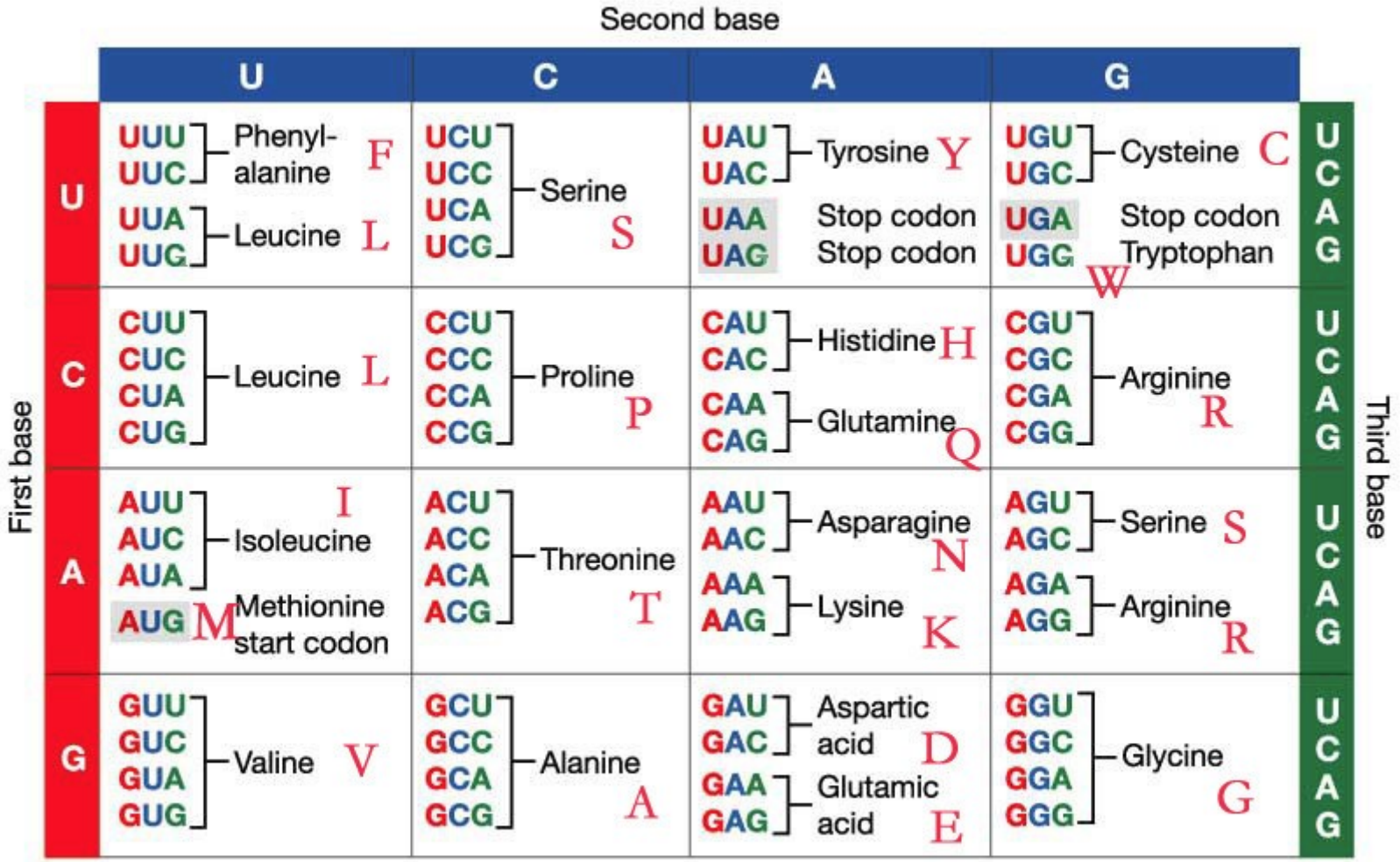
$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

Solution for continuous time $t$:

$$P = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} \end{pmatrix}$$

- Since 1969, many more models: K80, F81, T92, etc, all generalizing for more than just one parameter

# Genetic code is degenerate




- 64 DNA triplets encodes only 20 aminoacids

# Question?

We know two types of mutations in DNA *silent* and *coding*

- Which of them are more interesting for calculating divergence between species?
- And which are more interesting for paternity testing?

# Evolution models based on protein alphabet

- We are still assuming time-reversible Markov chain, but now in space of protein sequences.
- Matrix entries contain log-probabilities, leading to additive measures of similarity
- PAM (Point accepted mutations) matrices (Dayhoff, 1978) describe observed probabilities of occurence of point mutations for a given average divergence (PAM1 = one mutation/100 bases, mostly used PAM250)
- BLOSUM (BLOcks Substitution Matrix) (Henikoff, Henikoff 1992) were constructed using short protein alignments (Blocks) of given sequence identity. e.g.BLOSUM80 was derived from sequences of $\geqslant 80\%$ identity

# Hamming distance

- Hamming distance: a metric originating from Information theory
- Given two vectors of the same length, it returns the number of positions where they differ.
- $$D_H(s_1, s_2) = \Sigma_{i=1}^{n}\{1 : s_1[i] \neq s_2[i]; 0 : otherwise\}$$
- A proper distance (satisfies triangle inequality)

# Errors in DNA are not just substitutions

- DNA polymerase can (rarely) slide over nucleotides
- especially over stretches of low complexity
- this leads to short deletions of DNA after replication
- Transposable elements lead to insertions of larger segments
- Chromosome recombination leads to duplications and deletions on different chromosomes at the same time

# Edit distance

- We can introduce *edit distance*: the number of editing operations needed to transform one sequence into the other. These operations are:
  - Substitutions
  - Insertions
  - Deletions
- The *procedural* definition of the distance makes it difficult to work with
- Does it matter in what order I make the operations (*If i delete a character, I cannot substitute it anymore...*)
- It turns out the *optimal* edit distances are simpler and can be described in a formal way as sequence *alignments*

# Sequence alignment

For a given sequences $s, t$ over an alphabet $\Sigma$, their alignment is a pair of words $s', t'$ over the extended alphabet $\Sigma' = \Sigma \cup \{-\}$. Sequences $s', t'$ need to satisfy the following:

- $|s'| = |t'|$
- $s'_{|\Sigma} = s$ and $t'_{|\Sigma} = t$
- for no position $i$, $s'[i] = t'[i] = -$

For example, one of the words `HEAGAWGHEE` and `PAWHEAE` is

```
HEAGAWGHE-E
--P-AW-HEAE
```

Number of possible alignments for words of length $n$

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \simeq \frac{2^{2n}}{\sqrt{\pi n}}$$

# Simple sequence comparison by dot-plotting

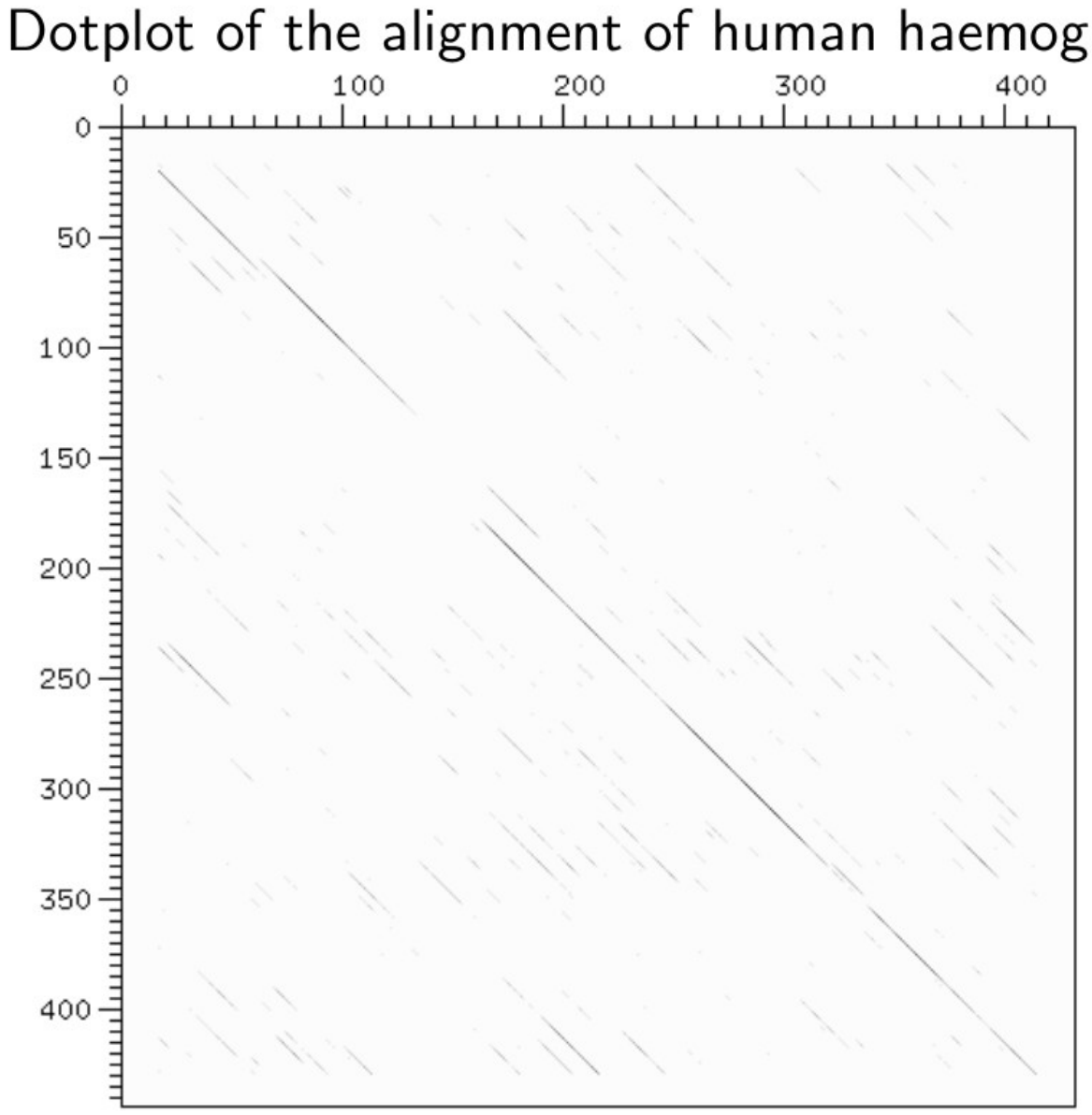
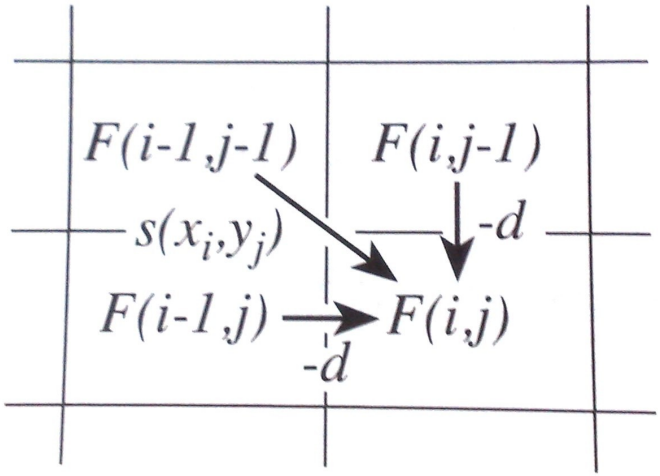



image (c) Hannes Luz

# Needleman-Wunsch dynamic algorithm

| | H | E | A | G | A | W | G | H | E | E |
|---|---|---|---|---|---|---|---|---|---|---|
| P | −2 | −1 | −1 | −2 | −1 | −4 | −2 | −2 | −1 | −1 |
| A | −2 | −1 | **5** | 0 | **5** | −3 | 0 | −2 | −1 | −1 |
| W | −3 | −3 | −3 | −3 | −3 | **15** | −3 | −3 | −3 | −3 |
| H | **10** | 0 | −2 | −2 | −2 | −3 | −2 | **10** | 0 | 0 |
| E | 0 | **6** | −1 | −3 | −1 | −3 | −3 | 0 | **6** | **6** |
| A | −2 | −1 | **5** | 0 | **5** | −3 | 0 | −2 | −1 | −1 |
| E | 0 | **6** | −1 | −3 | −1 | −3 | −3 | 0 | **6** | **6** |

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j), \\ F(i-1, j) - d, \\ F(i, j-1) - d. \end{cases}$$




| | | H | E | A | G | A | W | G | H | E | E |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 ← | **−8** ← | **−16** ← | −24 ← | −32 ← | −40 ← | −48 ← | −56 ← | −64 ← | −72 ← | −80 |
| P | −8 | −2 | −9 | **−17** ← | **−25** | −33 ← | −41 ← | −49 ← | −57 | −65 | −73 |
| A | −16 | −10 | −3 | −4 ← | −12 | **−20** ← | −28 ← | −36 ← | −44 ← | −52 ← | −60 |
| W | −24 | −18 | −11 | −6 | −7 | −15 | **−5** ← | **−13** ← | −21 ← | −29 ← | −37 |
| H | −32 | −14 | −18 | −13 | −8 | −9 | −13 | −7 | **−3** ← | −11 ← | −19 |
| E | −40 | −22 | −8 ← | −16 | −16 | −9 | −12 | −15 | −7 | **3** | −5 |
| A | −48 | −30 | −16 | −3 ← | −11 | −11 | −12 | −12 | −15 | **−5** | 2 |
| E | −56 | −38 | −24 | −11 | −6 | −12 | −14 | −15 | −12 | −9 | **1** |

Images adapted from Durbin et al.

# Smith-Waterman – local version of alignment

- If we add 0 to the dynamic algorithm formula
- We get a local version of the algorithm, giving us the best matching substrings

$$F(i,j) = \max \begin{cases} 0, \\ F(i-1, j-1) + s(x_i, y_j), \\ F(i-1, j) - d, \\ F(i, j-1) - d. \end{cases}$$

| | | H | E | A | G | A | W | G | H | E | E |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A | 0 | 0 | 0 | 5 | 0 | **5** | 0 | 0 | 0 | 0 | 0 |
| W | 0 | 0 | 0 | 0 | 2 | 0 | **20** | **12** | 4 | 0 | 0 |
| H | 0 | 10 | 2 | 0 | 0 | 0 | 12 | 18 | **22** | 14 | 6 |
| E | 0 | 2 | 16 | 8 | 0 | 0 | 4 | 10 | 18 | **28** | 20 |
| A | 0 | 0 | 8 | 21 | 13 | 5 | 0 | 4 | 10 | 20 | 27 |
| E | 0 | 0 | 6 | 13 | 18 | 12 | 4 | 0 | 4 | 16 | 26 |

# Inconsistencies in pairwise alignments

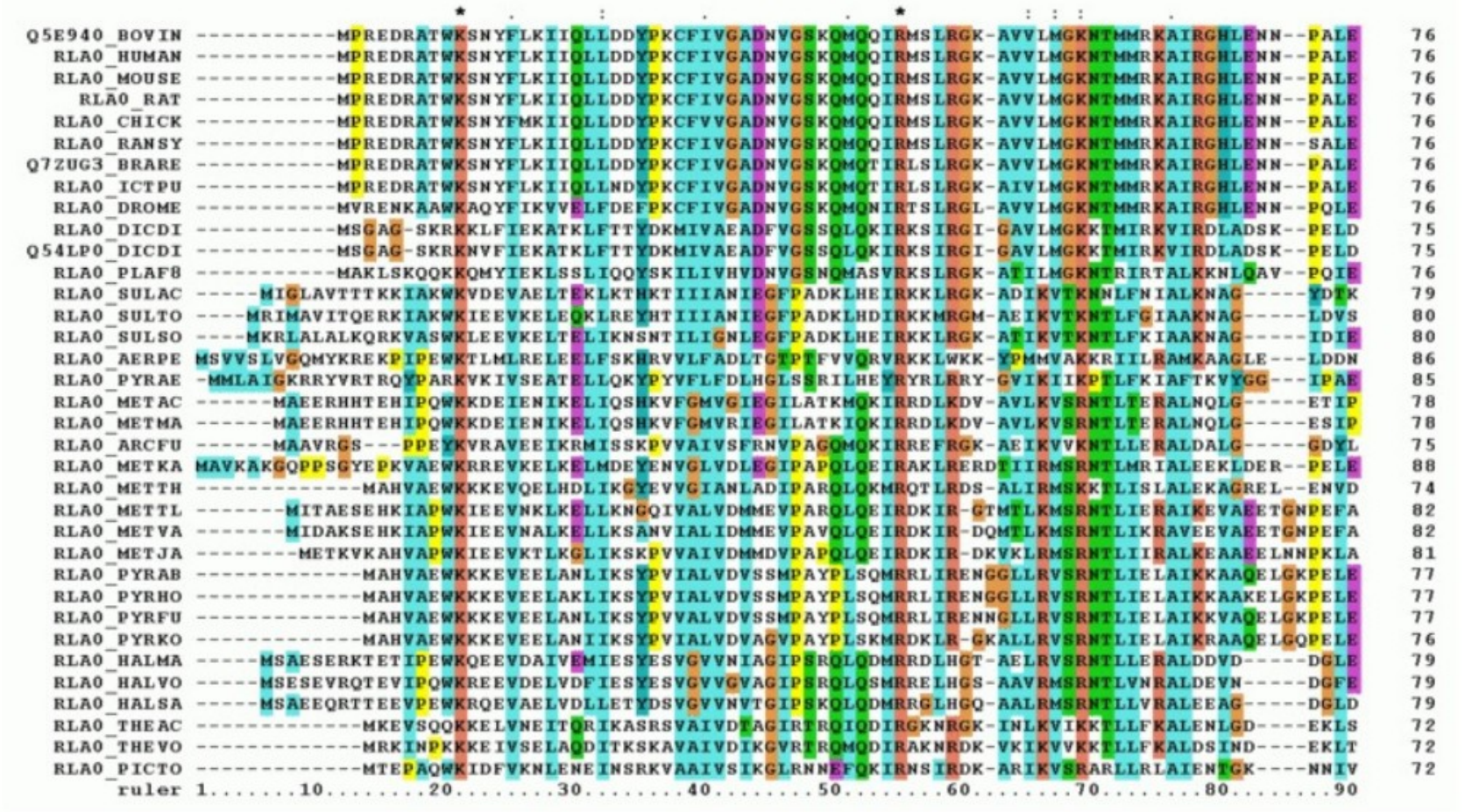
Pairwise alignments used to calculate distances (and reconstruct a tree) may lead to inconsistent picture. For example consider alignment of all pairs of 3 sequences: `CAAC`, `AACA`, `ACAA`

- `CAAC-` `AACA-` `ACAA-`
- `-AACA` `-ACAA` `-CAAC`

Which `C` in `CAAC` was in the ancestral sequence?

# A consistent alignment of many sequences

Can we make a generalization of the pairwise alignment idea?




*Note: the correspondence with edit distance is lost*

image (c) P. Winter

# Scoring multiple sequence alignments (MSAs)

What is the natural way to score multiple alignment "quality"?

- Assume column independence (as usual)
- Sum of pairs (SP) score

$$S(m_i) = \Sigma_{k<l} s(m_i^k, m_i^l)$$

- Does it work well (for counting parsimonous mutations)?
  *hint: Consider a column with all characters different.*
- How much does it overestimate the number of necessary mutations?

# Complexity of finding the optimal multiple alignment

- Dynamic algorithm has the cost of $\mathcal{O}(n^k)$
- In general, the problem is NP-Complete
- We can still try to slightly improve its performance by looking at the lower bound of alignment score (Carillo-lipman)

$$\sigma(a) \leq S(a^{kl}) - S(\hat{a}^{kl}) + \sum_{k'<l'} S(\hat{a}^{k'l'})$$

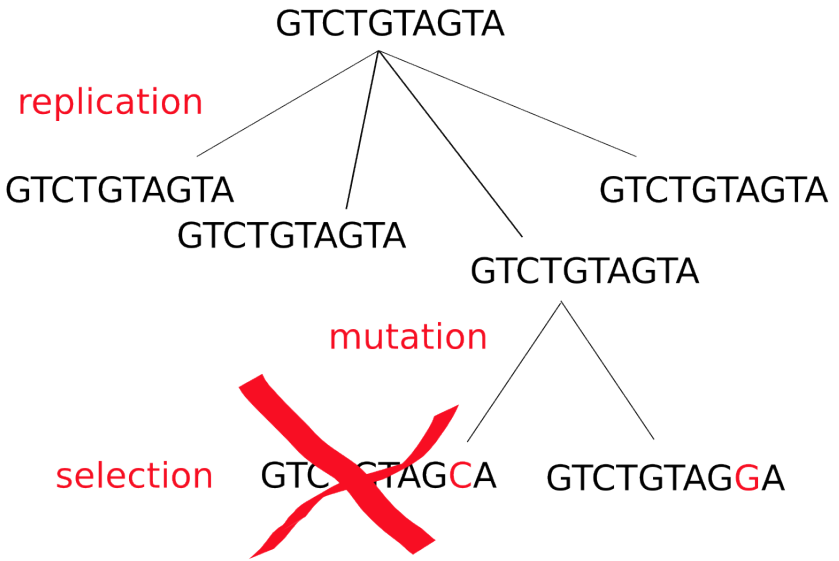
$$S(a^{kl}) \geq \beta^{kl}$$

$$\text{where } \beta^{kl} = \sigma(a) + S(\hat{a}^{kl}) - \sum_{k'<l'} S(\hat{a}^{k'l'}).$$

# Can we overcome the complexity issue?

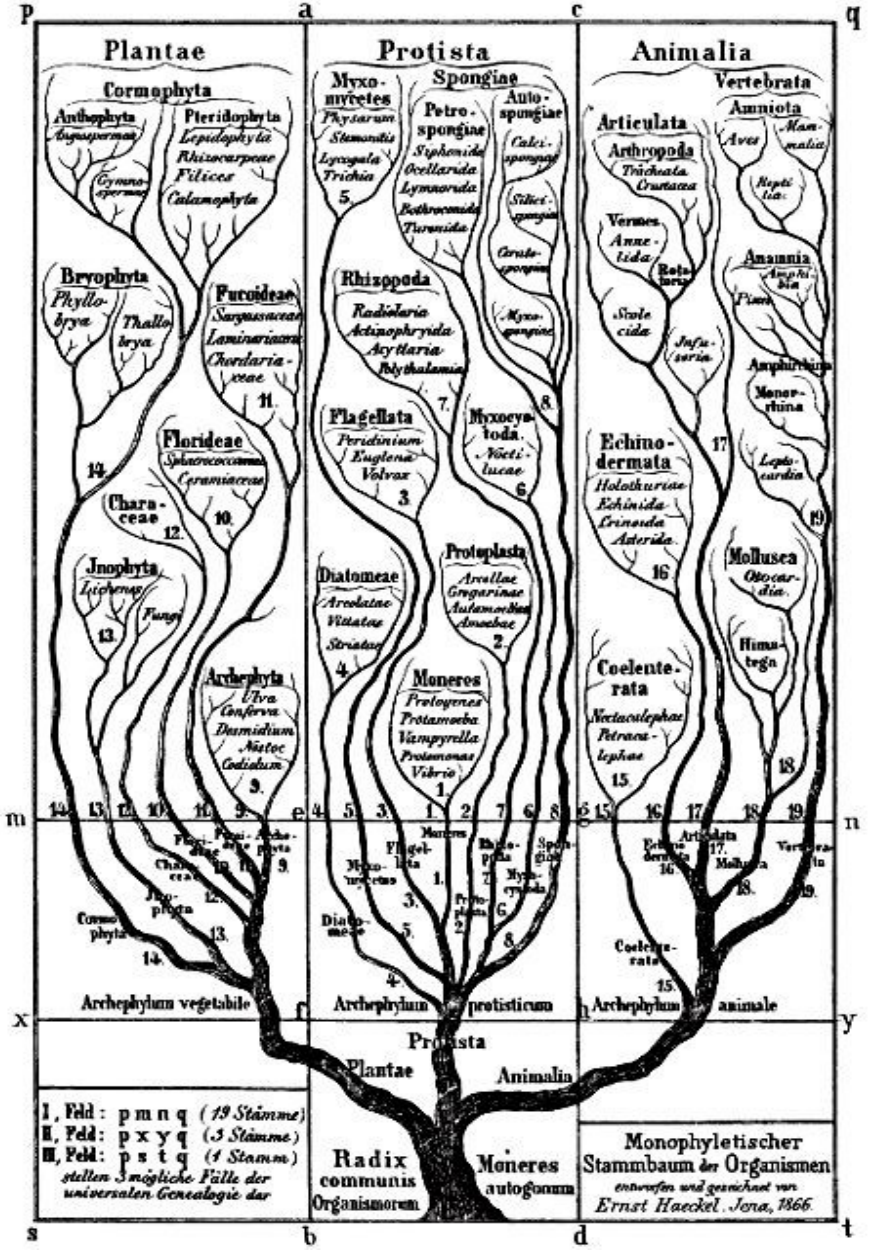
- Theoretically, we could try to prove that P=NP, and then solve MSA
- In practice, we are not (usually) making multiple alignments of random sequences. Usually we know they are related
- Can we use the knowledge that they originated from an evolutionary process to guide our search for optimal MSA?

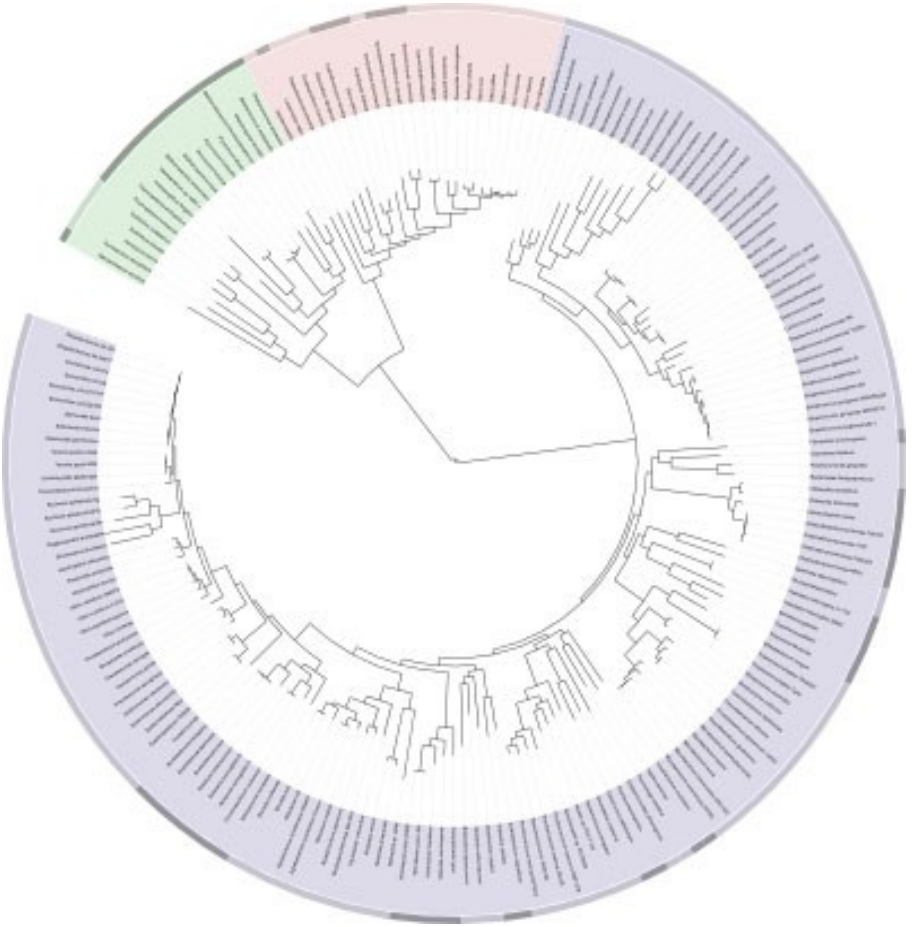
# Back to how evolution works




- Tree-like model of sequence evolution
- Common ancestor - root
- Internal nodes – ancestral sequences
- Leafs – curently available sequence pool or dead-ends

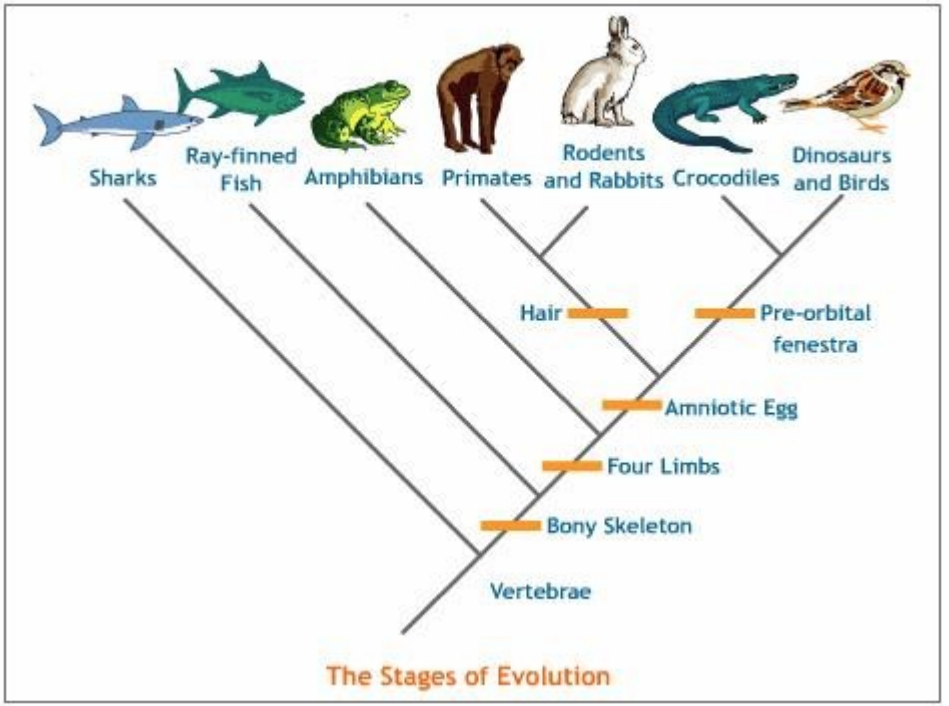
# The tree of life hypothesis



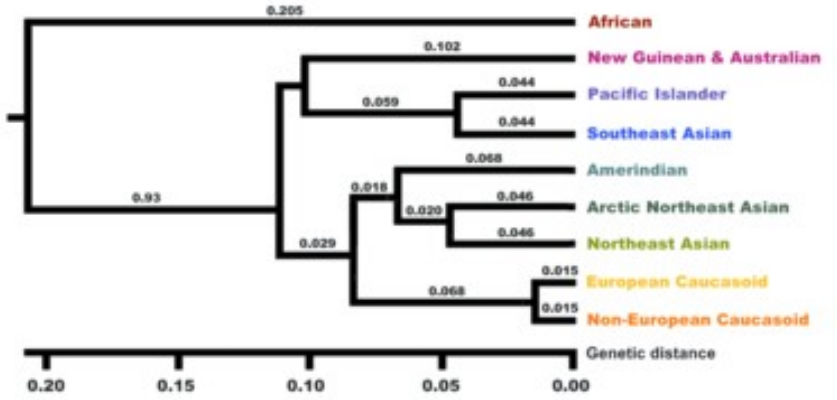



Interactive Tree of Life
http://itol.embl.de/

# Evolution of species and within species




Linkage tree for 9 population clusters showing genetic distances ($F_{ST}$)
(Cavalli-Sforza *et al.*, 1994:80)




$F_{ST}$ distance matrix for the 9 clusters shown above
(x10,000 with standard errors obtained by bootstrap analysis)

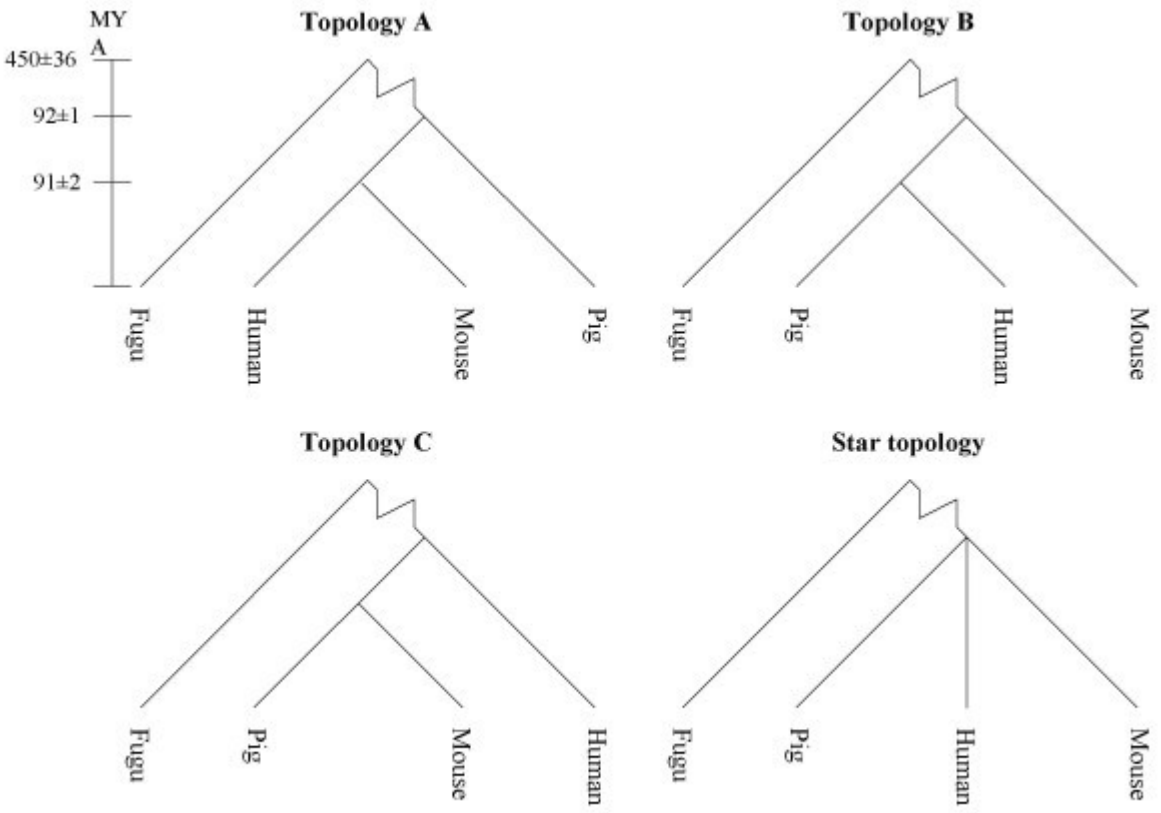
| | AFR | NEC | EUC | NEA | ANE | AME | SEA | PAI | NGA |
|---|---|---|---|---|---|---|---|---|---|
| African | 0.0 | | | | | | | | |
| Non-European Caucasian | 1340.0 ± 301 | 0.0 | | | | | | | |
| European Caucasian | 1655.6 ± 416 | 154.7 ± 29 | 0.0 | | | | | | |
| Northeast Asian | 1979.1 ± 452 | 640.4 ± 134 | 938.2 ± 217 | 0.0 | | | | | |
| Arctic Northeast Asian | 2008.5 ± 387 | 708.2 ± 160 | 746.7 ± 210 | 459.7 ± 98 | 0.0 | | | | |
| Amerindian | 2261.4 ± 434 | 955.5 ± 204 | 1038.2 ± 276 | 746.5 ± 183 | 577.4 ± 89 | 0.0 | | | |
| Southeast Asian | 2206.3 ± 529 | 939.6 ± 262 | 1240.4 ± 339 | 630.5 ± 299 | 1039.4 ± 326 | 1341.7 ± 418 | 0.0 | | |
| Pacific Islander | 2505.4 ± 648 | 953.7 ± 230 | 1344.7 ± 354 | 723.8 ± 262 | 1181.2 ± 331 | 1740.7 ± 544 | 436.7 ± 87 | 0.0 | |
| New Guinean and Australian | 2472.0 ± 536 | 1179.1 ± 189 | 1345.7 ± 231 | 734.4 ± 118 | 1012.5 ± 257 | 1457.9 ± 283 | 1237.9 ± 277 | 808.7 ± 264 | 0.0 |

# Finding the phylogenetic tree

- We are interested in measuring evolutionary distances by looking at molecular sequences
- We expect *distances* to *grow* with **decreasing similarity**
- The sequence alignment problem allows us to find the optimal alignment, however the score of the alignment is a measure of *similarity*, rather than distance
- problem of *maximizing similarity* is similar to *minimizing distance*
- However:
  - We expect $d(x, x) = 0$ while for most $a, b$ , $sim(a, a) \neq sim(b, b)$
  - Distances have triangle inequality, and similarities not

# Bifurcating or multifurcating trees

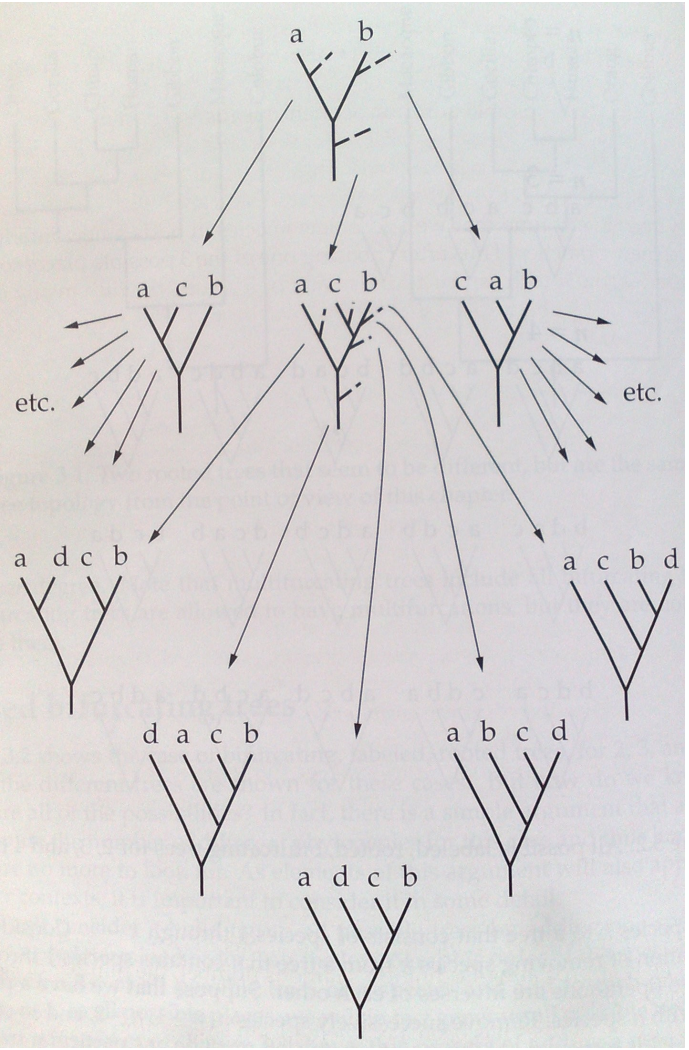
- Even though real evolution might very well include multifurcating nodes (i.e. the speciation events involving more species)
- It is enough to consider binary trees (which may lead to mutliple binary tree topologies)



# How many different binary trees?

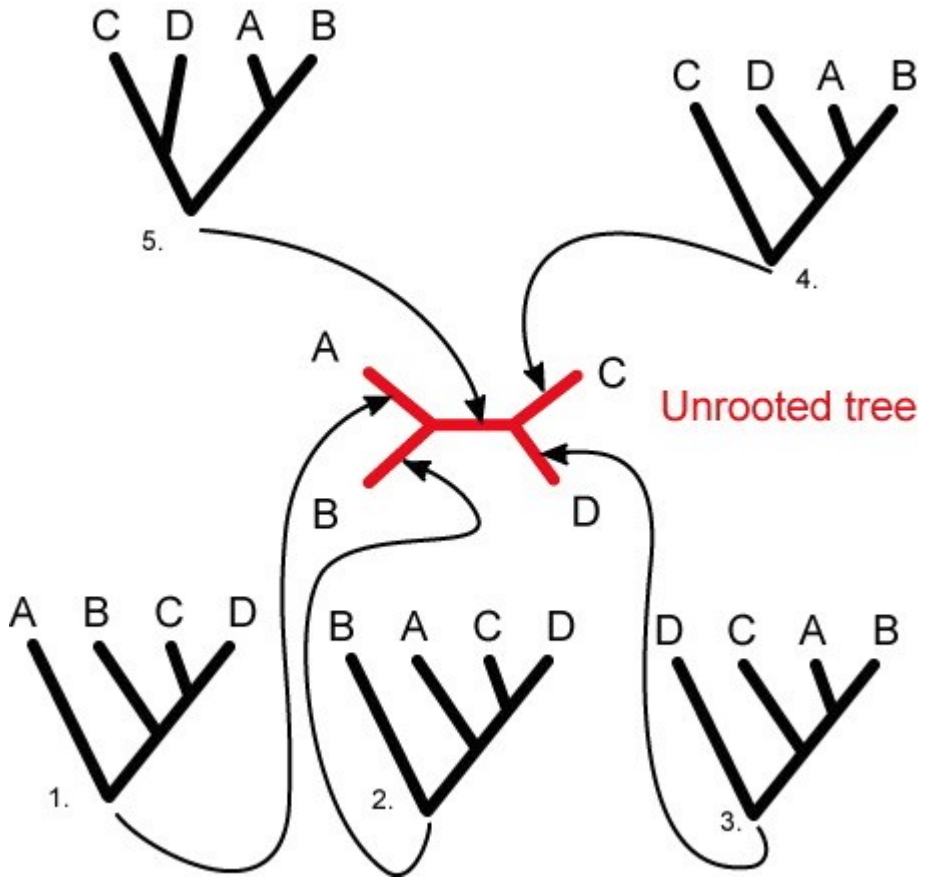
- How many different binary trees can there be for the given N sequences?
- The answer is the Catalan number sequence

  (2(n-1))!/((n-1)!n!)



# Rooted vsa unrooted trees

- Many different rooted trees actually correspond to the same unrooted tree topology
- This unrooted tree with branch lengths can correspond to a distance matrix



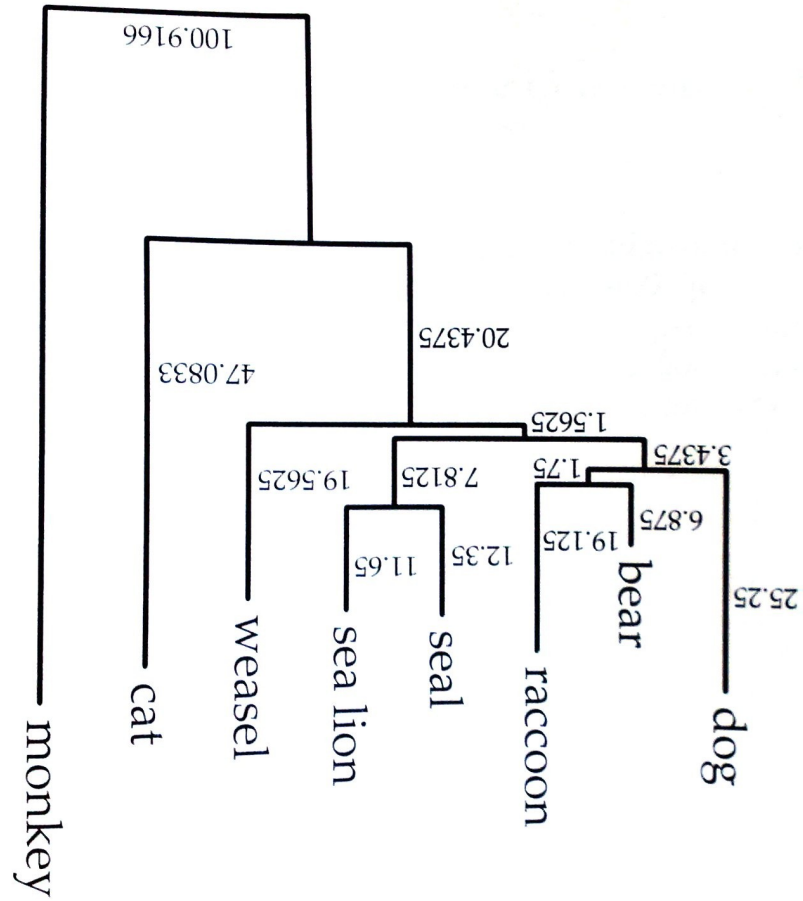
# Reconstructing a tree from distance matrix

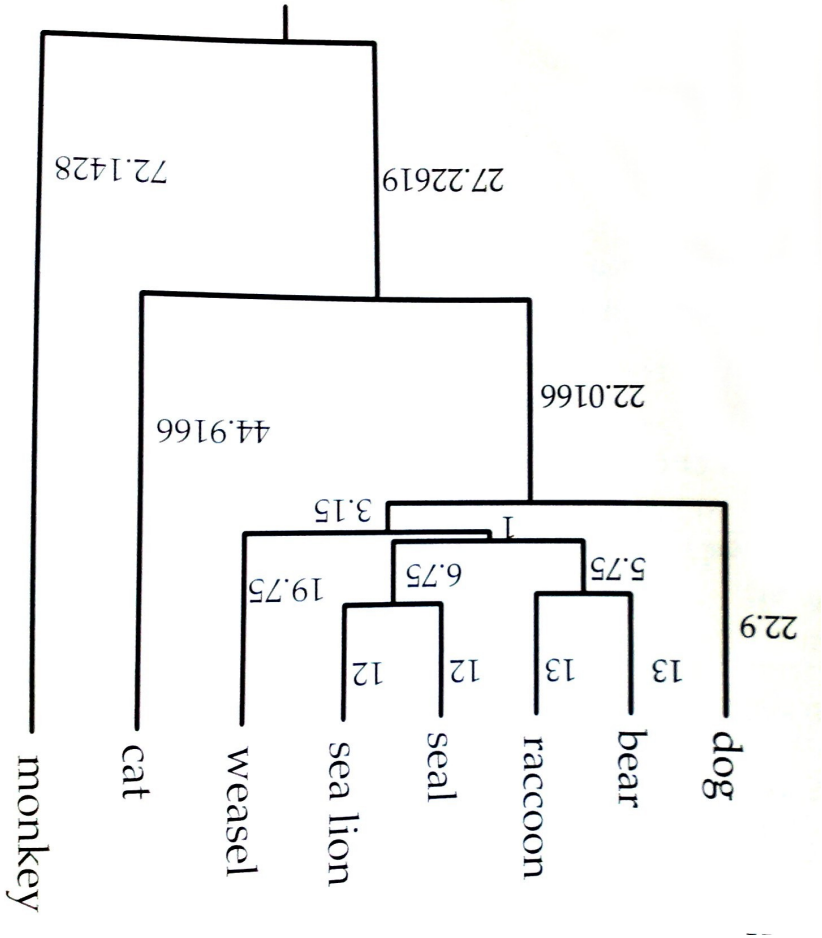
- Given a tree with branch lengths $T$, we can easily generate distance matrix $d_{ij}$
- Can we solve the reverse problem, and how does it relate to the original problem?
- Formally, for a given distance matrix $D$, we want to find a labelled tree $T$, optimizing the least squares criterion:

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (D_{ij} - d_{ij})^2$$

- In general, it is equivalent to solving the Steiner tree problem – one of the the original NP-complete problems
- Can we find any approximate or specialized solutions?

# Non-ultrametric vs Ultrametric trees



# Ultrametric vs metric

- Any metric requires:

$$
\begin{aligned}
d(x,y) &> 0 \qquad \text{for } x \neq y \\
d(x,y) &= 0 \qquad \text{for } x = y \\
d(x,y) &= d(y,x) \qquad \forall\, x,y \\
d(x,y) &\leq d(x,z) + d(y,z) \qquad \forall\, x,y,z \quad \text{(triangle inequality)}
\end{aligned}
$$

- If it is ultrametric it also satisfies, that any 3 leaves can be renamed x,y,z so that:

$$d(x,y) \leq d(x,z) = d(y,z)$$

# UPGMA (Unweighted Pair Group Method with Arithmetic Mean)

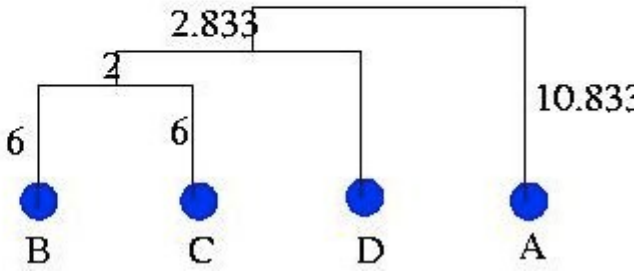
1. Find the $i$ and $j$ that have the smallest distance, $D_{ij}$.
2. Create a new group, $(ij)$, which has $n_{(ij)} = n_i + n_j$ members.
3. Connect $i$ and $j$ on the tree to a new node [which corresponds to the new group $(ij)$]. Give the two branches connecting $i$ to $(ij)$ and $j$ to $(ij)$ each length $D_{ij}/2$.
4. Compute the distance between the new group and all the other groups (except for $i$ and $j$) by using:

$$D_{(ij),k} = \left(\frac{n_i}{n_i + n_j}\right) D_{ik} + \left(\frac{n_j}{n_i + n_j}\right) D_{jk}$$

5. Delete the columns and rows of the data matrix that correspond to groups $i$ and $j$, and add a column and row for group $(ij)$.
6. If there is only one item in the data matrix, stop. Otherwise, return to step 1.

# How does it work?

| | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 17 | 21 | 27 |
| B | | 0 | 12 | 18 |
| C | | | 0 | 14 |
| D | | | | 0 |




- We start from a matrix and finish with an ultrametric tree
- If the matrix is not ultrametric, the result might not be optimal

# Neighbor-joining

1. For each tip, compute $u_i = \sum_{j:j\neq i}^{n} D_{ij}/(n-2)$. Note that the denominator is (deliberately) not the number of items summed.

2. Choose the $i$ and $j$ for which $D_{ij} - u_i - u_j$ is smallest.

3. Join items $i$ and $j$. Compute the branch length from $i$ to the new node ($v_i$) and from $j$ to the new node ($v_j$) as

$$\begin{aligned} v_i &= \tfrac{1}{2}D_{ij} + \tfrac{1}{2}(u_i - u_j) \\ v_j &= \tfrac{1}{2}D_{ij} + \tfrac{1}{2}(u_j - u_i) \end{aligned}$$

4. Compute the distance between the new node $(ij)$ and each of the remaining tips as

$$D_{(ij),k} = (D_{ik} + D_{jk} - D_{ij}) \Big/ 2$$

5. Delete tips $i$ and $j$ from the tables and replace them by the new node, $(ij)$, which is now treated as a tip.

6. If more than two nodes remain, go back to step 1. Otherwise, connect the two remaining nodes (say, $\ell$ and $m$) by a branch of length $D_{\ell m}$.

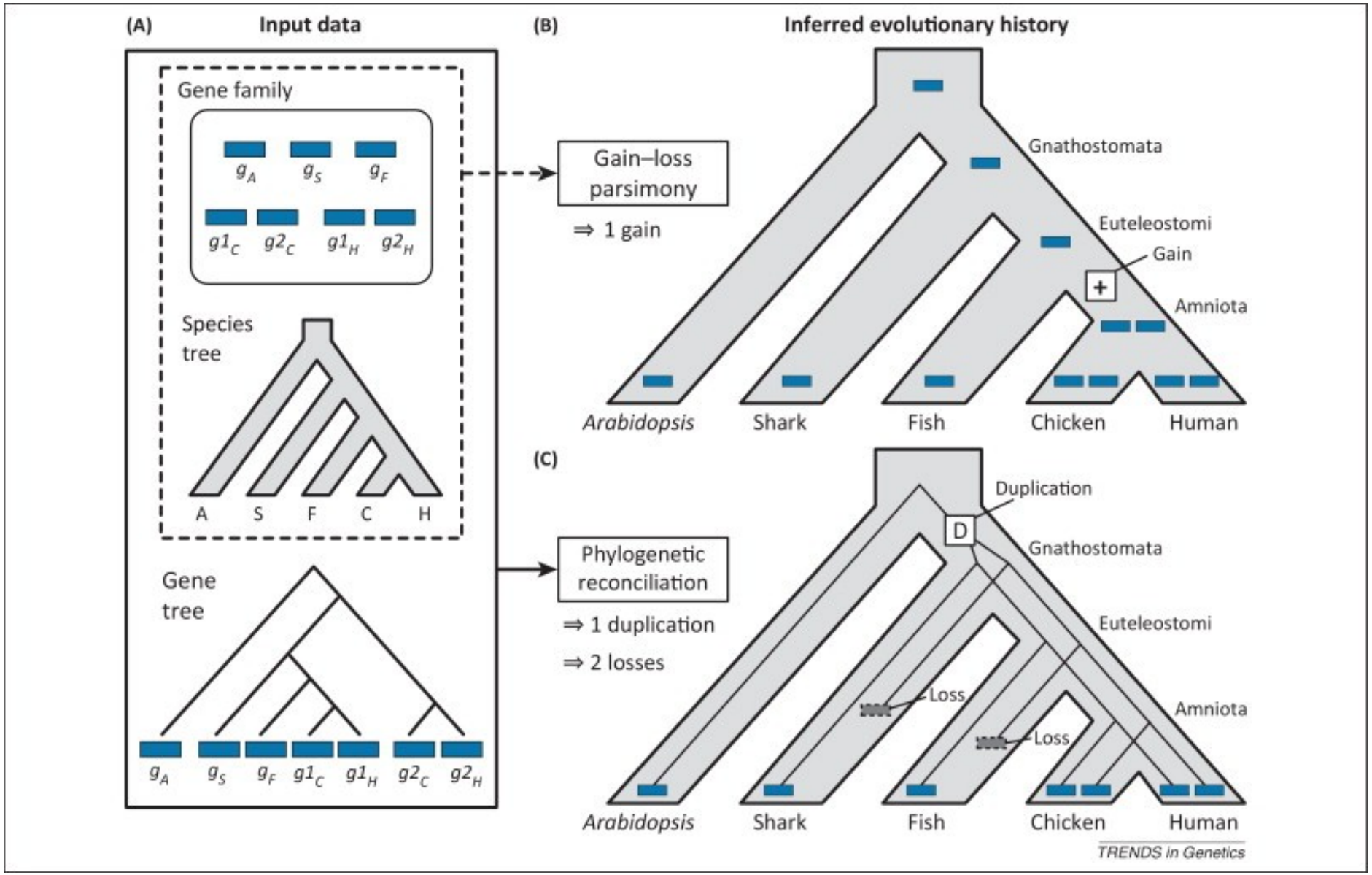
# Properties of NJ algorithm

- The same complexity as average linkage hierarchical clustering $\mathcal{O}(n^3)$
- Guaranteed to return the correct answer if the distance matrix $D$ originates from a tree
- Works also for non-ultrametric trees

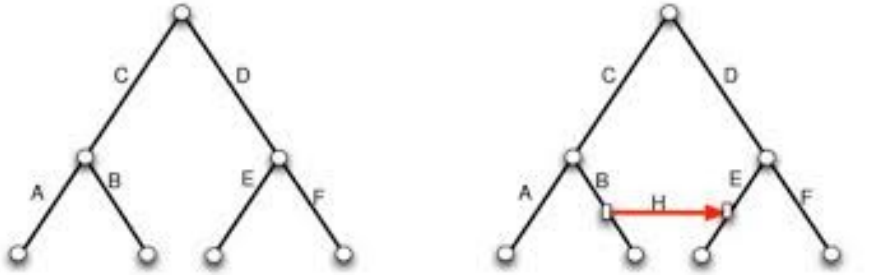
# Further tree-related problems

- Gene-species tree reconciliation
- Tree refinement
- Horizontal gene transfer - Phylogenetic networks
- Comparison of large trees
- Optimality measures for phylogenetic trees
- True Ancestral sequence reconstruction
- Etc...

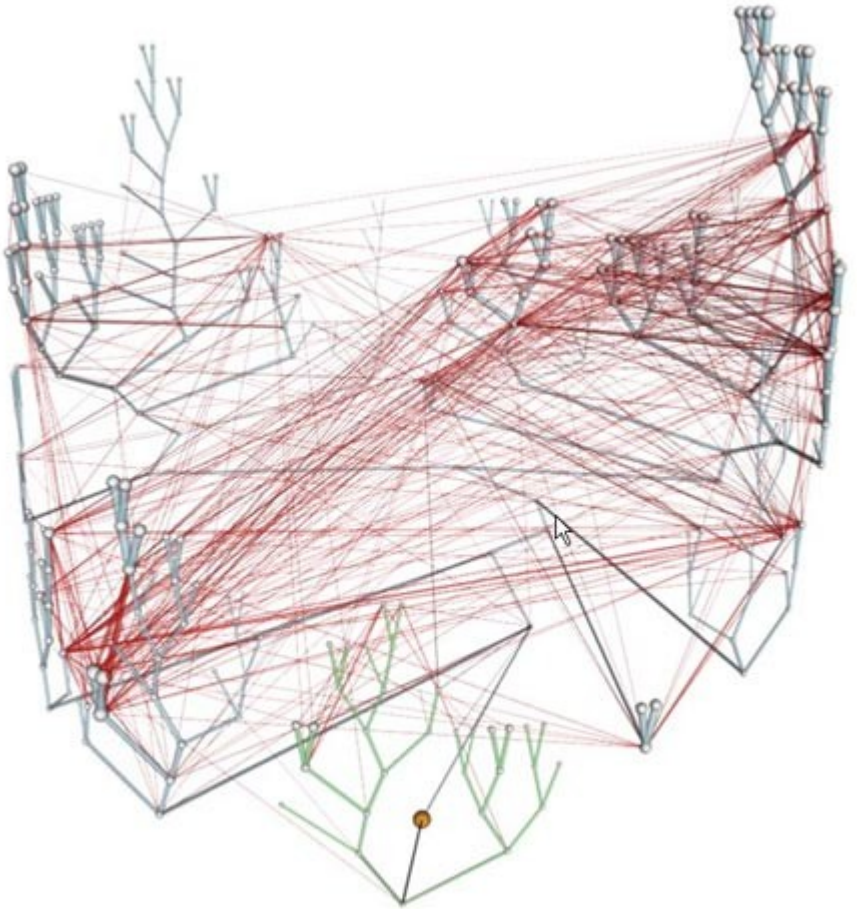
# Gene- species-tree reconciliation



# Horizontal gene transfer




(a) Phylogenetic tree (b) Phylogenetic network



A bird's-eye view of the tree of life, showing the vines in red and the tree's branches in grey [Bacteria] and green [Archaea]. The last universal common ancestor is shown as a yellow sphere.

# Now back to multiple alignments

- Theoretically, we could try to prove that P=NP, and then solve MSA
- In practice, we are not (usually) making multiple alignments of random sequences. Usually we know they are related
- Can we use the knowledge that they originated from an evolutionary process to guide our search for optimal MSA?

# Feng-Doolitle approach

- We can use the greedy approach similar to UPGMA
- In each step we choose the nearest pair (using pairwise sequence distances)
- And we can merge alignments based on the pairwise alignment between the sequences
- Uses the principle of "once a gap, always a gap".
- *We need to "elegantly" align alignments of more than one sequence*

# Score for profile alignment

In the process of incremental alignment we need to align profiles (alignments)

$$\sum_i S(m_i) = \sum_i \sum_{k<l\leq N} s(m_i^k, m_i^l)$$

$$= \sum_i \sum_{k<l\leq n} s(m_i^k, m_i^l) + \sum_i \sum_{n<k<l\leq N} s(m_i^k, m_i^l) + \sum_i \sum_{k\leq n, n<l\leq N} s(m_i^k, m_i^l).$$

image (c) Durbin et al

# A first proper approach - CLUSTALW

**Algorithm:** CLUSTALW **progressive alignment**

(i) Construct a distance matrix of all $N(N-1)/2$ pairs by pairwise dynamic programming alignment followed by approximate conversion of similarity scores to evolutionary distances using the model of Kimura [1983].

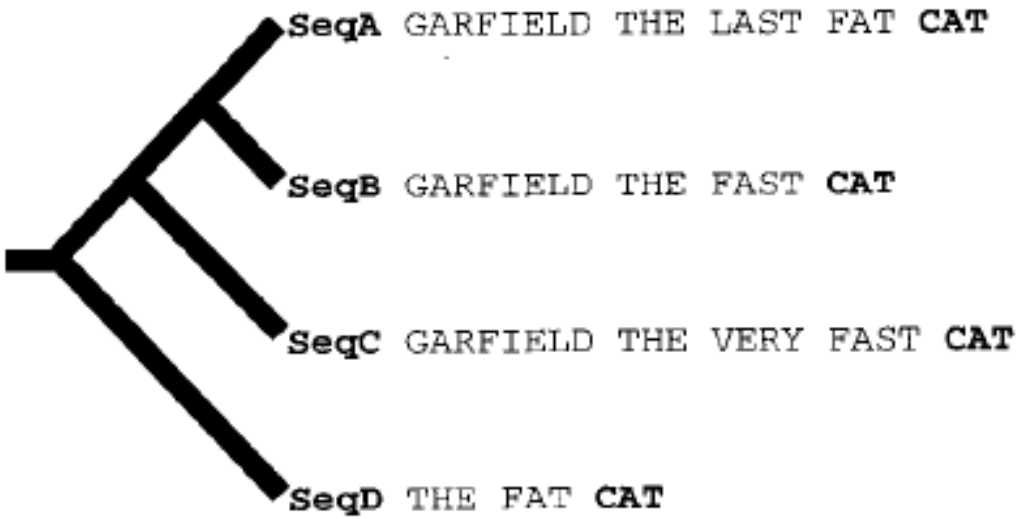
(ii) Construct a guide tree by a neighbour-joining clustering algorithm by Saitou & Nei [1987].

(iii) Progressively align at nodes in order of decreasing similarity, using sequence–sequence, sequence–profile, and profile–profile alignment. ◁

image (c) Durbin et al

# Practical issues with the simple incremental approach

a)Regular Progressive Alignment Strategy



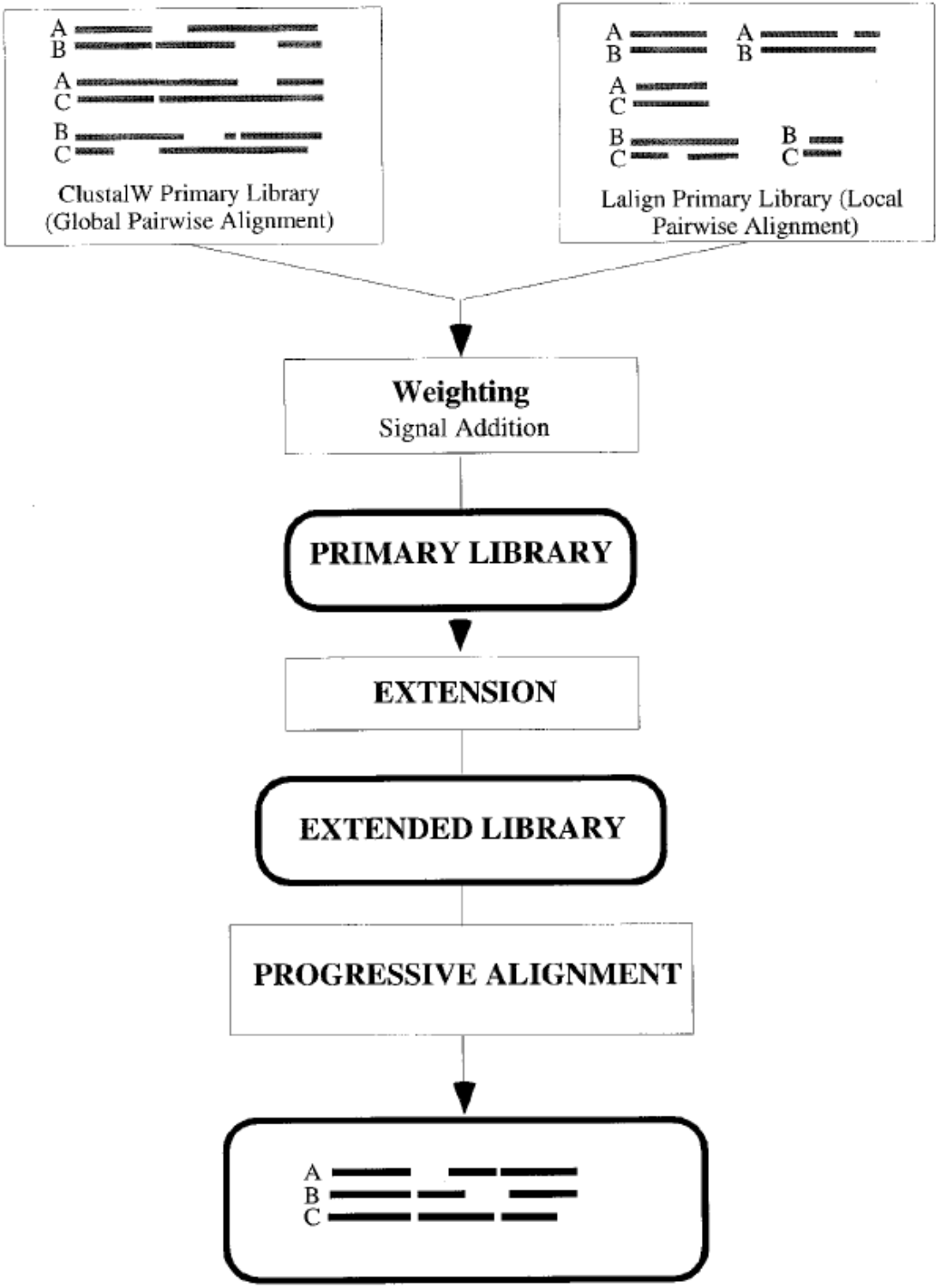

```
SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE FAST CA-T ---
SeqC GARFIELD THE VERY FAST CAT
SeqD -------  THE ---- FA-T CAT
```

# T-Coffee algorithm (Notredamme 2000)

Create one library of global pairwise alignments

And one library of local pairwise alignments

Use the signals in both for imptrovement of the progressive alignment



# T-Coffee in action

b)Primary Library

```
SeqA GARFIELD THE LAST FAT CAT
SeqB GARFIELD THE FAST CAT ---
```
Prim. Weight = 88

```
SeqA GARFIELD THE LAST FA-T CAT
SeqC GARFIELD THE VERY FAST CAT
```
Prim. Weight = 77

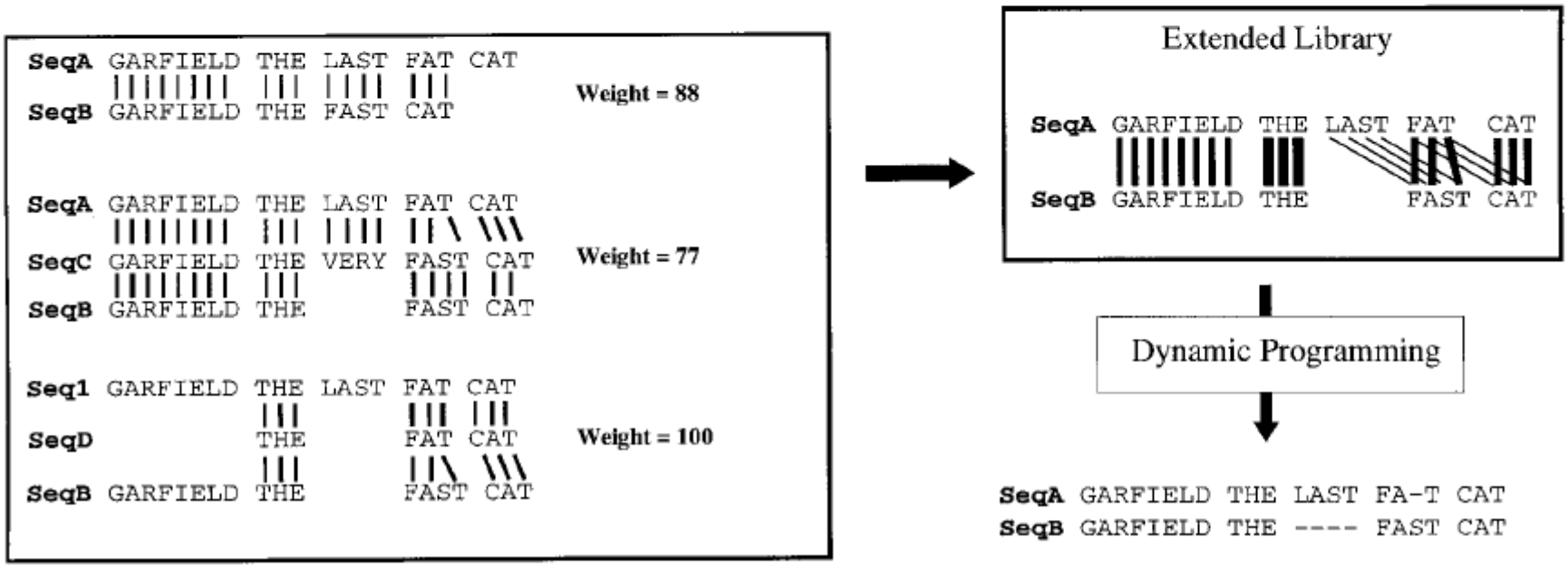
```
SeqA GARFIELD THE LAST FAT CAT
SeqD ------- THE ---- FAT CAT
```
Prim. Weight =100

```
SeqB GARFIELD THE ---- FAST CAT
SeqC GARFIELD THE VERY FAST CAT
```
Prim Weight = 100

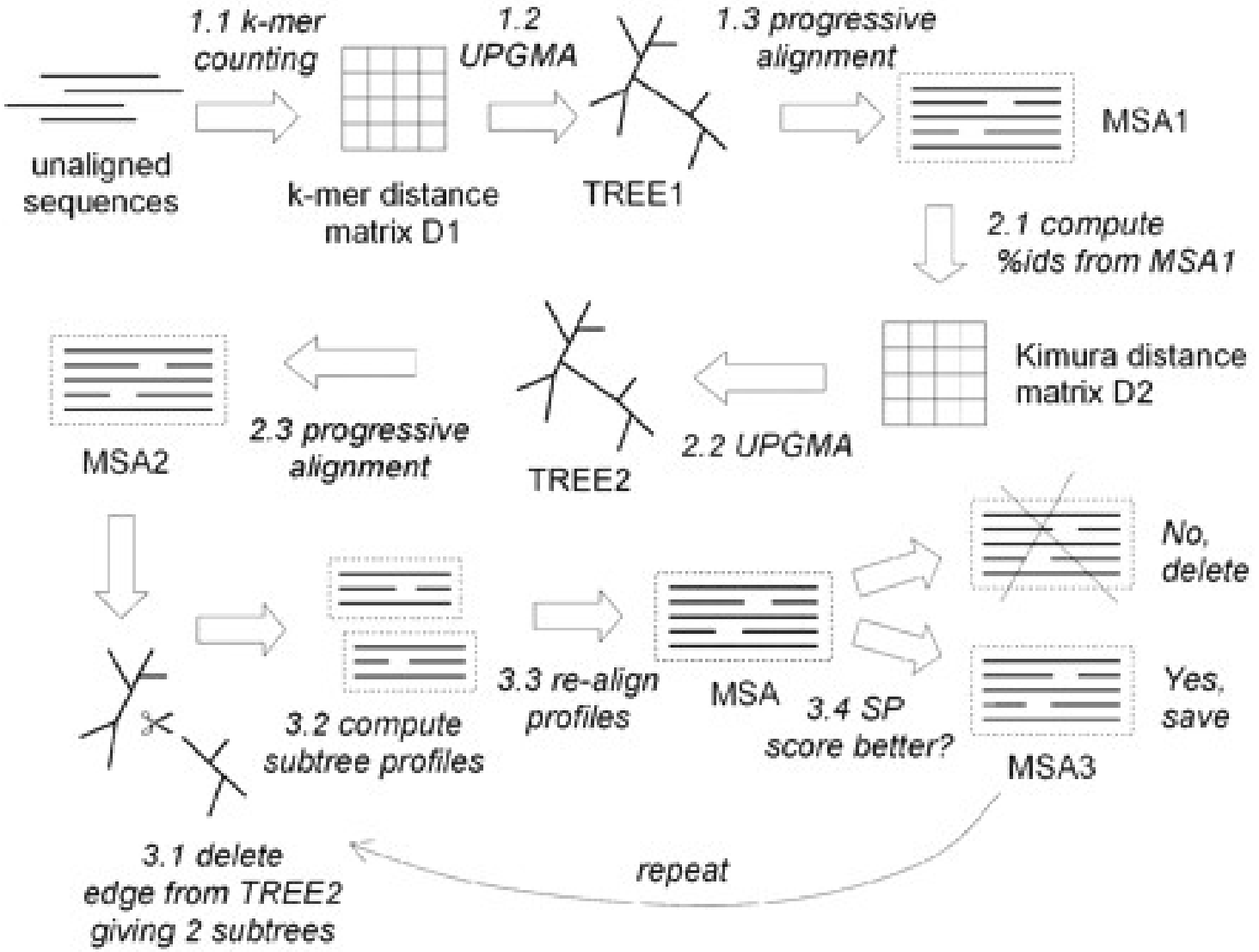
```
SeqB GARFIELD THE FAST CAT
SeqD ------- THE FA-T CAT
```
Prim. Weight = 100

```
SeqC GARFIELD THE VERY FAST CAT
SeqD ------- THE ---- FA-T CAT
```
Prim. Weight = 100

c)Extended Library for seq1 and seq2



# Muscle method (Edgar 2004)



# Books to read more