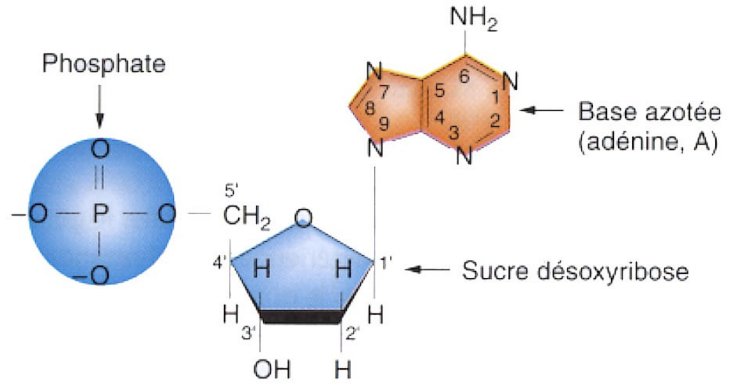
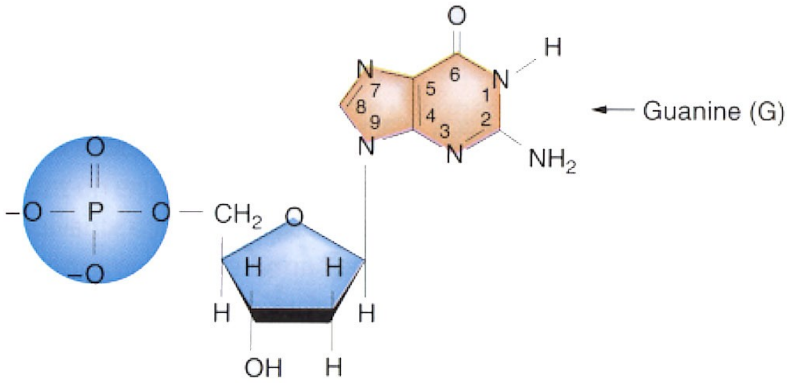
La théorie synthétique de l'évolution

La théorie synthétique de l'évolution

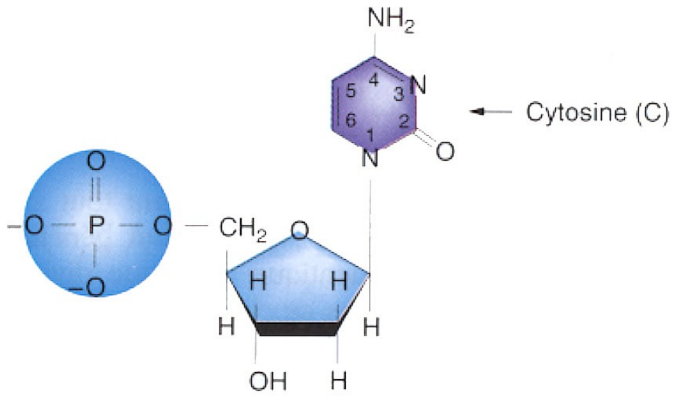
1. L'ADN est le support de l'information génétique



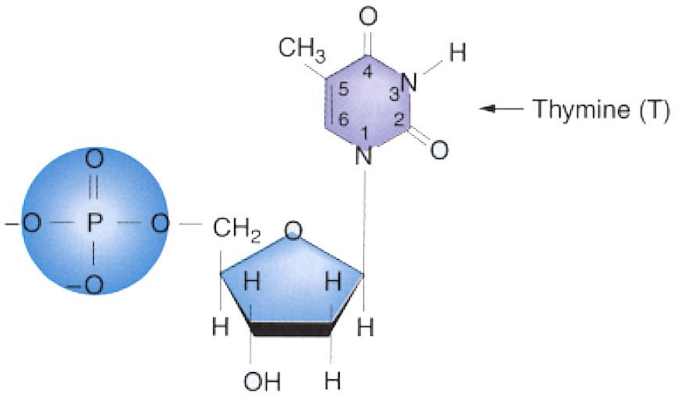
Désoxyadénosine 5'-phosphate (dAMP)



Désoxyguanosine 5'-phosphate (dGMP)



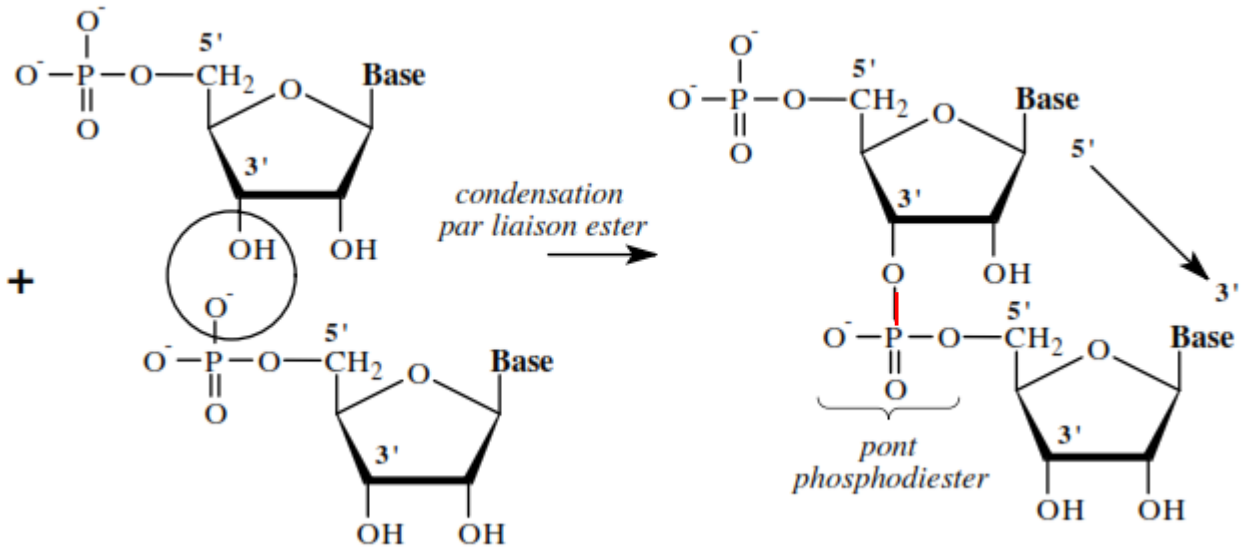
Désoxycytidine 5'-phosphate (dCMP)



Désoxythymidine 5'-phosphate (dTMP)

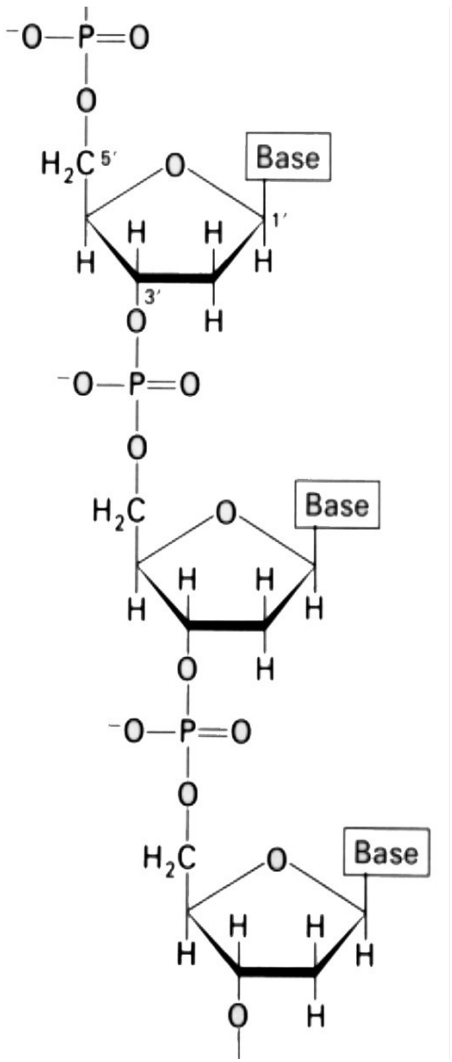
La théorie synthétique de l'évolution

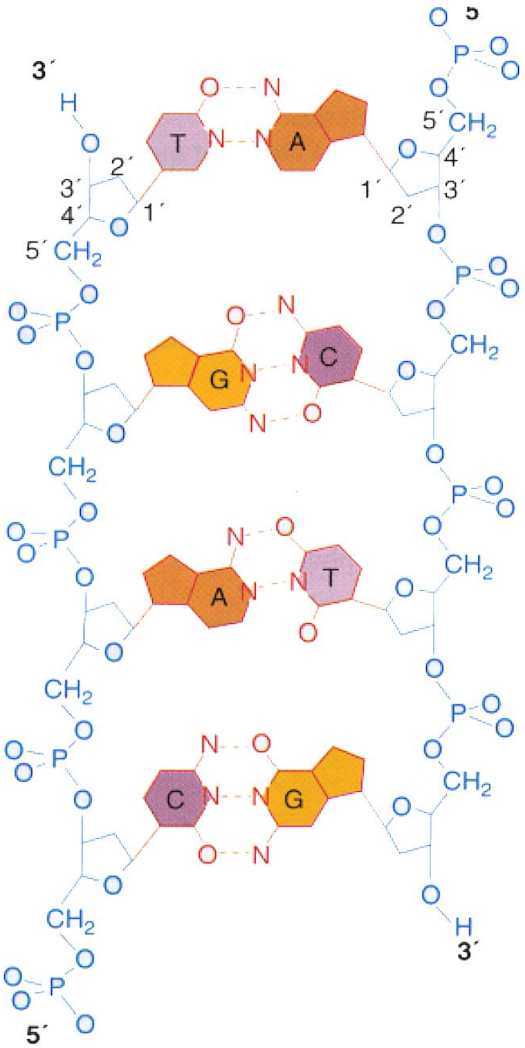
1. L'ADN est le support de l'information génétique

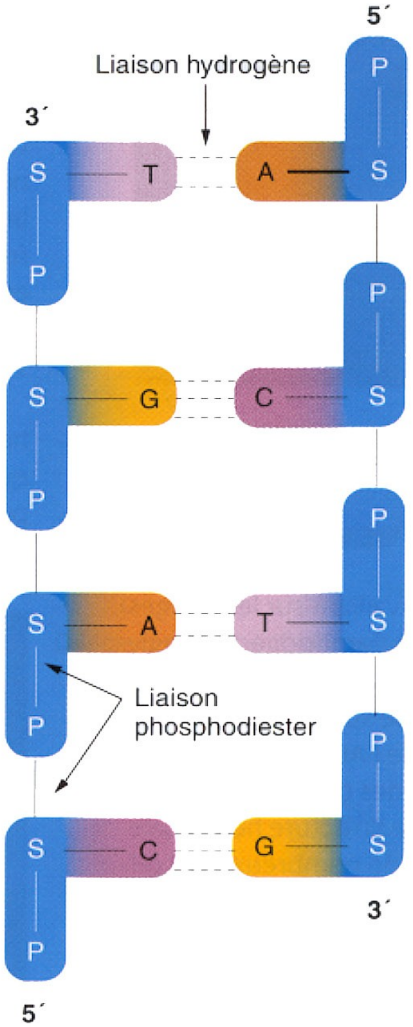


La théorie synthétique de l'évolution

1. L'ADN est le support de l'information génétique







La théorie synthétique de l'évolution

1. L'ADN est le support de l'information génétique

2. L'ADN peut subir des modifications (mutations)

La théorie synthétique de l'évolution

1. L'ADN est le support de l'information génétique

2. L'ADN peut subir des modifications (mutations)

3. Les mutations peuvent être héritables.

La théorie synthétique de l'évolution

1. L'ADN est le support de l'information génétique

2. L'ADN peut subir des modifications (mutations)

3. Les mutations peuvent être héritables.

4. En fonction de leur patrimoine génétique (Génotype) les individus d'une même espèce vont réagir différemment à leur environnement (Phénotype).

50% de diabétiques, dont ½ est obèse

Moins de 10% de diabétiques, aucun obèse

La théorie synthétique de l'évolution

1. L'ADN est le support de l'information génétique

2. L'ADN peut subir des modifications (mutations)

3. Les mutations peuvent être héritables.

4. En fonction de leur patrimoine génétique (Génotype) les individus d'une même espèce vont réagir différemment à leur environnement (Phénotype).

5. Dans un environnement donné, seuls les individus les mieux adaptés à cet environnement surivivent et se reproduisent (Sélection naturelle).


Biston betularia

% Relachés			% Recapturés	
Noirs	Blancs		Noirs	Blancs
50	50	Troncs clairs	1	15
50	50	Troncs bruns	25	12

D’après Kettlewell, 1955

La théorie synthétique de l'évolution

1. L'ADN est le support de l'information génétique

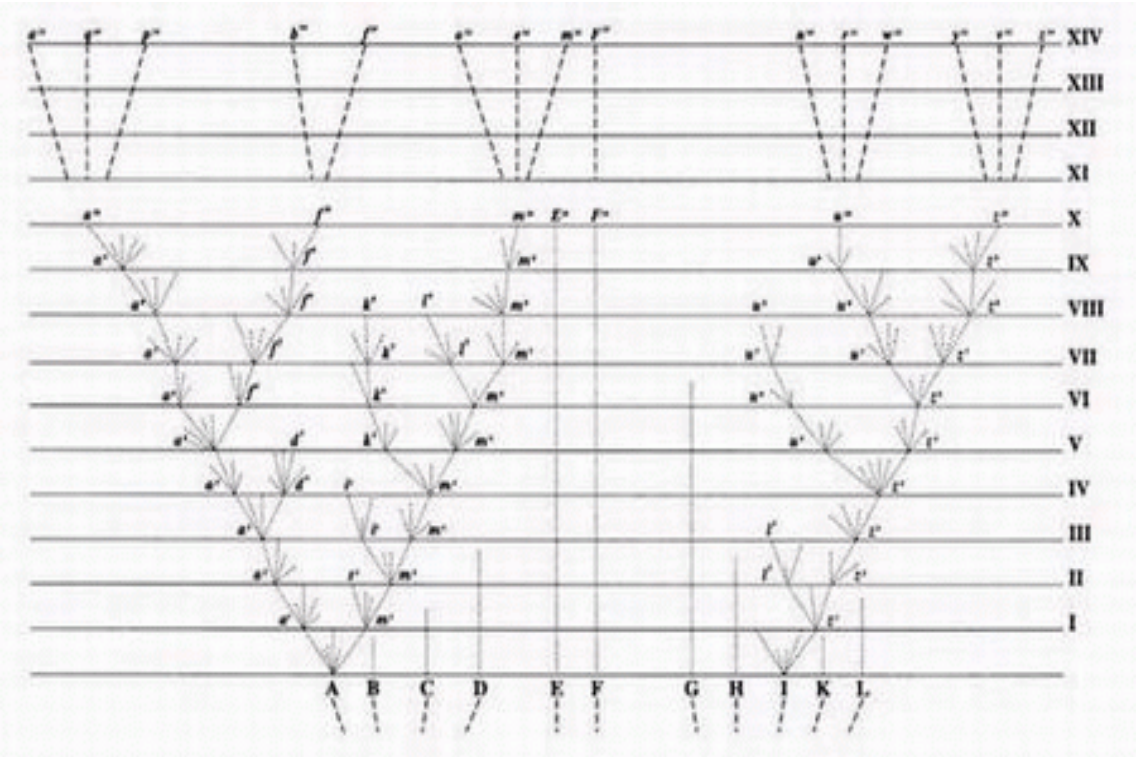
2. L'ADN peut subir des modifications (mutations)

3. Les mutations peuvent être héritables.

4. En fonction de leur patrimoine génétique (Génotype) les individus d'une même espèce vont réagir différemment à leur environnement (Phénotype).

5. Dans un environnement donné, seuls les individus les mieux adaptés à cet environnement surivivent et se reproduisent (Sélection naturelle).

6. L'accumulation d'un grand nombre de mutations au cours du temps, la sélection naturelle ainsi que la dérive génétique contribuent à l'apparition d'espèces nouvelles (Evolution)



"The affinities of all the beings of the same class have sometimes been represented by a great tree... As buds give rise by growth to fresh buds, and these if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications."

Charles Darwin, 1859

“Molecules as Documents of Evolutionary History”

Zuckerkandl and Pauling (1965)

Collecting Sequences in the Laboratory

- Protein-sequencing methods: 1951

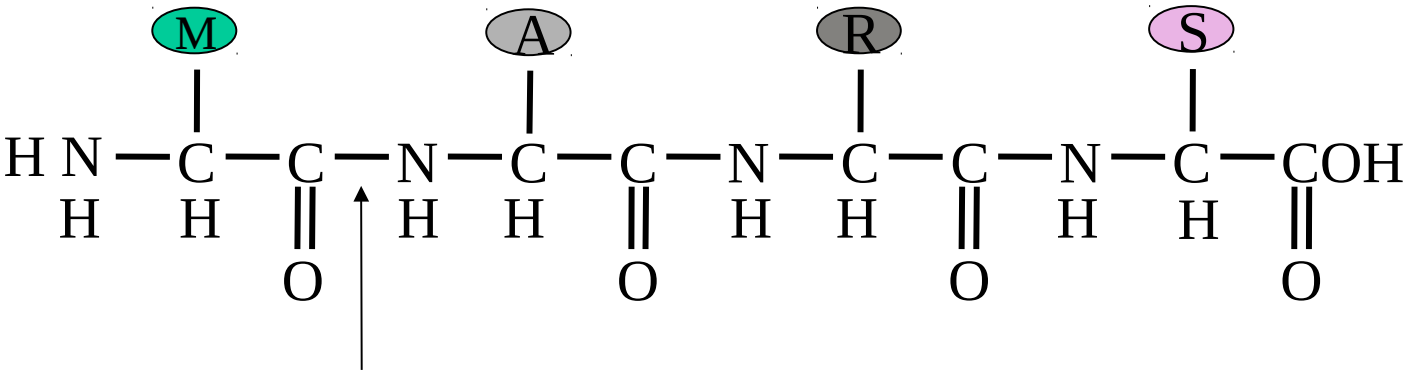
Sanger F*. and Tuppy H.

The amino acid sequences of the phenylalanyl chain of insulin. (1951) *Biochem. J.* **49**:481-490

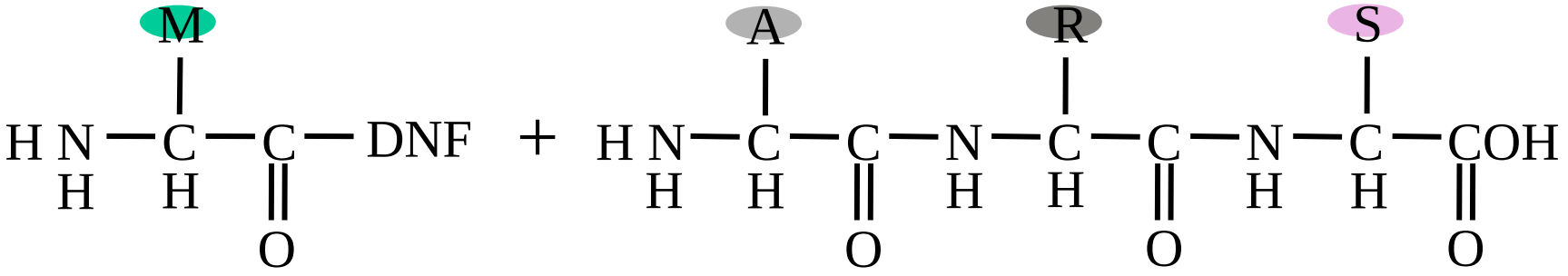
* *Nobel Prize in Chemistry 1958*

The amino acid sequences of the phenylalanyl chain of insulin

Sanger F. and Tuppy H. (1951)



2,4 DinitroFluoroBenzene (DNF)



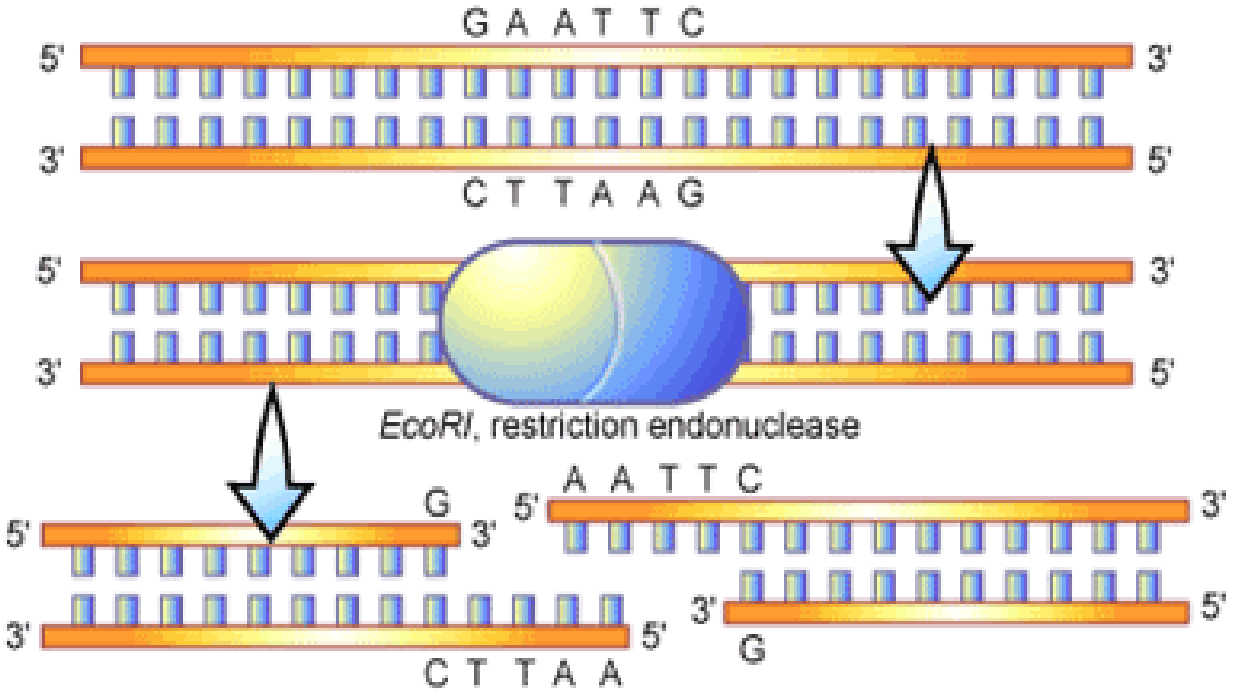
Collecting Sequences in the Laboratory

- DNA-sequencing methods: 1977

Maxam A.M. and Gilbert W.*
A new method for sequencing DNA.
(1977) *Proc. Natl. Acad. Sci. USA.* **74**:560-564

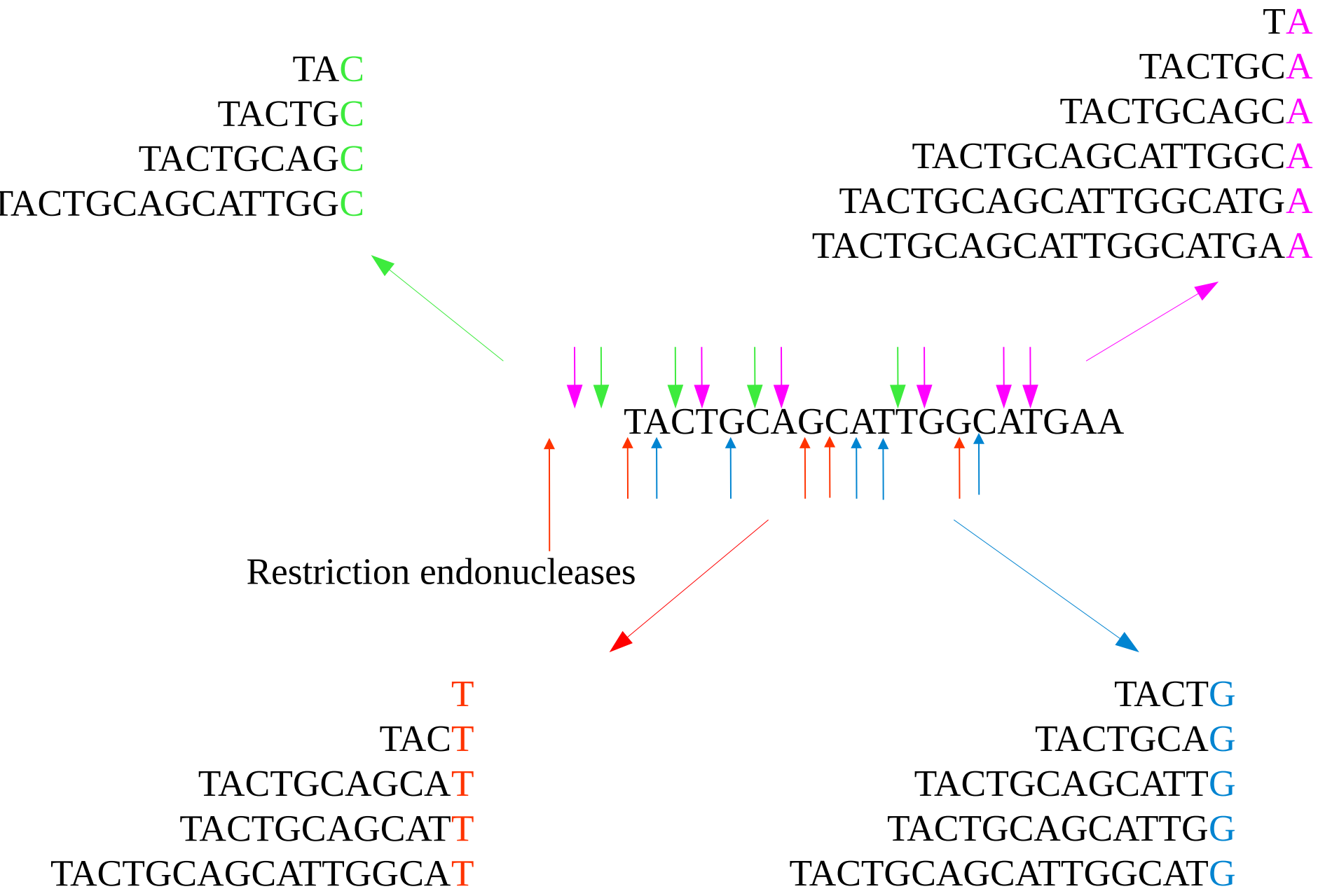
Sanger F.*, Nicklen S., and Coulson A.R.
DNA sequencing with chain terminating inhibitors.
(1977) *Proc. Natl. Acad. Sci. USA.* **74**:5463-5467

** ¼ Nobel Prize in Chemistry 1980*

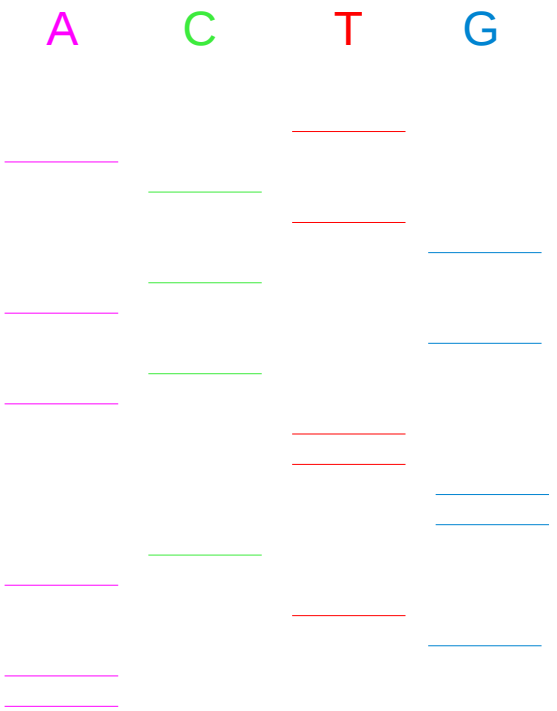

G A A T T C
5'
3'
3'
5'
C T T A A G
5'
3'
3'
5'
EcoRI, restriction endonuclease
G
5'
3'
3'
5'
C T T A A
A A T T C
5'
3'
3'
5'
G

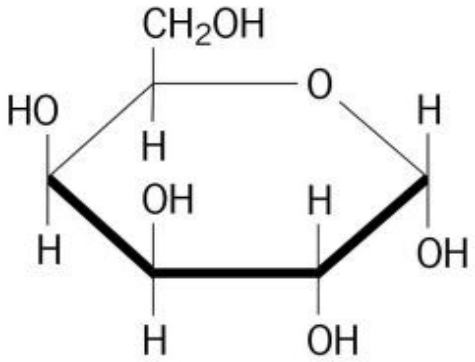
A new method for sequencing DNA

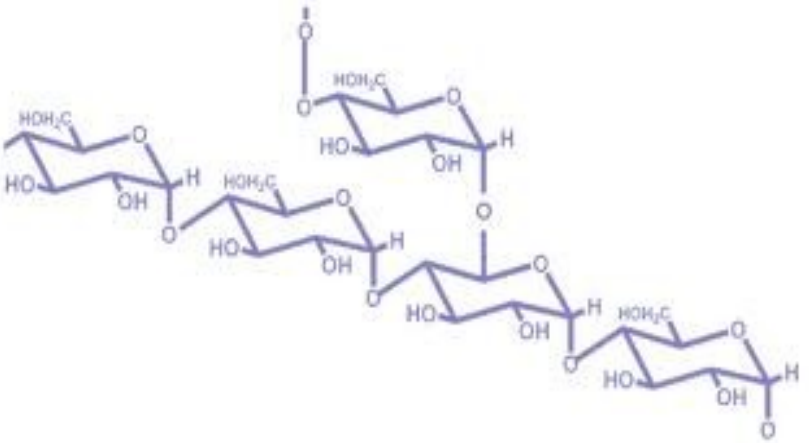
Maxam A.M. and Gilbert W. (1977)

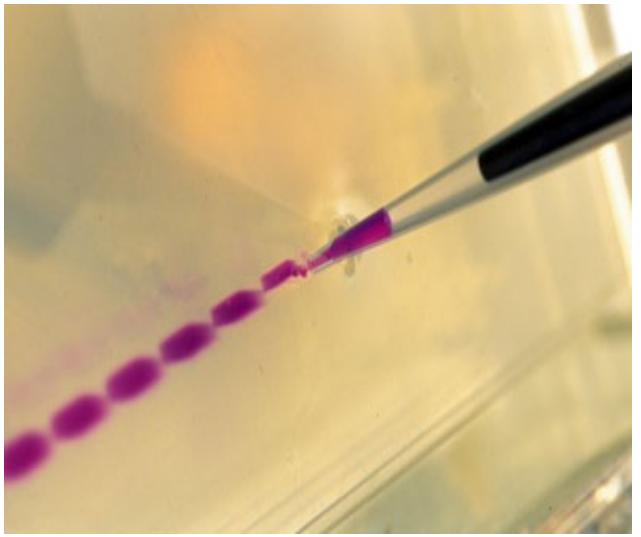


TA
TACTGCA
TACTGCAGCA
TACTGCAGCATTGGCA
TACTGCAGCATTGGCATGA
TACTGCAGCATTGGCATGAA
TAC
TACTGC
TACTGCAGC
TACTGCAGCATTGGC
TACTG
TACTGCAG
TACTGCAGCATTG
TACTGCAGCATTGG
TACTGCAGCATTGGCATG
T
TACT
TACTGCAGCAT
TACTGCAGCATT
TACTGCAGCATTGGCAT

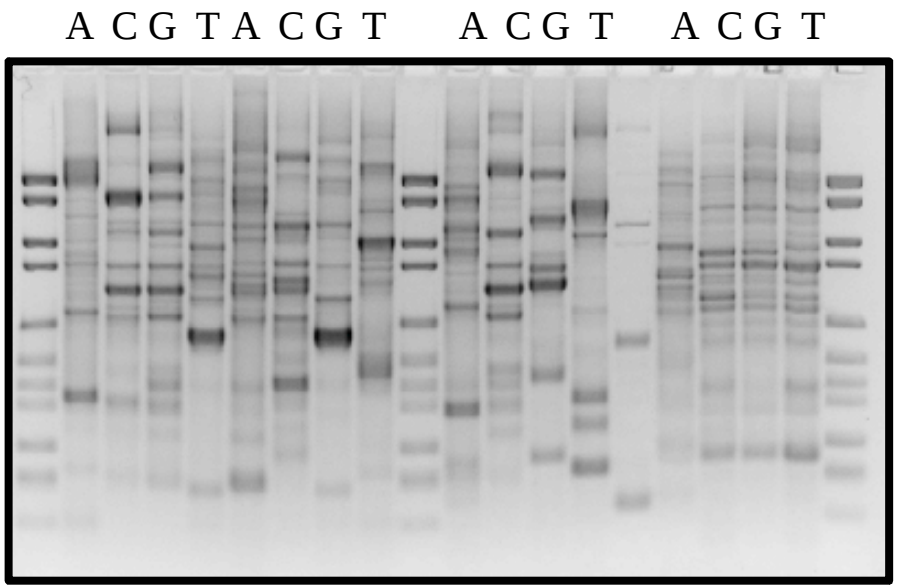



CH2OH
HO
O
H
H
OH
H
H
OH
H
OH



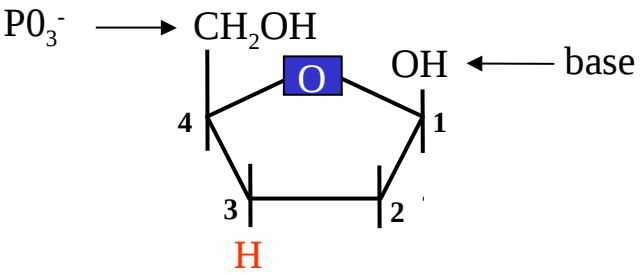



Agarose 1%
Minnie·the·Gel Cicle
SUBMARINE AGAROSE GEL UNIT
HE33
Hoefer
WARNING

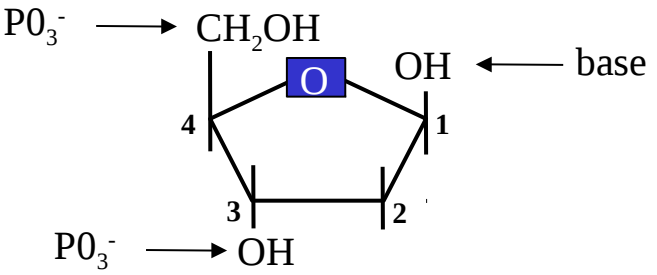

A C G T A C G T
A C G T
A C G T

DNA sequencing with chain terminating inhibitors

Sanger F., Nicklen S., and Coulson A.R. (1977)



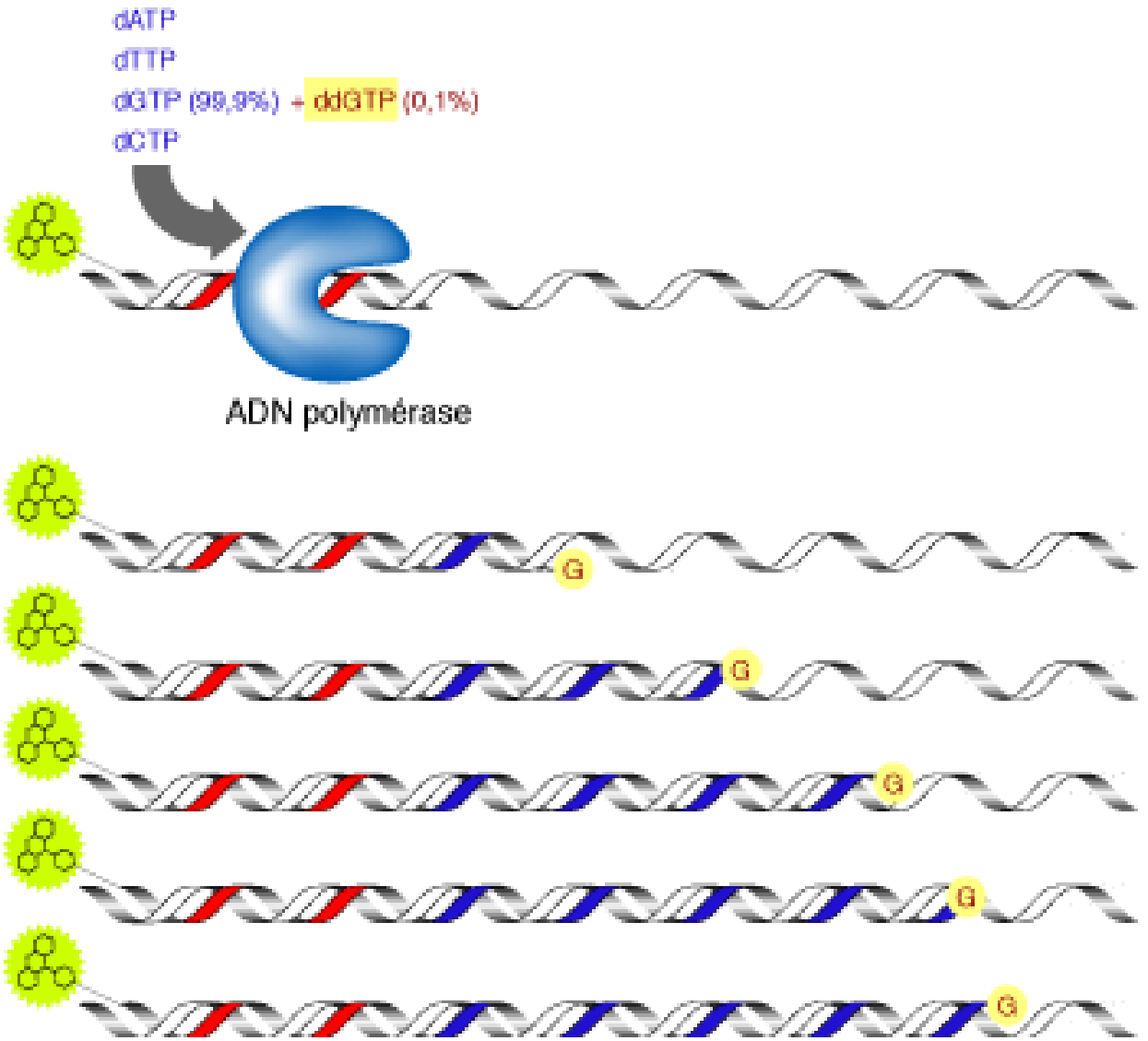
ddNTP

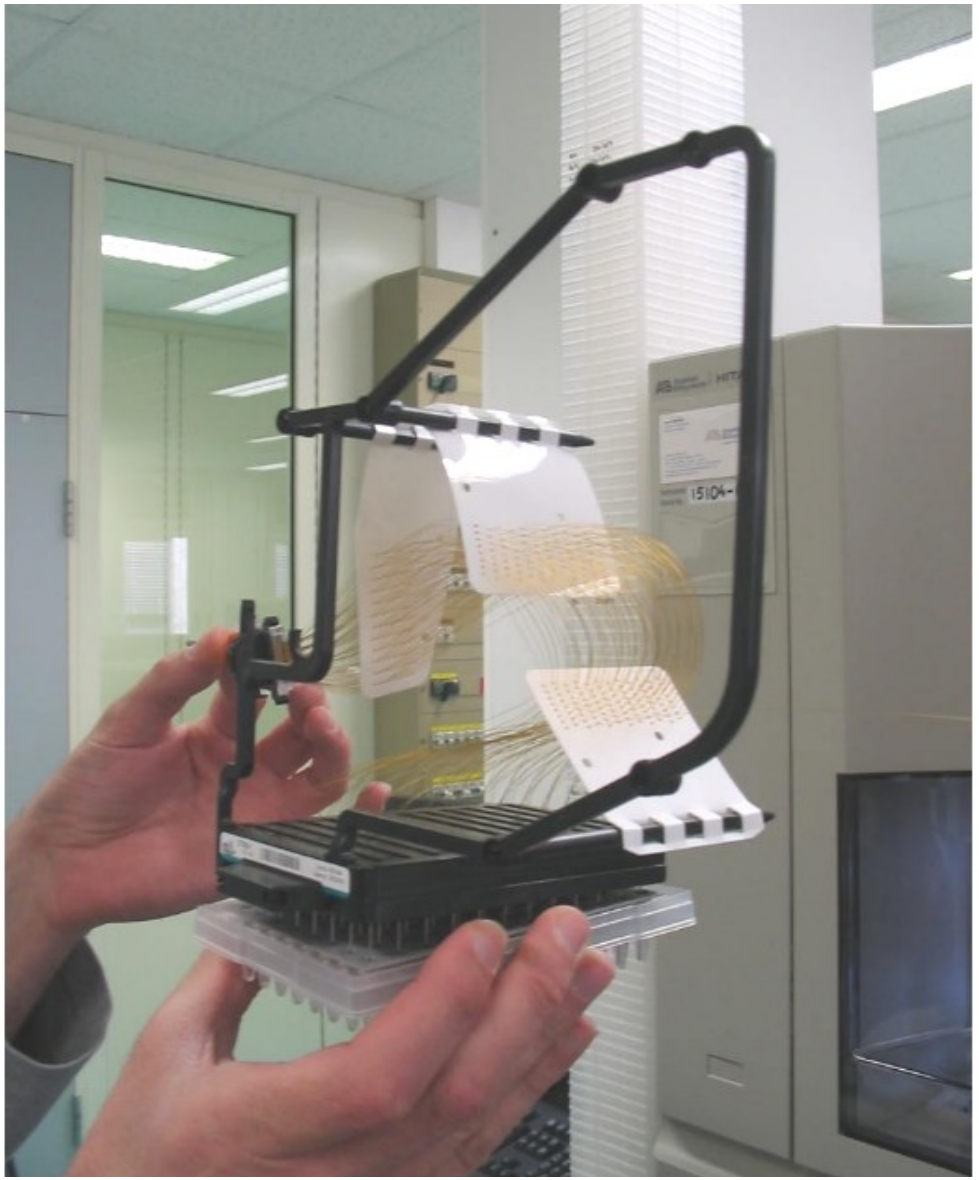


dNTP

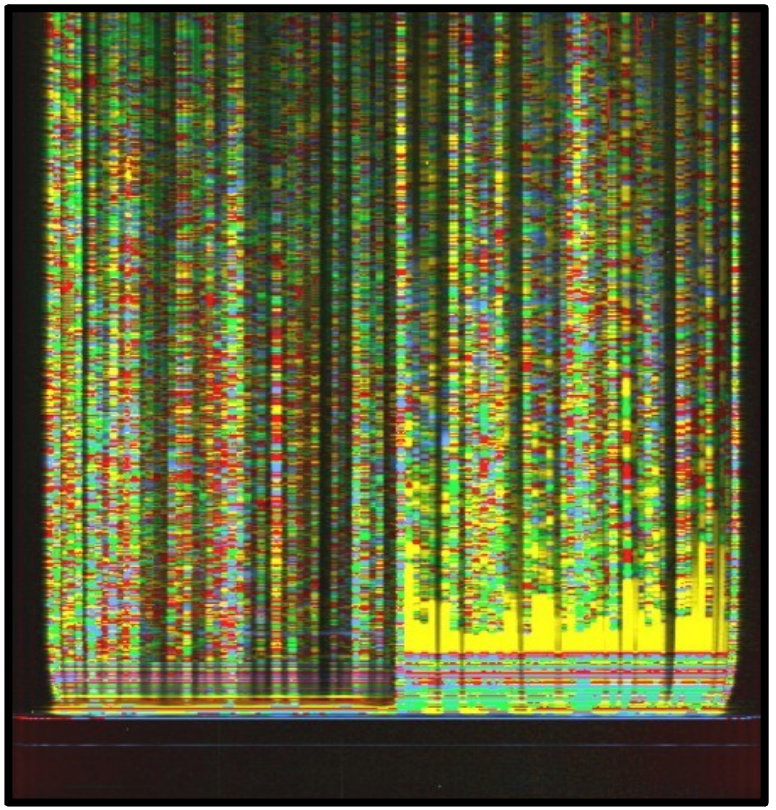
TAGCATGATGCGGTCCAA

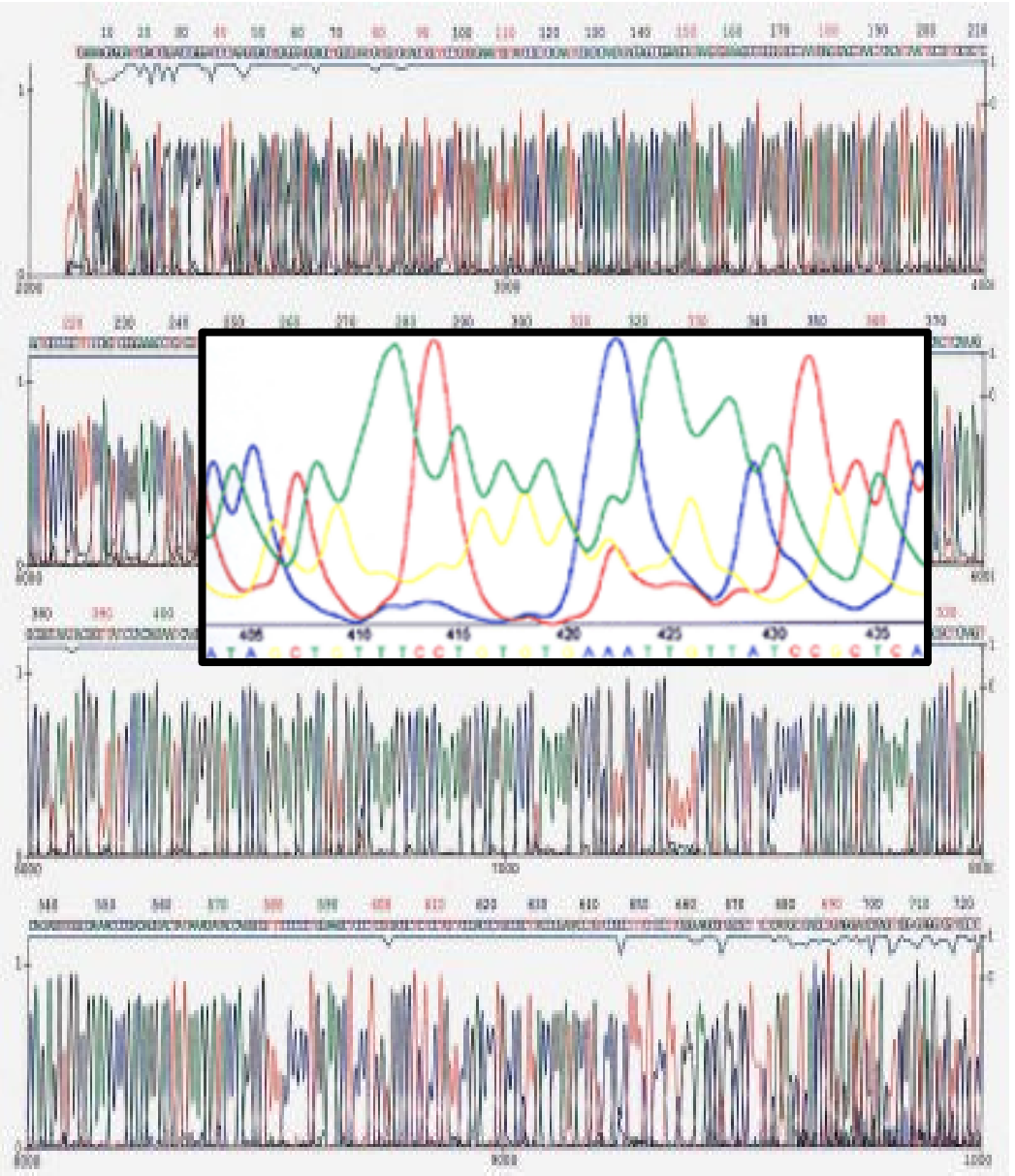
TAGCATGATGCGGTCCAA
TAGCATGATGCGGTCCA
TAGCATGA
TAGCA
TA


dATP
dTTP
dGTP (99,9%) + ddGTP (0,1%)
dCTP
ADN polymérase
G
G
G
G
G

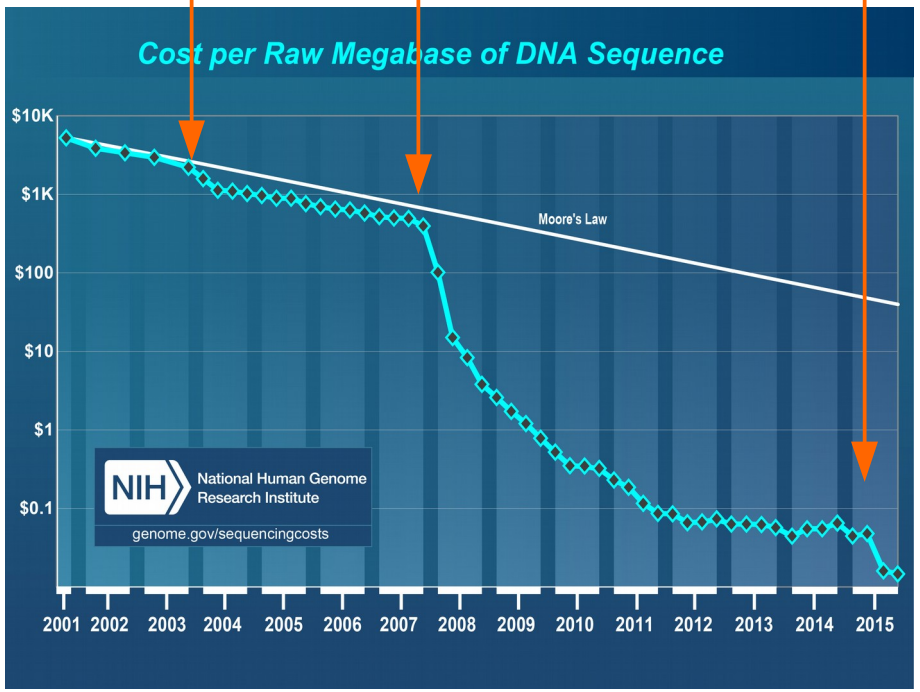


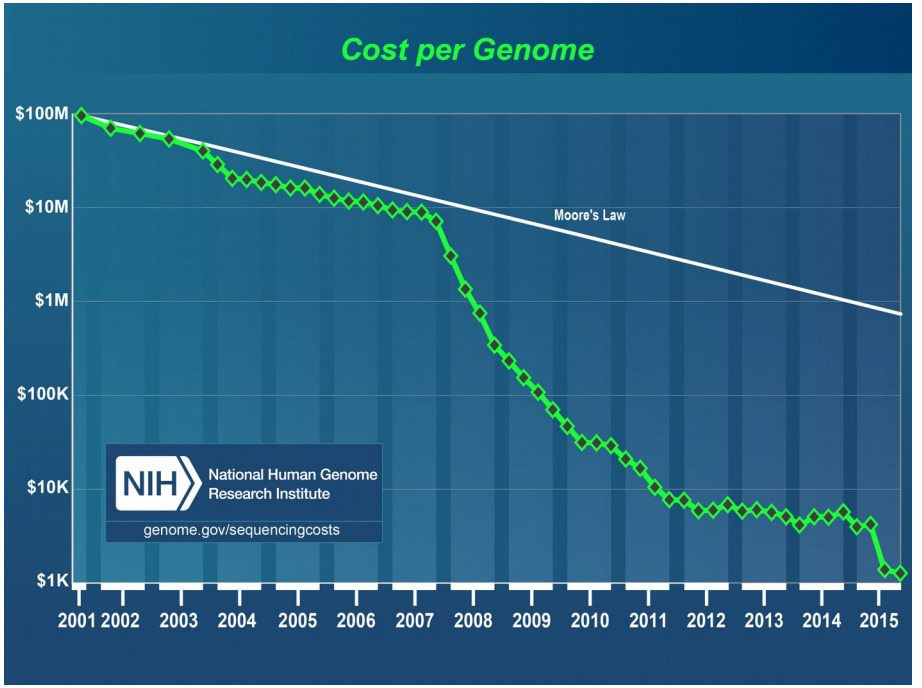

3100 Genetic Analyzer



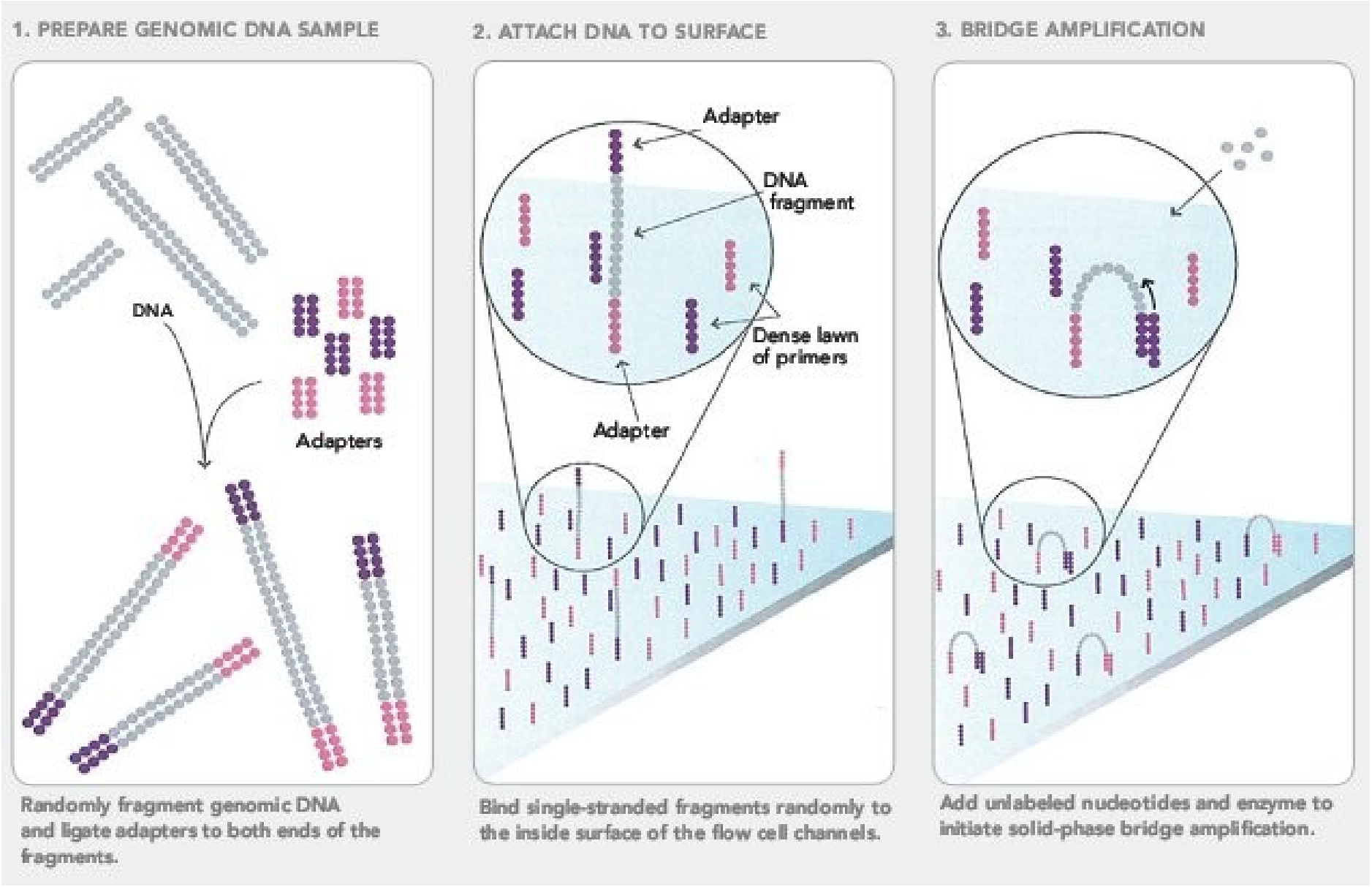


454 Illumina PacBio

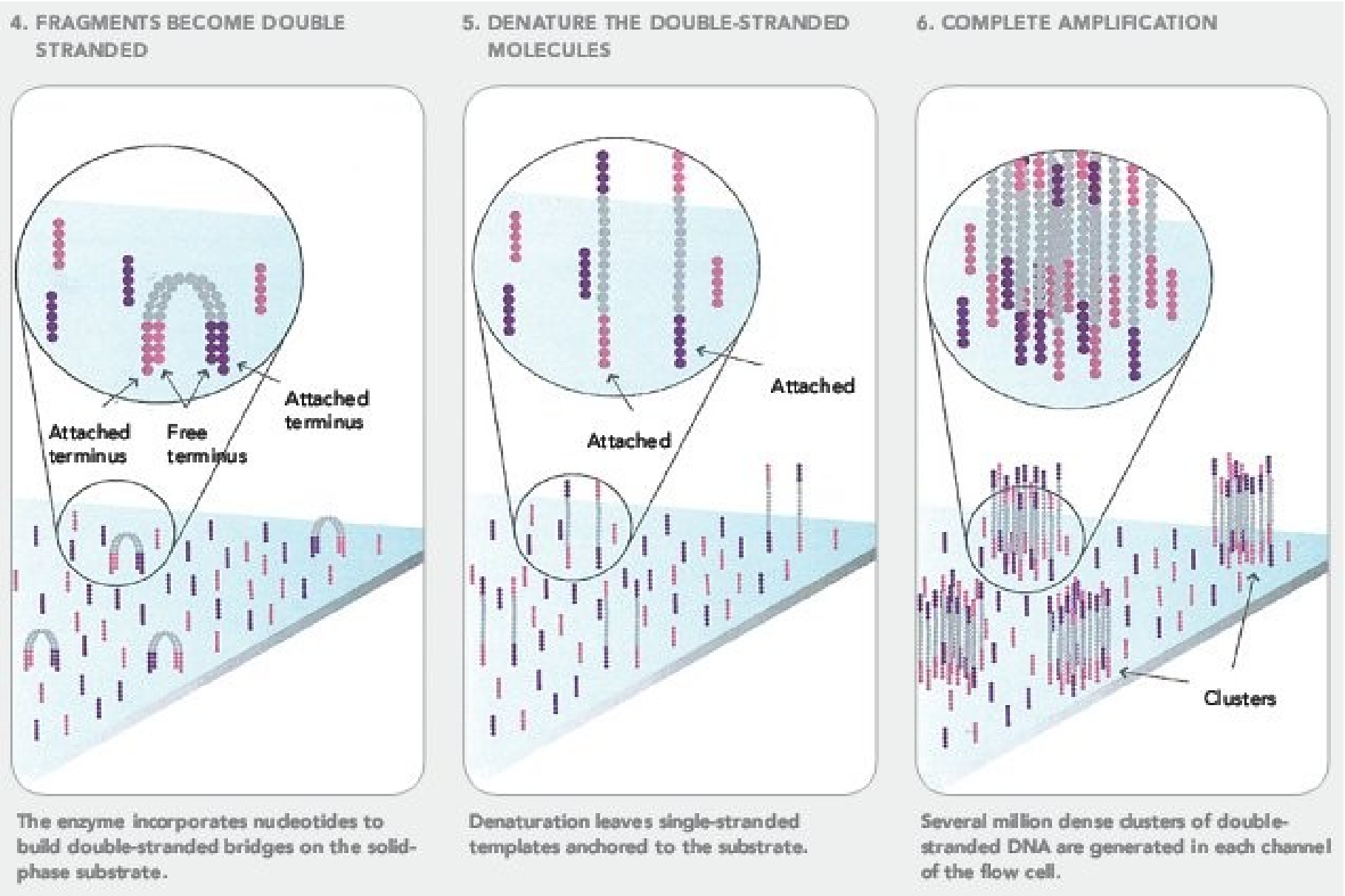




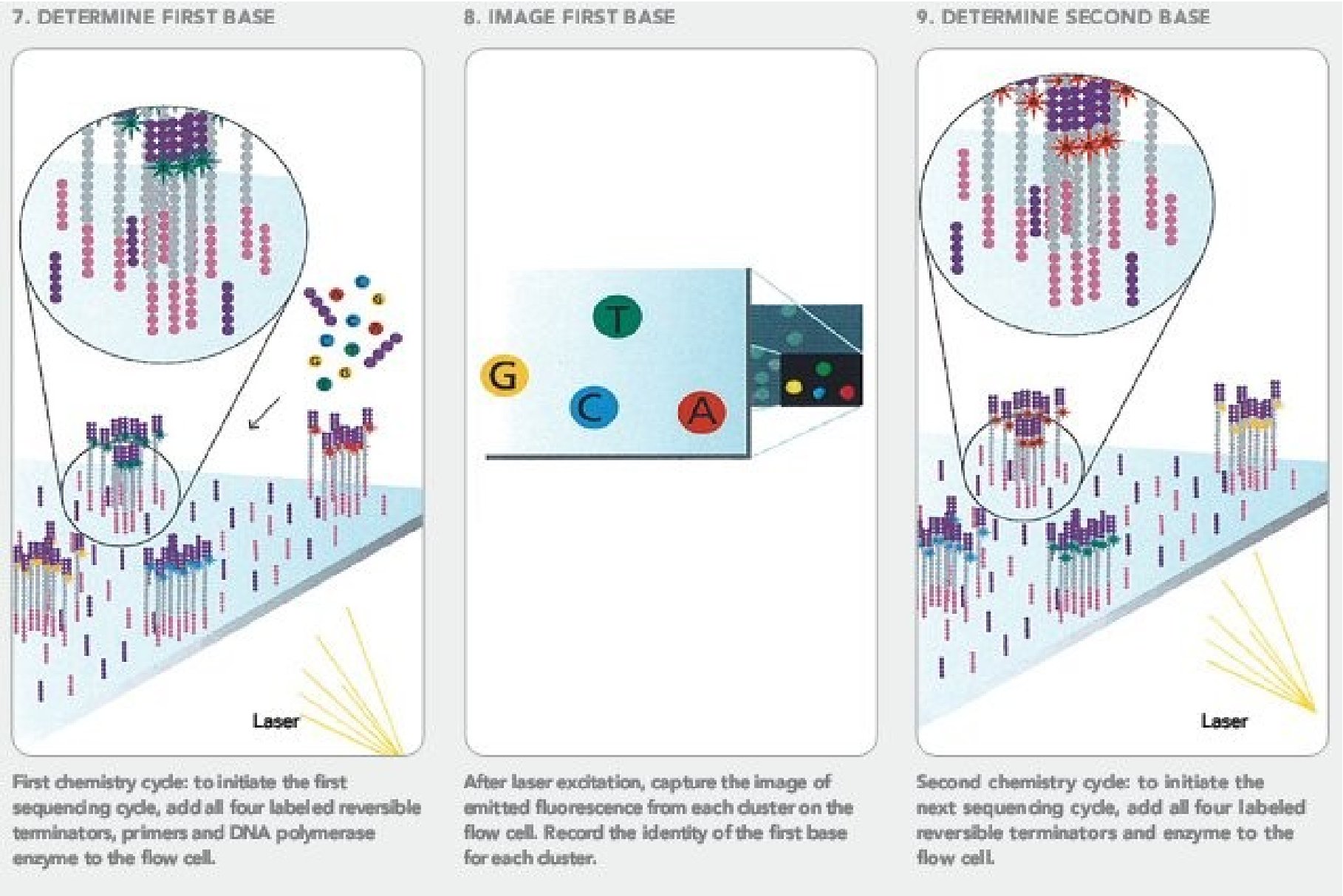
Le séquençage des acides nucléiques (Haut débit Illumina)



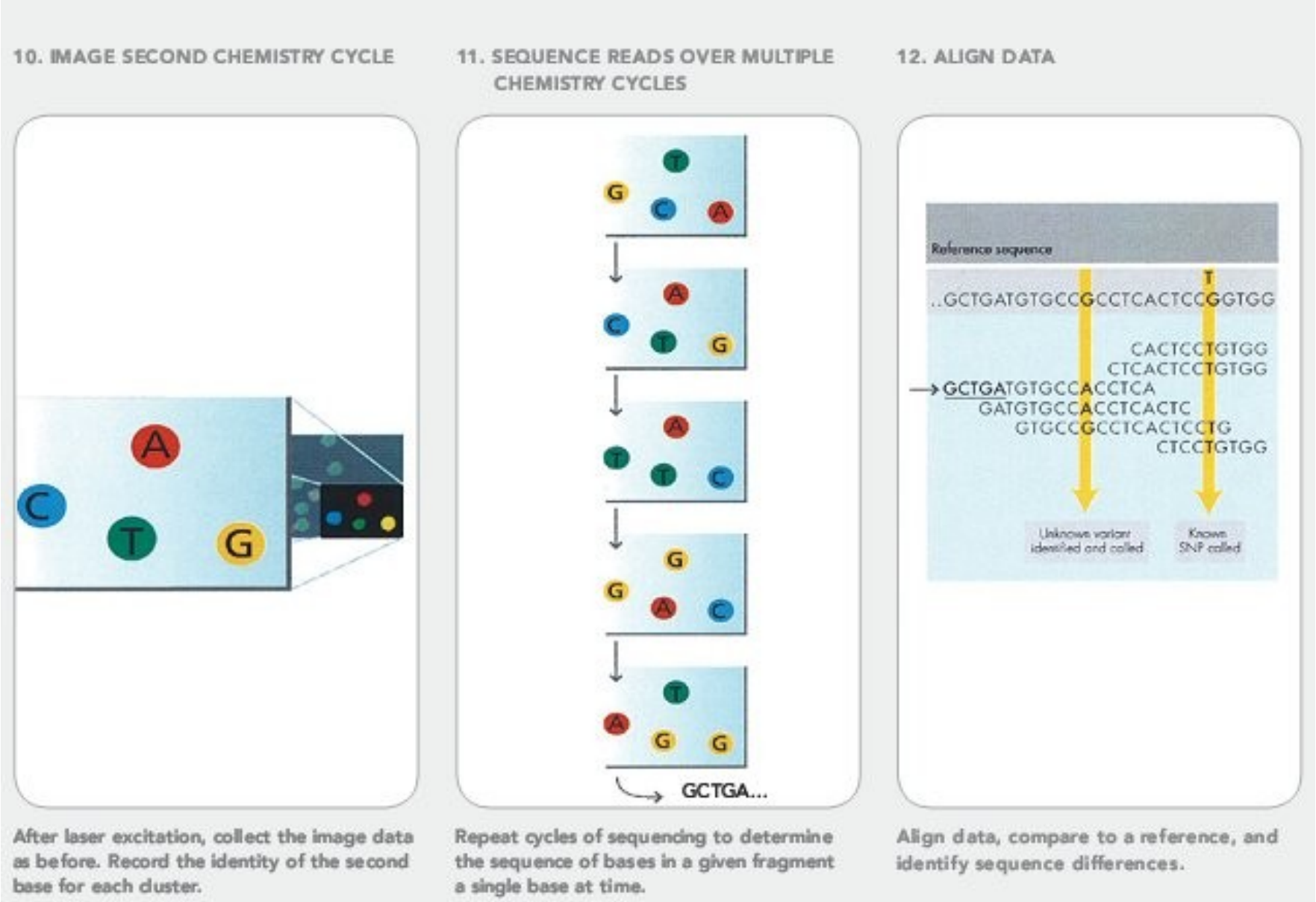
Le séquençage des acides nucléiques (Haut débit Illumina)



Le séquençage des acides nucléiques (Haut débit Illumina)

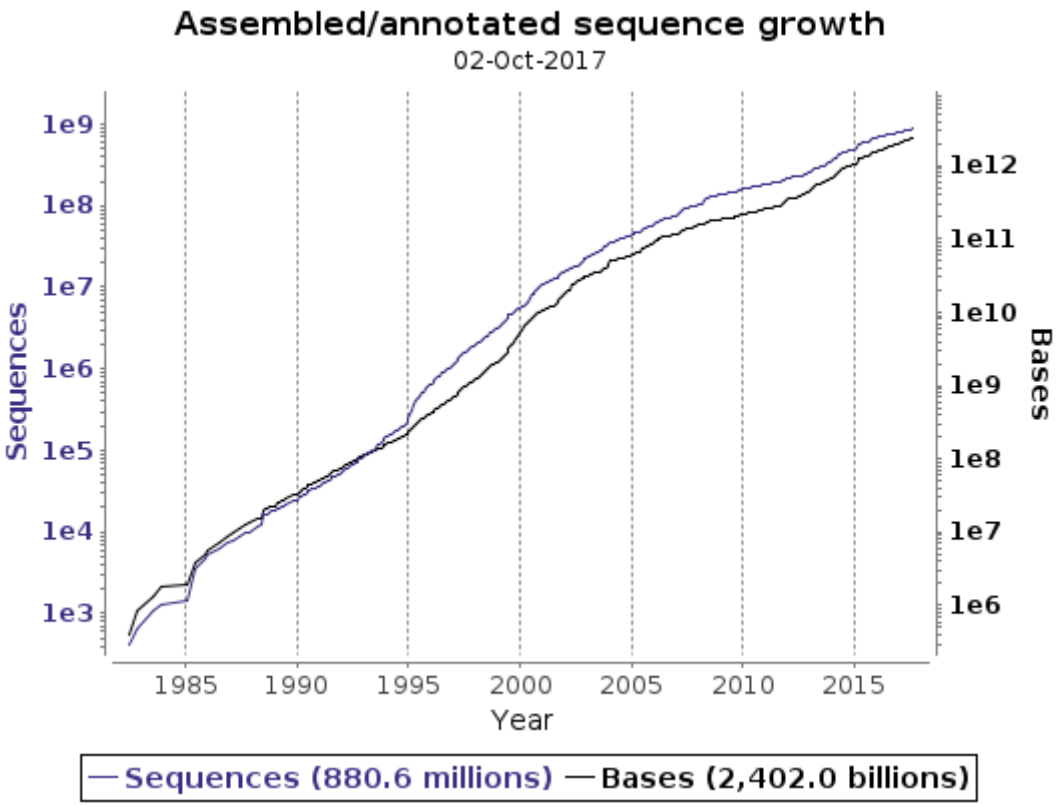


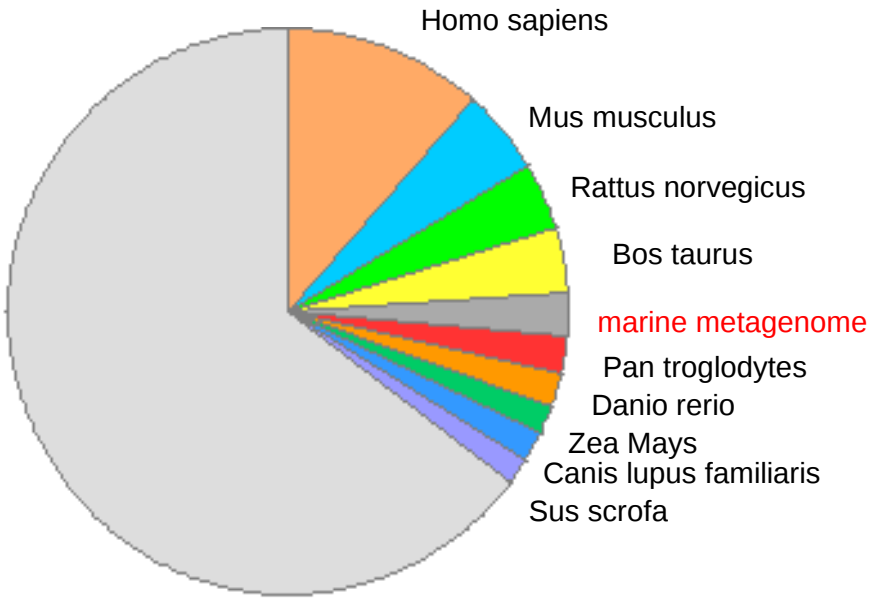
Le séquençage des acides nucléiques (Haut débit Illumina)



Le séquençage des acides nucléiques (Haut débit Illumina)

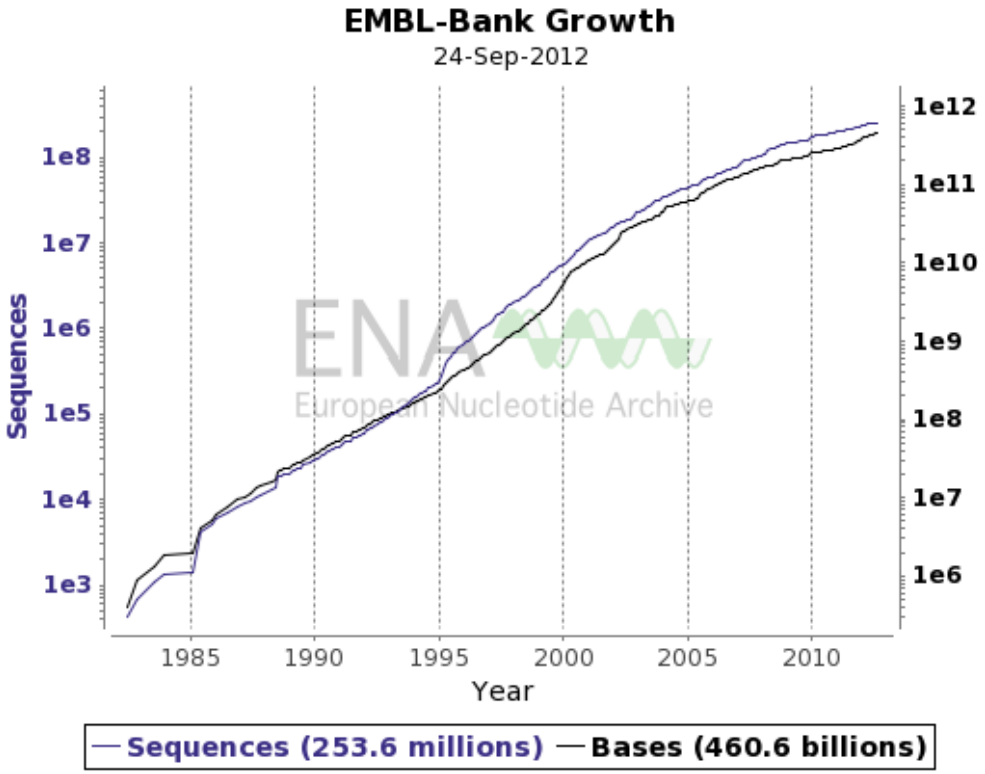
Séquençage des génomes et banques de données

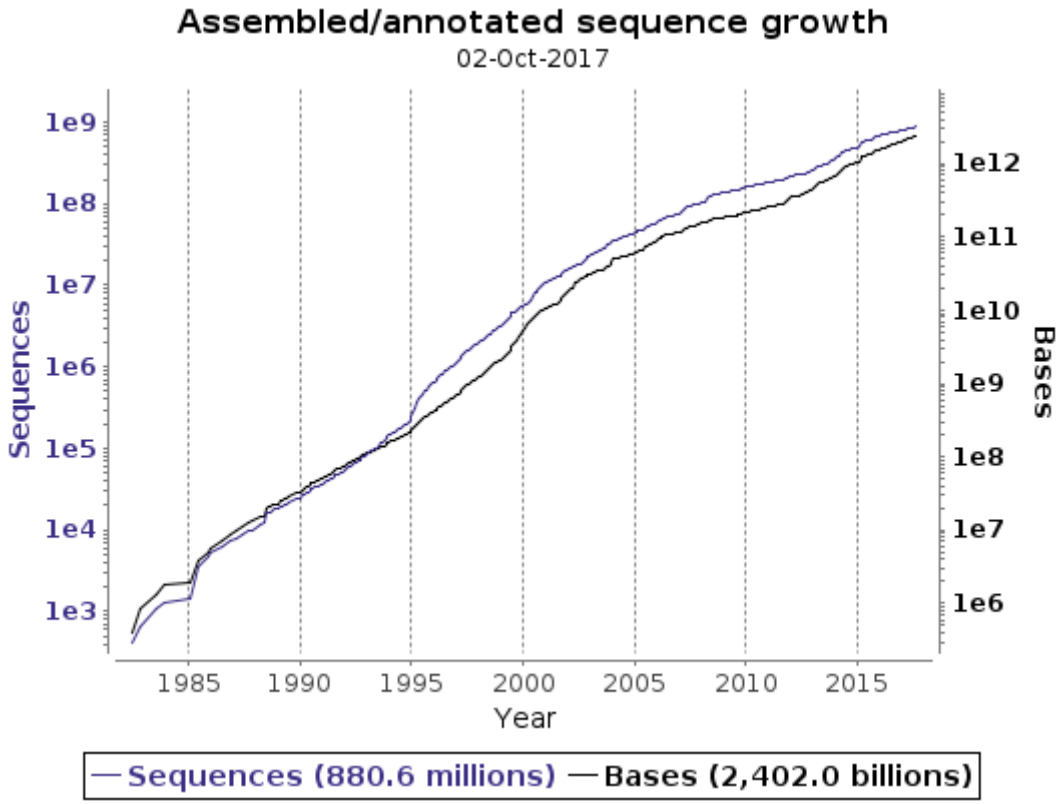




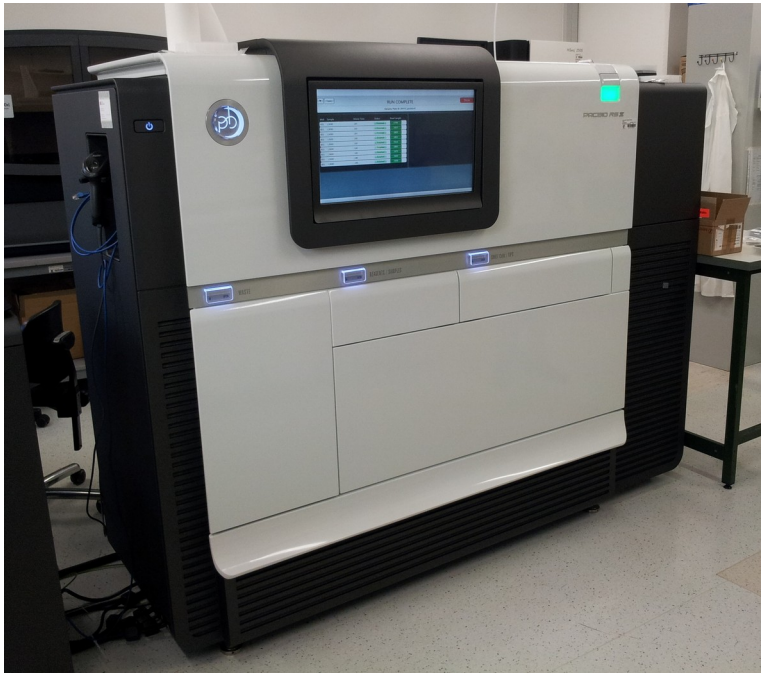
EMBL Stats 02/10/17

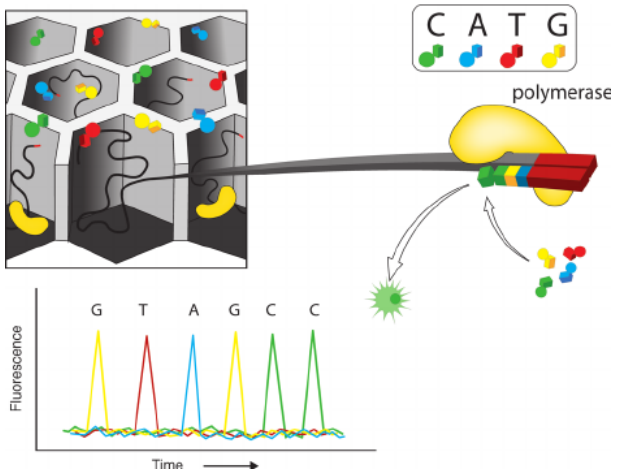
Séquençage des génomes et banques de données

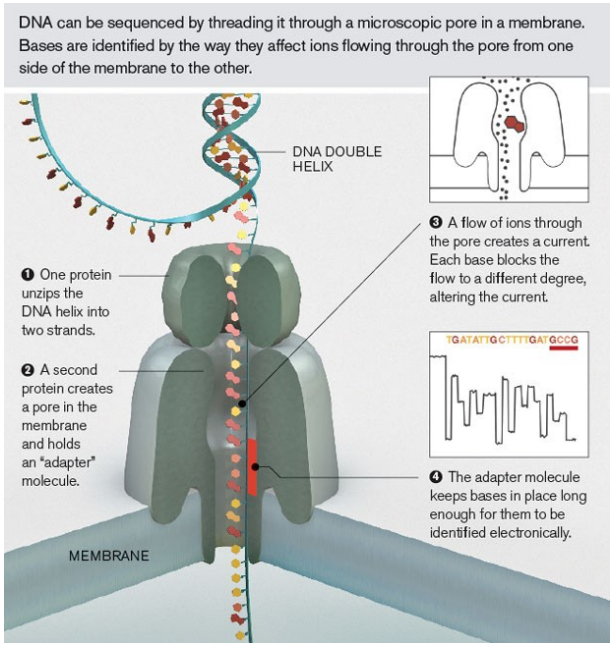




Séquençage des génomes et banques de données







Le séquençage des génomes

Nombre de Génomes Séquencés au 06/10/17
(source GOLD https://gold.jgi.doe.gov)

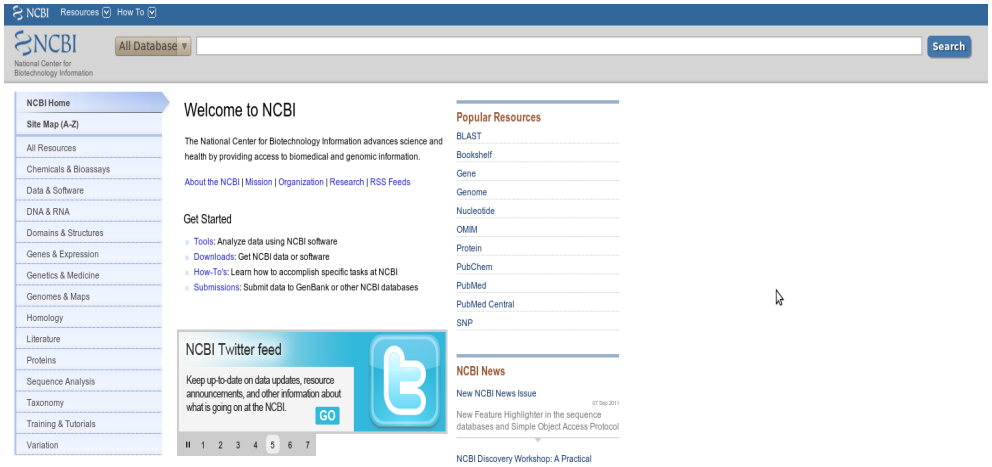
2 414 Archées		250 222 Bactéries		20 800 Eucaryotes	
Methanocaldococcus jannaschii	1996	*Haemophilus influenzae*	1995	*Saccharomyces cerevisiae*	1997
Methanobacterium thermo.	1997	*Helicobacter pylori*	1997	*Caenorhabditis elegans*	1998
Archeoglobus fulgidus	1997	*Escherichia coli K12*	1997	*Drosophila melanogaster*	2000
Pyrococcus horikoshii	1998	*Bacillus subtilis*	1997	***Homo sapiens***	**2001**
Sulfolobus solfataricus	2001	*Mycobacterium tuberculosis*	1998	*Arabidopsis thaliana*	2002
…		*Treponema pallidum*	1998	*Mus musculus*	2002
		Bacillus anthracis	2004	*Oryza sativa*	2002
		…		*Takifugu rubripes*	2002
				Plasmodium falciparum	2002
				Anopheles gambiae	2002
				Neurospora crassa	2003
				Rattus norvegicus	2004
				…	

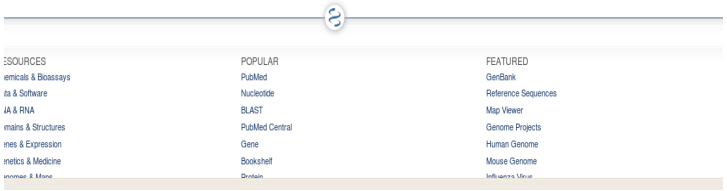
+ 1 073 Métagénomes

Taille des génomes

Virus	Virus de la grippe	13 000	11
	Bactériophage λ	50 000	60
	Mimivirus	1 200 000	1 260
Bactéries	*Haemophilus influenzae*	1 800 000	1 657
	Escherichia coli	4 640 000	4 243
Archées	*Pyrococcus abyssi*	1 770 000	1 898
Eucaryotes	*Saccharomyces cerevisiae*	12 000 000	5 863
	Plasmodium falciparum	21 800 000	5 314
	Drosophila melanogaster	118 000 000	16 548
	Homo sapiens	3 400 000 000	26 517
	Zea mais	5 000 000 000	50 000

La sauvegarde des séquences





Le décodage des génomes

aaccccccccc tcccccgctt ctggccacag cacttaaaca catctctgcc aaaccccaaa aaccccccccc tcccccgctt ctggccacag cacttaaaca catctctgcc aaaccccaaa
aacaaagaac cctaacacca gcctaaccag atttcaaatt ttatcttttg gcggtatgca aacaaagaac cctaacacca gcctaaccag atttcaaatt ttatcttttg gcggtatgca
cttttaacag tcacccccca actaacacat tattttcccc tcccactccc atactactaa cttttaacag tcacccccca actaacacat tattttcccc tcccactccc atactactaa
tctcatcaat acaaccccca cccatcctac ccagcacaca cacaccgctg ctaaccccat tctcatcaat acaaccccca cccatcctac ccagcacaca cacaccgctg ctaaccccat
accccgaacc aaccaaaccc caaagacacc ccccacagtt tatgtagctt acctcctcaa accccgaacc aaccaaaccc caaagacacc ccccacagtt tatgtagctt acctcctcaa
agcaatacac tgaaaatgtt tagacgggct cacatcaccc cataaacaaa taggtttggt agcaatacac tgaaaatgtt tagacgggct cacatcaccc cataaacaaa taggtttggt
cctagccttt ctattagctc ttagtaagat tacacatgca agcatccccg ttccagtgag cctagccttt ctattagctc ttagtaagat tacacatgca agcatccccg ttccagtgag
ttcaccctct aaatcaccac gatcaaaagg gacaagcatc aagcacgcag caatgcagct ttcaccctct aaatcaccac gatcaaaagg gacaagcatc aagcacgcag caatgcagct
caaaacgctt agcctagcca cacccccacg ggaaacagca gtgattaacc tttagcaata caaaacgctt agcctagcca cacccccacg ggaaacagca gtgattaacc tttagcaata
aacgaaagtt taactaagct atactaaccc cagggttggt caatttcgtg ccagccaccg aacgaaagtt taactaagct atactaaccc cagggttggt caatttcgtg ccagccaccg
cggtcacacg attaacccaa gtcaatagaa gccggcgtaa agagtgtttt agatcacccc cggtcacacg attaacccaa gtcaatagaa gccggcgtaa agagtgtttt agatcacccc
ctccccaata aagctaaaac tcacctgagt tgtaaaaaac tccagttgac acaaaataga ctccccaata aagctaaaac tcacctgagt tgtaaaaaac tccagttgac acaaaataga
ctacgaaagt ggctttaaca tatctgaaca cacaatagct aagacccaaa ctgggattag ctacgaaagt ggctttaaca tatctgaaca cacaatagct aagacccaaa ctgggattag
ataccccact atgcttagcc ctaaacctca acagttaaat caacaaaact gctcgccaga ataccccact atgcttagcc ctaaacctca acagttaaat caacaaaact gctcgccaga
acactacgag ccacagctta aaactcaaag gacctggcgg tgcttcatat ccccctagag acactacgag ccacagctta aaactcaaag gacctggcgg tgcttcatat ccccctagag
gagcctgttc tgtaatcgat aaaccccgat caacctcacc acctcttgct cagcctatat gagcctgttc tgtaatcgat aaaccccgat caacctcacc acctcttgct cagcctatat
accgccatct tcagcaaacc ctgatgaagg ctacaaagta agcgcaagta cccacgtaaa accgccatct tcagcaaacc ctgatgaagg ctacaaagta agcgcaagta cccacgtaaa
gacgttaggt caaggtgtag cccatgaggt ggcaagaaat gggctacatt ttctacccca gacgttaggt caaggtgtag cccatgaggt ggcaagaaat gggctacatt ttctacccca
gaaaactacg atagccctta tgaaacttaa gggtcgaagg tggatttagc agtaaactga gaaaactacg atagccctta tgaaacttaa gggtcgaagg tggatttagc agtaaactga
gagtagagtg cttagttgaa cagggccctg aagcgcgtac acaccgcccg tcaccctcct gagtagagtg cttagttgaa cagggccctg aagcgcgtac acaccgcccg tcaccctcct
caagtatact tcaaaggaca tttaactaaa acccctacgc atttatatag aggagacaag caagtatact tcaaaggaca tttaactaaa acccctacgc atttatatag aggagacaag
gatcacaggt ctatcaccct attaaccact cacgggagct ctccatgcat ttggtatttt gatcacaggt ctatcaccct attaaccact cacgggagct ctccatgcat ttggtatttt
cgtctggggg gtgtgcacgc gatagcattg cgagacgctg gagccggagc accctatgtc cgtctggggg gtgtgcacgc gatagcattg cgagacgctg gagccggagc accctatgtc
gcagtatctg tctttgattc ctgccccatc ccattattta tcgcacctac gttcaatatt gcagtatctg tctttgattc ctgccccatc ccattattta tcgcacctac gttcaatatt
acaggcgaac atacctacta aagtgtgtta attaattaat gcttgtagga cataataata acaggcgaac atacctacta aagtgtgtta attaattaat gcttgtagga cataataata
acaattgaat gtctgcacag ccgctttcca cacagacatc ataacaaaaa atttccacca acaattgaat gtctgcacag ccgctttcca cacagacatc ataacaaaaa atttccacca
aaccccccccc tcccccgctt ctggccacag cacttaaaca catctctgcc aaaccccaaa aaccccccccc tcccccgctt ctggccacag cacttaaaca catctctgcc aaaccccaaa
aacaaagaac cctaacacca gcctaaccag atttcaaatt ttatcttttg gcggtatgca aacaaagaac cctaacacca gcctaaccag atttcaaatt ttatcttttg gcggtatgca
cttttaacag tcacccccca actaacacat tattttcccc tcccactccc atactactaa cttttaacag tcacccccca actaacacat tattttcccc tcccactccc atactactaa
tctcatcaat acaaccccca cccatcctac ccagcacaca cacaccgctg ctaaccccat tctcatcaat acaaccccca cccatcctac ccagcacaca cacaccgctg ctaaccccat
accccgaacc aaccaaaccc caaagacacc ccccacagtt tatgtagctt acctcctcaa accccgaacc aaccaaaccc caaagacacc ccccacagtt tatgtagctt acctcctcaa
agcaatacac tgaaaatgtt tagacgggct cacatcaccc cataaacaaa taggtttggt agcaatacac tgaaaatgtt tagacgggct cacatcaccc cataaacaaa taggtttggt
cctagccttt ctattagctc ttagtaagat tacacatgca agcatccccg ttccagtgag cctagccttt ctattagctc ttagtaagat tacacatgca agcatccccg ttccagtgag
ttcaccctct aaatcaccac gatcaaaagg gacaagcatc aagcacgcag caatgcagct ttcaccctct aaatcaccac gatcaaaagg gacaagcatc aagcacgcag caatgcagct
caaaacgctt agcctagcca cacccccacg ggaaacagca gtgattaacc tttagcaata caaaacgctt agcctagcca cacccccacg ggaaacagca gtgattaacc tttagcaata
aacgaaagtt taactaagct atactaaccc cagggttggt caatttcgtg ccagccaccg aacgaaagtt taactaagct atactaaccc cagggttggt caatttcgtg ccagccaccg
cggtcacacg attaacccaa gtcaatagaa gccggcgtaa agagtgtttt agatcacccc cggtcacacg attaacccaa gtcaatagaa gccggcgtaa agagtgtttt agatcacccc
ctccccaata aagctaaaac tcacctgagt tgtaaaaaac tccagttgac acaaaataga ctccccaata aagctaaaac tcacctgagt tgtaaaaaac tccagttgac acaaaataga
ctacgaaagt ggctttaaca tatctgaaca cacaatagct aagacccaaa ctgggattag ctacgaaagt ggctttaaca tatctgaaca cacaatagct aagacccaaa ctgggattag
ataccccact atgcttagcc ctaaacctca acagttaaat caacaaaact gctcgccaga ataccccact atgcttagcc ctaaacctca acagttaaat caacaaaact gctcgccaga
acactacgag ccacagctta aaactcaaag gacctggcgg tgcttcatat ccccctagag acactacgag ccacagctta aaactcaaag gacctggcgg tgcttcatat ccccctagag
gagcctgttc tgtaatcgat aaaccccgat caacctcacc acctcttgct cagcctatat gagcctgttc tgtaatcgat aaaccccgat caacctcacc acctcttgct cagcctatat
accgccatct tcagcaaacc ctgatgaagg ctacaaagta agcgcaagta cccacgtaaa accgccatct tcagcaaacc ctgatgaagg ctacaaagta agcgcaagta cccacgtaaa
gacgttaggt caaggtgtag cccatgaggt ggcaagaaat gggctacatt ttctacccca gacgttaggt caaggtgtag cccatgaggt ggcaagaaat gggctacatt ttctacccca
gaaaactacg atagccctta tgaaacttaa gggtcgaagg tggatttagc agtaaactga gaaaactacg atagccctta tgaaacttaa gggtcgaagg tggatttagc agtaaactga
gagtagagtg cttagttgaa cagggccctg aagcgcgtac acaccgcccg tcaccctcct gagtagagtg cttagttgaa cagggccctg aagcgcgtac acaccgcccg tcaccctcct

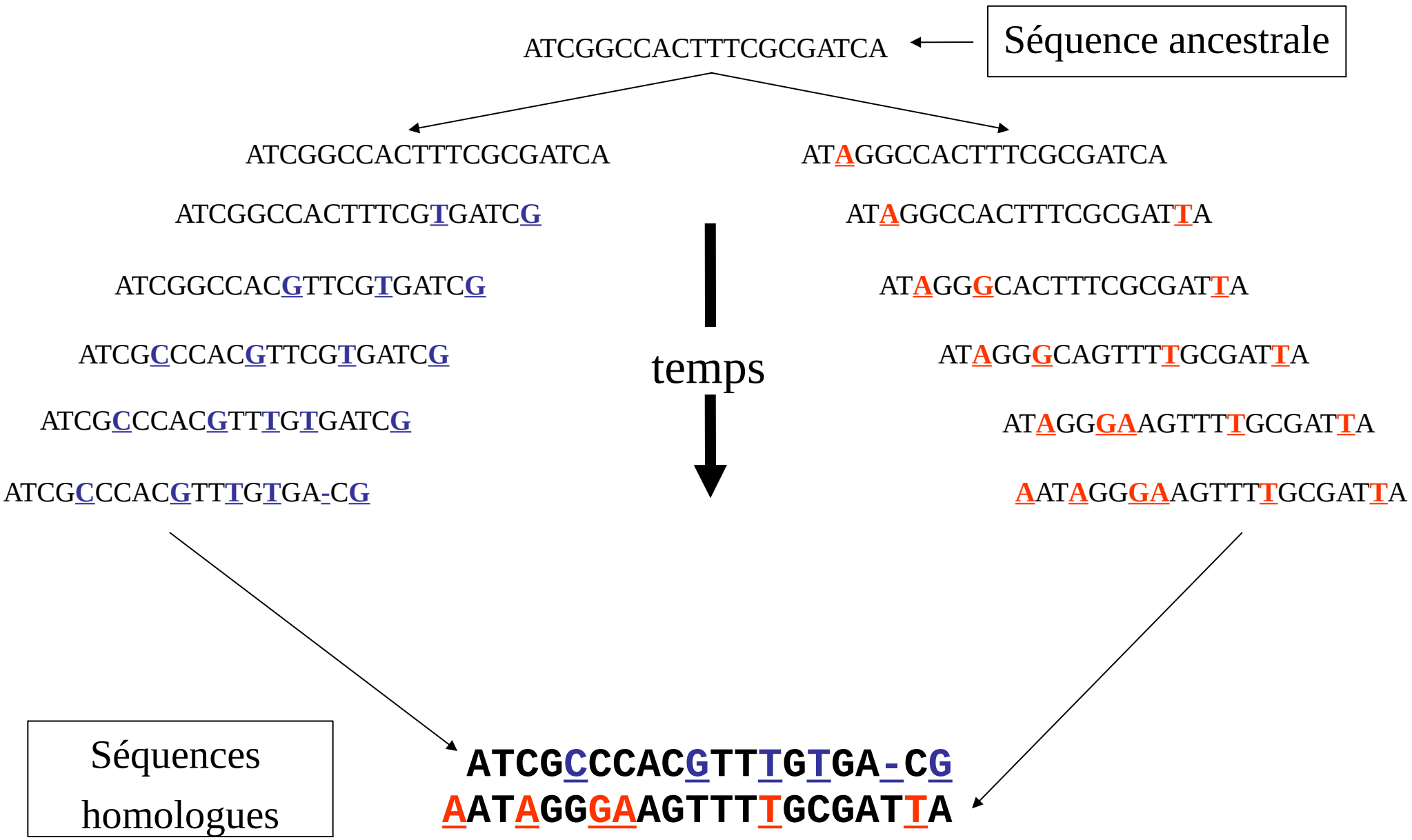
Le décodage des génomes - Annotation structurale

Recherche des ORF

Recherche des CDS

Recherche des éléments régulateurs

Evolution des séquences



Identité / similarité de séquence

On recherche les caractères qui dans les deux séquences suivent le même ordre.
(Homologie des caractères élémentaires ~ résidus)

seq 1	MARSATTACKS	**MAR**S-**A**TTACK**S**	**42% Id.**
seq2	MARTIANS	**MAR**TI**A**N----**S**	

Pour un alignement optimal, on utilise un nombre minimal de mésappariements et d'insertions/délétions afin de maximiser le nombre de résidus identiques.

L'**identité** / **similarité** de deux séquences est une **information quantitative** (%)

seq 1	MARSATTACKS	**MAR**S-**A**TTACK**S**	**42% Id.**
Seq 2	MARTIANS	**MAR**TI**A**N----**S**	

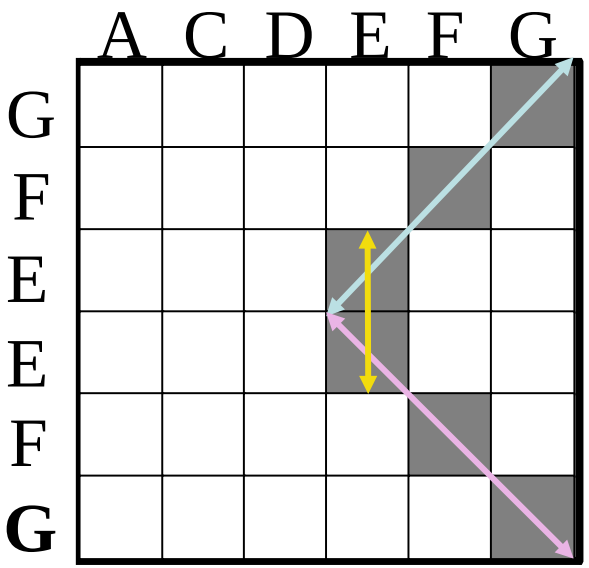
seq 1	MARSATTACKS	**M**AR**S**-**A**TT**A**CK**S**	**42% Id.**
seq 3	MTGSTALVAS	**M**TG**S**T**A**LV**A**--**S**	

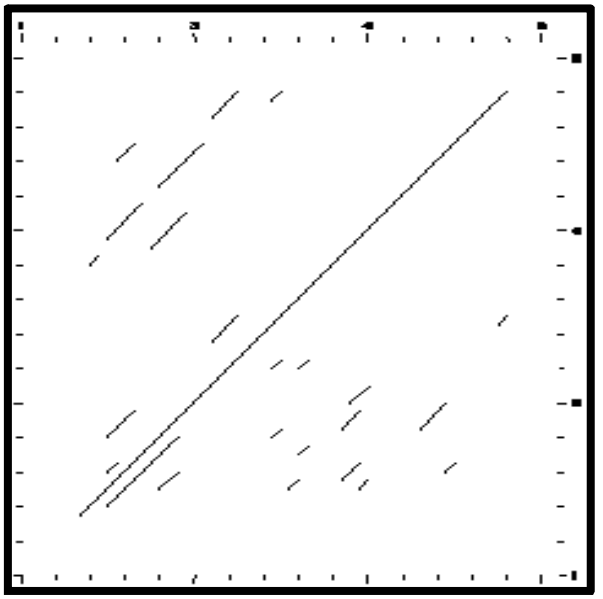
Dot Matrix Sequence Comparison

Gibbs A.J. and McIntyre G.A. (1970)

The Diagram, a method for comparing sequences. Its use with amino acids sequences.

Eur. J. Biochem. (16):1-11





Alignment of pairs of sequences

Sequence Alignments: Search for individual characters that are in the same order in the sequences.

In an optimal alignment, non-identical characters (mismatches) and gaps are placed to bring as many identical or similar characters as possible.

7 identities
2 gaps
10 mismatches

```
LGPSSKQTGKGS-SRIWDN
|     |  |||   |  |
LN-ITKSAGKGAIMRLGDA
```

Scoring sequence alignments

```
LGPSSKQTGKGS-SRIWDN
|     |  |||   |  |
LN-ITKSAGKGAIMRLGDA
```

7	Identities	+ 7
2	Gaps	- 1
10	Mismatches	- 10
		- 4

Use of scoring Matrix:

- PAM 250
- BLOSUM 62

Dayhoff Amino Acid Substitution Matrices

Percent Accepted Mutation (PAM)

Margaret Oakley Dayhoff (1925-1983)

List the likelihood of change from one amino acid to another.

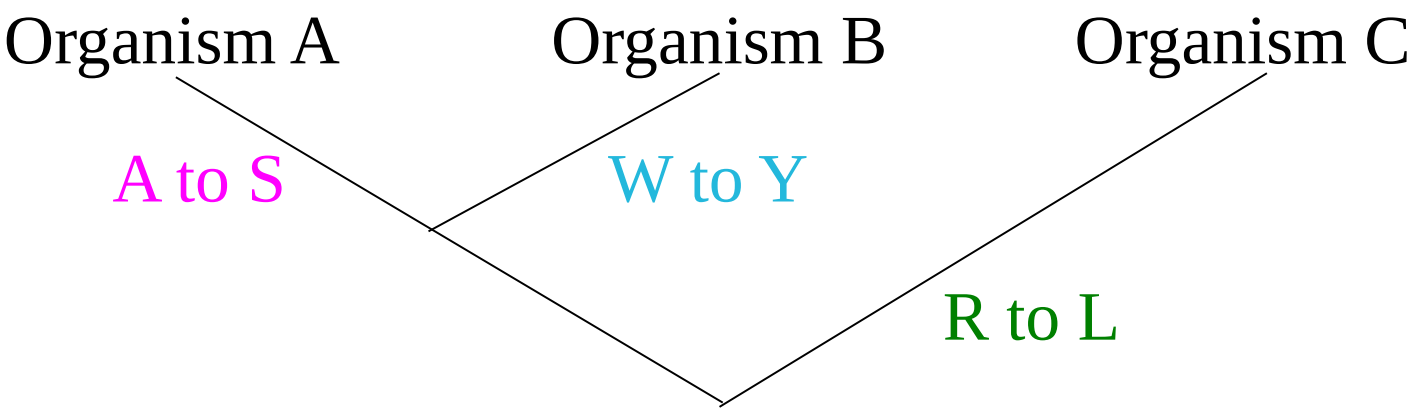
Based on a small data set .

- 1572 changes in 71 groups of proteins
- sequence similarity ≥ 85% (Accepted Mutations)

Amino acid substitution observed over short periods of evolutionary history can be extrapolated to longer distances. Each Matrix gives the changes expected for a given period of Evolutionary time.

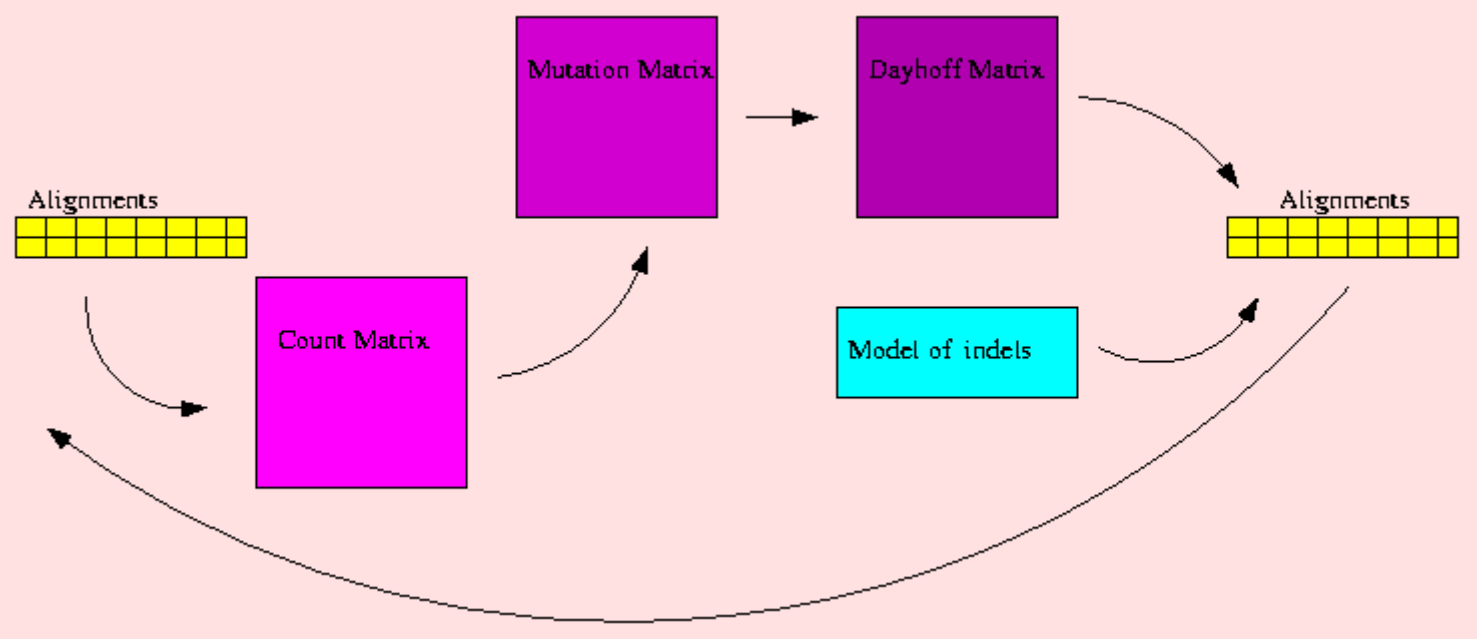
The PAM1 score matrix

Organism A	A	W	T	V	A	S	A	V	R	T	S	I
Organism B	A	Y	T	V	A	A	A	V	R	T	S	I
Organism C	A	W	T	V	A	A	A	V	L	T	S	I



Score of changing

$$S_{A \to B} = \#\ \text{Changes}_{A \to B} \times (\#\ \text{Changes}_{aa})_A \times \frac{(\%\ \text{all changes})_A}{(f_{aa})_A}$$



	A	G	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C
A	4372	121	113	162	96	140	324	387	40	136	212	219	72	50	213	596	300	6	39	428
G	121	2748	100	70	20	188	170	88	69	48	85	679	38	26	46	129	110	10	30	59
V	113	100	1727	322	16	92	166	190	89	37	42	209	17	18	38	207	160	4	40	45
L	162	70	322	3092	8	127	732	154	68	18	36	176	11	14	67	204	120	4	24	28
I	96	20	16	8	680	6	6	20	6	30	30	12	10	22	6	61	54	7	16	83
P	140	188	92	127	6	1367	346	58	70	35	100	310	52	18	48	128	103	6	30	66
S	324	170	166	732	6	346	3378	138	63	40	68	411	29	13	93	212	192	6	28	104
T	387	88	190	154	20	58	138	5788	30	19	36	136	14	17	46	230	78	8	12	40
D	40	69	89	68	6	70	63	30	1059	14	39	92	16	54	34	48	40	13	96	27
E	136	48	37	18	30	35	40	19	14	3217	858	72	198	144	24	55	156	12	37	1300
N	212	85	42	36	30	100	68	36	39	858	5061	124	410	331	62	104	133	20	92	593
Q	219	679	209	176	12	310	411	136	92	72	124	3054	48	23	84	216	172	4	32	84
K	72	38	17	11	10	52	29	14	16	198	410	48	923	82	12	38	64	5	24	134
R	50	26	18	14	22	18	13	17	54	144	331	23	82	2191	14	38	39	59	367	101
H	213	46	38	67	6	48	93	46	34	24	62	84	12	14	2757	147	66	2	10	54
F	596	129	207	204	61	128	212	230	48	55	104	216	38	38	147	2326	490	13	28	100
Y	300	110	160	120	54	103	192	78	40	156	133	172	64	39	66	490	2808	4	31	299
W	6	10	4	4	7	6	6	8	13	12	20	4	5	59	2	13	4	582	48	14
M	39	30	40	24	16	30	28	12	96	37	92	32	24	367	10	28	31	48	1612	42
C	428	59	45	28	83	66	104	40	27	1300	593	84	134	101	54	100	299	14	42	3784

	A	G	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C
A	0.54	0.03	0.03	0.03	0.08	0.04	0.05	0.05	0.02	0.02	0.03	0.04	0.03	0.01	0.06	0.11	0.06	0.01	0.01	0.06
G	0.02	0.57	0.03	0.01	0.02	0.06	0.03	0.01	0.04	0.01	0.01	0.11	0.02	0.01	0.01	0.02	0.02	0.01	0.01	0.01
V	0.01	0.02	0.48	0.06	0.01	0.03	0.03	0.03	0.05	0.01	0.01	0.03	0.01	0.00	0.01	0.04	0.03	0.00	0.02	0.01
L	0.02	0.01	0.09	0.57	0.01	0.04	0.11	0.02	0.03	0.00	0.00	0.03	0.01	0.00	0.02	0.04	0.02	0.00	0.01	0.00
I	0.01	0.00	0.00	0.00	0.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.01	0.01
P	0.02	0.04	0.03	0.02	0.01	0.42	0.05	0.01	0.04	0.01	0.01	0.05	0.02	0.00	0.01	0.02	0.02	0.01	0.01	0.01
S	0.04	0.04	0.05	0.13	0.01	0.11	0.52	0.02	0.03	0.01	0.01	0.07	0.01	0.00	0.02	0.04	0.04	0.01	0.01	0.01
T	0.05	0.02	0.05	0.03	0.02	0.02	0.02	0.77	0.02	0.00	0.00	0.02	0.01	0.00	0.01	0.04	0.01	0.01	0.00	0.01
D	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.00	0.54	0.00	0.00	0.01	0.01	0.02	0.01	0.01	0.01	0.02	0.04	0.00
E	0.02	0.01	0.01	0.00	0.03	0.01	0.01	0.00	0.01	0.50	0.10	0.01	0.09	0.04	0.01	0.01	0.03	0.02	0.01	0.18
N	0.03	0.02	0.01	0.01	0.03	0.03	0.01	0.00	0.02	0.13	0.60	0.02	0.19	0.09	0.02	0.02	0.02	0.02	0.03	0.08
Q	0.03	0.14	0.06	0.03	0.01	0.09	0.06	0.02	0.05	0.01	0.01	0.50	0.02	0.01	0.02	0.04	0.03	0.01	0.01	0.01
K	0.01	0.01	0.00	0.00	0.01	0.02	0.00	0.00	0.01	0.03	0.05	0.01	0.42	0.02	0.00	0.01	0.01	0.01	0.01	0.02
R	0.01	0.01	0.00	0.00	0.02	0.01	0.00	0.00	0.03	0.02	0.04	0.00	0.04	0.61	0.00	0.01	0.01	0.07	0.14	0.01
H	0.03	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.02	0.00	0.01	0.01	0.01	0.00	0.72	0.03	0.01	0.00	0.00	0.01
F	0.07	0.03	0.06	0.04	0.05	0.04	0.03	0.03	0.02	0.01	0.01	0.04	0.02	0.01	0.04	0.43	0.09	0.02	0.01	0.01
Y	0.04	0.02	0.04	0.02	0.04	0.03	0.03	0.01	0.02	0.02	0.02	0.03	0.03	0.01	0.02	0.09	0.52	0.01	0.01	0.04
W	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.71	0.02	0.00
M	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.05	0.01	0.01	0.01	0.01	0.10	0.00	0.01	0.01	0.06	0.61	0.01
C	0.05	0.01	0.01	0.01	0.07	0.02	0.02	0.01	0.01	0.20	0.07	0.01	0.06	0.03	0.01	0.02	0.06	0.02	0.02	0.51

The PAM1 score matrix for group of sequence *i*

Exposure to Mutation

$$(EM_{AA})_i = \frac{(f_{AA})_i}{(\%\ \text{all changes})_i}$$

Relative Mutability

$$(RM_A)_i = \frac{(\#\ \text{Changes}_{AA})_i}{(EM_A)_i}$$

Score of changing

$$(S_{A \to B})_i = (\#\ \text{Changes}_{A \to B})_i \,.\, (RM_A)_i$$

The PAM1 score matrix

$$S_{A \rightarrow B} = \sum (S_{A \rightarrow B})_i$$

The resulting score AX were summed and normalized such that that their sum represented a probability of change of 1% (50 million years of evolution).

By multiplying PAM1 by itself *N* times we can obtain transition Matrices for sequences with lower sequence similarity: PAM60 (60%), PAM80 (50%), PAM120 (40%), PAM250 (20%)…

Each change in the current amino acid at a particular site is assumed to be independent of previous mutational event at this site. Amino Acid substitution in a protein sequence is viewed as a Markov model.

Log odds score matrix

Relative Frequency of Change

$$RFC_{A \rightarrow B} = \frac{S_{250,\, A \rightarrow B}}{f_A}$$

Mutation Data Matrix

$$\text{MDM } S_{AB} = \frac{10.\log_{10}(RFC_{A \rightarrow B}) + 10.\log_{10}(RFC_{B \rightarrow A})}{2}$$

The PAM250 score matrix

	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C
G	5																			
A	1	2																		
V	-1	0	4																	
L	-4	-2	2	6																
I	-3	-1	4	2	5															
P	0	1	-1	-3	-2	6														
S	1	1	-1	-3	-1	1	2													
T	0	1	0	-2	0	0	1	3												
D	1	0	-2	-4	-2	-1	0	0	4											
E	0	0	-2	-3	-2	-1	0	0	3	4										
N	0	0	-2	-3	-2	0	1	0	2	1	2									
Q	-1	0	-2	-2	-2	0	-1	-1	2	2	1	4								
K	-2	-1	-2	-3	-2	-1	0	0	0	0	1	1	5							
R	-3	-2	-2	-3	-2	0	0	-1	-1	-1	0	1	3	6						
H	-2	-1	-2	-2	-2	0	-1	-1	1	1	2	3	0	2	6					
F	-5	-3	-1	2	1	-5	-3	-3	-6	-5	-3	-5	-5	-4	-2	9				
Y	-5	-3	-2	-1	-1	-5	-3	-3	-4	-4	-2	-4	-4	-4	0	7	10			
W	-7	-6	-6	-2	-5	-6	-2	-5	-7	-7	-4	-5	-3	-2	-3	0	0	17		
M	-3	-1	2	4	2	-2	-2	-1	-3	-2	-2	-1	0	0	-2	0	-2	-4	6	
C	-3	-2	-2	-6	-2	-3	0	-2	-5	-5	-4	-5	-5	-4	-3	-4	0	-8	-5	12

PAM 250

Blocks Amino Acid Substitution Matrices (BLOSUM)

Henikkoff and Henikoff. (1992)

More than 500 families of related proteins (signatures).

Based on a large set (~2000) of conserved amino acid patterns (3 to 60 aa long) = blocks [Prosite / MOTIF / PROTOMAT].

Based on scoring substitutions found over a range of evolutionary periods.

BLOSUM62

Alignment of sequences by dynamic programming

- Global Sequence Alignment: 1970

  Wunsch C.D. And Needleman S.B.

  A general method applicable to the search for similarities in the amino acid sequence of two proteins.

  (1970) *J. Mol. Biol.* **48**:443-453

- Local Sequence Alignment: 1981

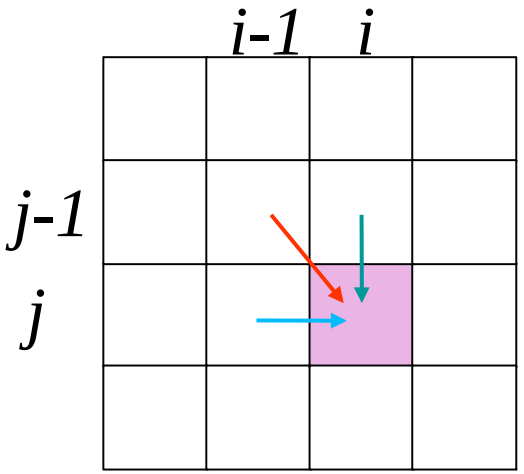
  Smith T.F. and Waterman M.S.

  Identification of common molecular subsequences.

  (1981) *J. Mol. Biol.* **147**:195-197

Dynamic Programming Algorithm

Global : Wunsch and Needleman S.B. C.D. (1970)



$$S_{ij} = \max \{ \; - S_{i\text{-}1,j\text{-}1} + s(a_i b_j)$$

$$- \max_{x \geq 1} (S_{i\text{-}x,j} - w_x)$$

$$- \max_{y \geq 1} (S_{i,j\text{-}y} - w_y)$$

$$\}$$

$$[\; W_x = W_{opening} + W_{addition}(x\text{-}1) \;]$$

Exemple

	GAP	M	N	A	L	S	D	R	T
GAP	0	-12	-16	-20	-24	-28	-32	-36	-40
M	-12	6	-6						
G	-16	-6	6						
S	-20								
D	-24								
R	-28								
T	-32								
T	-36								
E	-40								
T	-44								

GAP Penalty = - 12- 4 (x - 1)

	GAP	M	N	A	L	S	D	R	T
GAP	0	-12	-16	-20	-24	-28	-32	-36	-40
M	-12	6	-6						
G	-16	-6	6						
S	-20								
D	-24								
R	-28								
T	-32								
T	-36								
E	-40								
T	-44								

	GAP	M	N	A	L	S	D	R	T
GAP	0	-12	-16	-20	-24	-28	-32	-36	-40
M	-12	6	-6	**-10**	**-14**	**-18**	**-22**	**-26**	**-30**
G	-16	-6	6	**-5**	**-10**	**-13**	**-17**	**-22**	**-26**
S	-20	**-10**	**-5**	**7**	**-5**	**-8**	**-13**	**-17**	**-21**
D	-24	**-14**	**-8**	**-5**	**3**	**-5**	**-4**	**-14**	**-17**
R	-28	**-18**	**-14**	**-9**	**-8**	**3**	**-6**	**2**	**-10**
T	-32	**-22**	**-18**	**-13**	**-11**	**-7**	**3**	**-7**	**5**
T	-36	**-26**	**-22**	**-22**	**-15**	**-10**	**-7**	**2**	**-4**
E	-40	**-30**	**-25**	**-21**	**-20**	**-15**	**-7**	**-8**	**2**
T	-44	**-34**	**-30**	**-24**	**-23**	**-19**	**-15**	**-8**	**-5**

GAP Penalty = - 12- 4 (*x* - 1)

	GAP	M	N	A	L	S	D	R	T
GAP	0	-12	-16	-20	-24	-28	-32	-36	-40
M	-12	6	-6	-10	-14	-18	-22	-26	-30
G	-16	-6	6	-5	-10	-13	-17	-22	-26
S	-20	-10	-5	7	-5	-8	-13	-17	-21
D	-24	-14	-8	-5	3	-5	-4	-14	-17
R	-28	-18	-14	-9	-8	3	-6	2	-10
T	-32	-22	-18	-13	-11	-7	3	-7	5
T	-36	-26	-22	-22	-15	-10	-7	2	-4
E	-40	-30	-25	-21	-20	-15	-7	-8	2
T	-44	-34	-30	-24	-23	-19	-15	-8	-5

M-NALSDRT
| |
MGSDRTTET

Global *vs* Local

```
LGPSSKQTGKGS-SRIWDN
|    |  |||   |  |
LN-ITKSAGKGAIMRLGDA
```

Global

The entire sequence is aligned

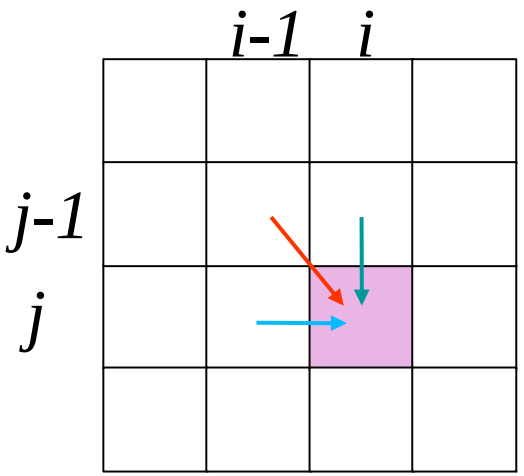
```
          LGPSSKQTGKGS-SRIWDN
                   |||
          LN-ITKSAGKGAIMRLGDA
```

Local

Stretches of sequences are aligned

Dynamic Programming Algorithm

Local : Smith T.F. and Waterman M.S. *(1981)*



$$S_{ij} = \max \{ \; - S_{i\text{-}1,j\text{-}1} + s(a_i b_j)$$

$$- \max_{x \geq 1} (S_{i\text{-}x,j} - w_x)$$

$$- \max_{y \geq 1} (S_{i,j\text{-}y} - w_y)$$

$$- 0 \; \leftarrow$$

$$\}$$

$$[\; W_x = W_{opening} + W_{addition}(x\text{-}1) \;]$$

Exemple

	GAP	M	N	A	L	S	D	R	T
GAP	0	0	0	0	0	0	0	0	0
M	0	6	0	0	4	0	0	0	0
G	0	0	6	1	0	5	1	0	0
S	0	0	1	7	0	2	5	1	1
D	0	0	2	1	3	0	6	4	1
R	0	0	0	0	0	3	0	12	3
T	0	0	0	1	0	1	3	0	15
T	0	0	0	1	0	1	1	2	3
E	0	0	1	0	0	0	4	0	2
T	0	0	0	2	0	1	0	3	3

SDRT

| | | | |

SDRT

BLAST sequence database similarity search

Basic Local Alignment Search Tool

Altschul S.F. and al. (1990) J. Mol. Biol. (215):403-410

http://www.ncbi.nlm.nih.gov/BLAST/

BLAST search Algorithm

Step 1: The query sequence is filtered to remove low-complexity region

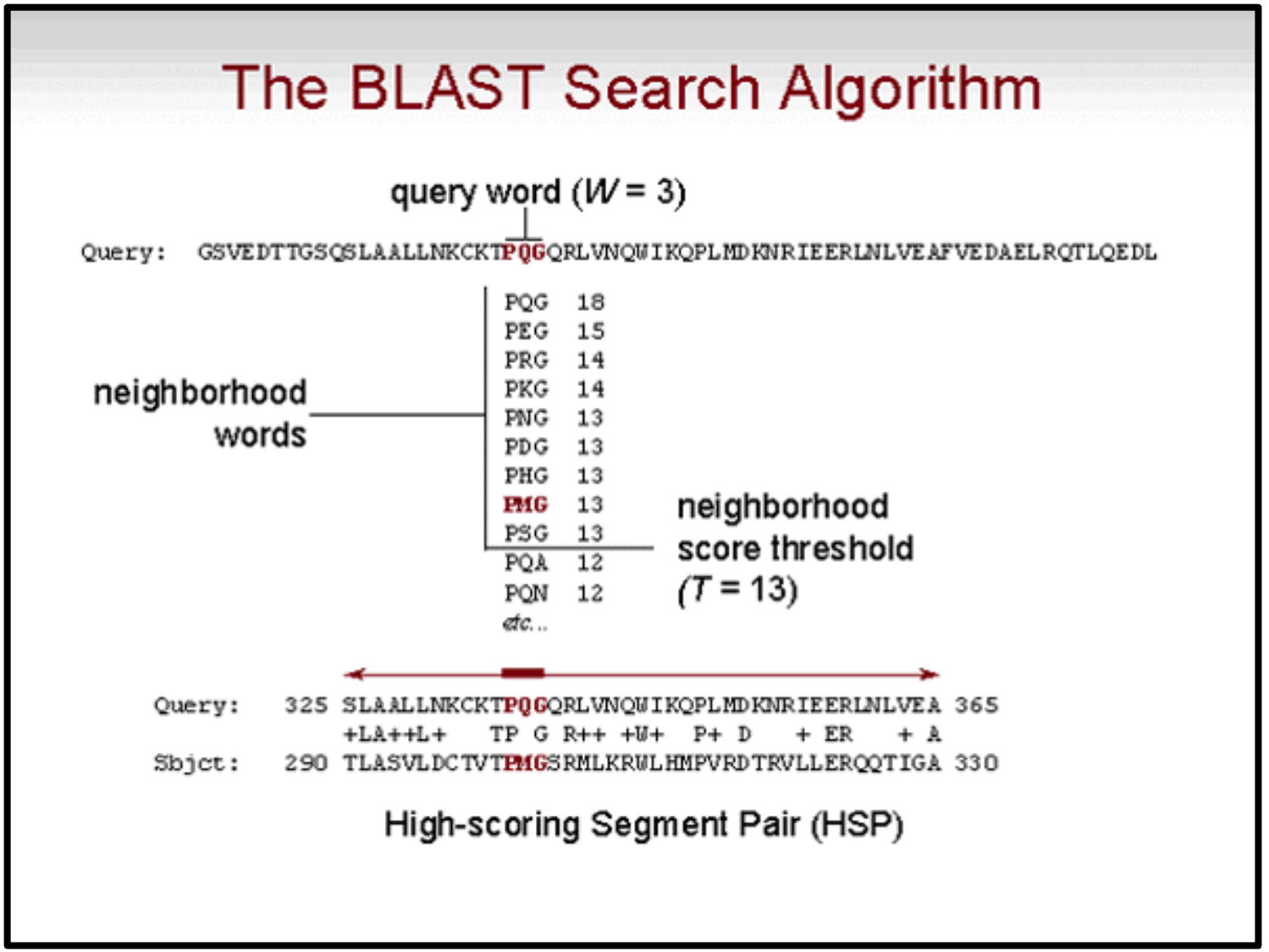
Step 2: A list of words of length 3 (11 for DNA) in the query proteins is made:

```
          MARSATTACKS
MAR ARS RSA SAT ATT TTA TAC ACK CKS
```

Step 3: Using the BLOSUM62 scoring matrix the words are evaluated for an exact match and also for matches with any other combination of three amino acids.

Step 4: A cutoff score called neighborhood word score threshold (T) is selected to reduce the number of possible matches to the most significant ones (score $\geq T$). The remaining high-scoring words are organized into an efficient search tree.

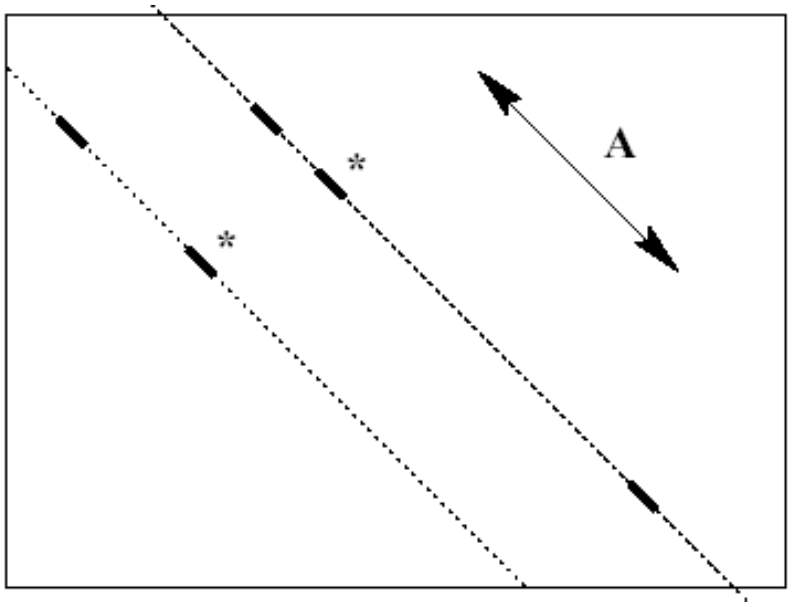
BLAST search Algorithm



BLAST search Algorithm

Step 5: Each database sequence is scanned for an exact match to one of the remaining words. If a match id found, this match is used to seed a possible ungapped alignment.

Step 6: Short ungapped matched regions lying on the same diagonal and within distance A of each other are used as starting point for a longer ungapped alignment between the words. (scores are summed)



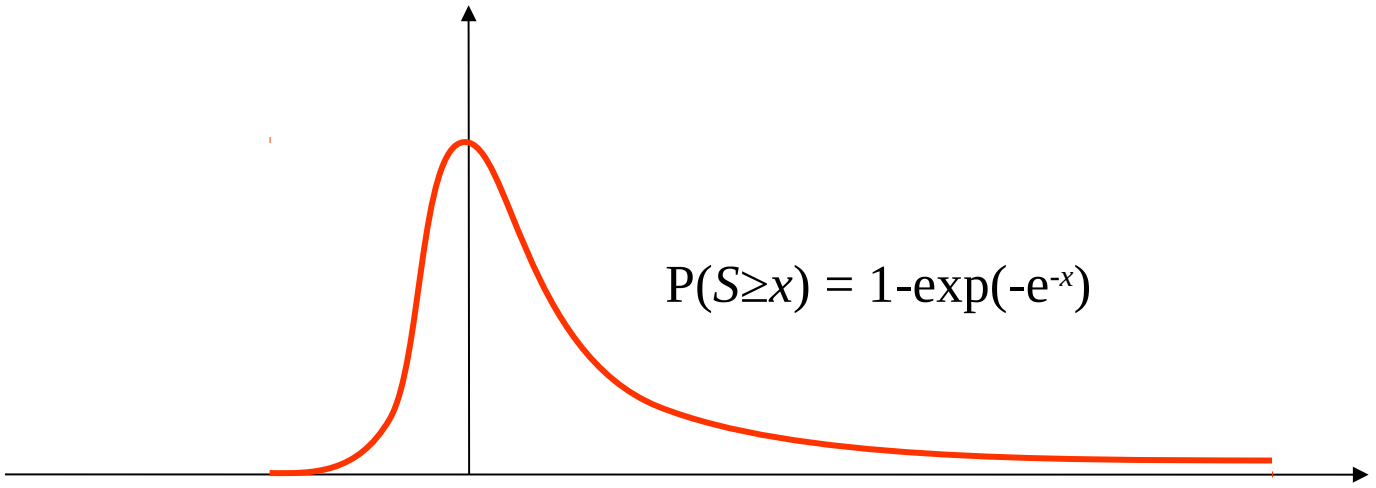
BLAST search Algorithm

Step 7: These joined regions are extended in each direction along the sequences, continuing for as long the score continued to increase. At this point, a larger stretch of sequence (called the HSP or High-scoring Segment Pair), may have been found.

Step 8: The HSPs scores greater than a cutoff score *S* are identified and listed.

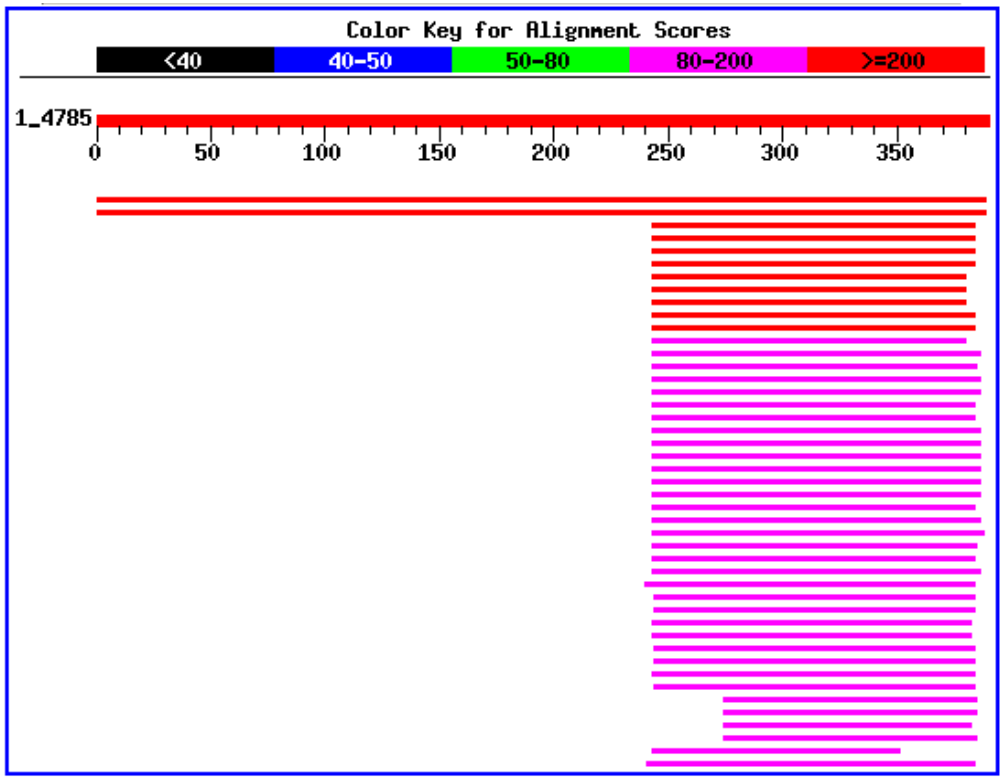
Step 9: The statistical significance of each HSPs is determined. The ones with expect scores greater than the user threshold parameter *E* are reported. A Smith and Waterman local alignment Is shown the query sequence with each of the matched sequences in the database.

When two sequences have been aligned optimally, the significance of a local alignment score can be tested on the basis of the distribution of scores expected by aligning two random sequences of the same length and composition as the two test sequences (Karlin and Altschul 1990).



The random sequence alignment scores follow a extreme value distribution.

Homologie, orthologie et paralogie



Organism Report

Drosophila melanogaster [flies] taxid 7227

gi|84957|pir||S06222 finger protein snail - fruit fly (Dro...	590	8e-168
gi|17136490|ref|NP_476732.1| CG3956-PA [Drosophila melanog...	589	2e-167
gi|7287904|gb|AAF44942.1| symbol=wor; synonym=BG:DS03023.1...	215	6e-55
gi|6465949|gb|AAF12733.1| zinc finger protein Worniu [Dros...	214	1e-54
gi|17136260|ref|NP_476601.1| CG4158-PA [Drosophila melanog...	214	2e-54
gi|19528395|gb|AAL90312.1| RE10012p [Drosophila melanogaster]	214	2e-54
gi|17136258|ref|NP_476600.1| CG3758-PA [Drosophila melanog...	205	7e-52
gi|17946356|gb|AAL49212.1| RE64266p [Drosophila melanogaster]	205	7e-52
gi|119567|sp|P25932|ESCA_DROME Escargot protein (Fleabag p...	205	8e-52
gi|24657086|ref|NP_523911.2| CG11130-PA [Drosophila melanog...	149	7e-35
gi|1022788|gb|AAA91035.1| neuron specific zinc finger tran...	149	7e-35
gi|24654785|ref|NP_612040.1| CG17181-PA [Drosophila melano...	149	8e-35
gi|24657081|ref|NP_647845.2| CG12605-PB [Drosophila melano...	148	1e-34
gi|17944451|gb|AAL48115.1| RH02885p [Drosophila melanogaster]	148	1e-34
gi|45552939|ref|NP_995996.1| CG12605-PC [Drosophila melano...	146	5e-34
gi|5052544|gb|AAD38602.1| scratch [Drosophila melanogaster]	112	7e-24

Patella vulgata [molluscs] taxid 6465

gi|17223770|gb|AAL06240.1| Sna1 [Patella vulgata]	200	2e-50
gi|17223774|gb|AAL12167.1| SNA2 [Patella vulgata]	193	3e-48
gi|17223772|gb|AAL12166.1| SNA2 [Patella vulgata]	154	2e-36

Nematostella vectensis (starlet sea anemone) [anthozoans] taxid 45351

gi|38569877|gb|AAR24456.1| snail family zinc-finger protei...	200	3e-50
gi|38569879|gb|AAR24457.1| snail-family zinc finger protei...	179	7e-44
gi|33621860|gb|AAQ23385.1| snail [Nematostella vectensis]	178	1e-43

Anopheles gambiae str. PEST [flies] taxid 180454

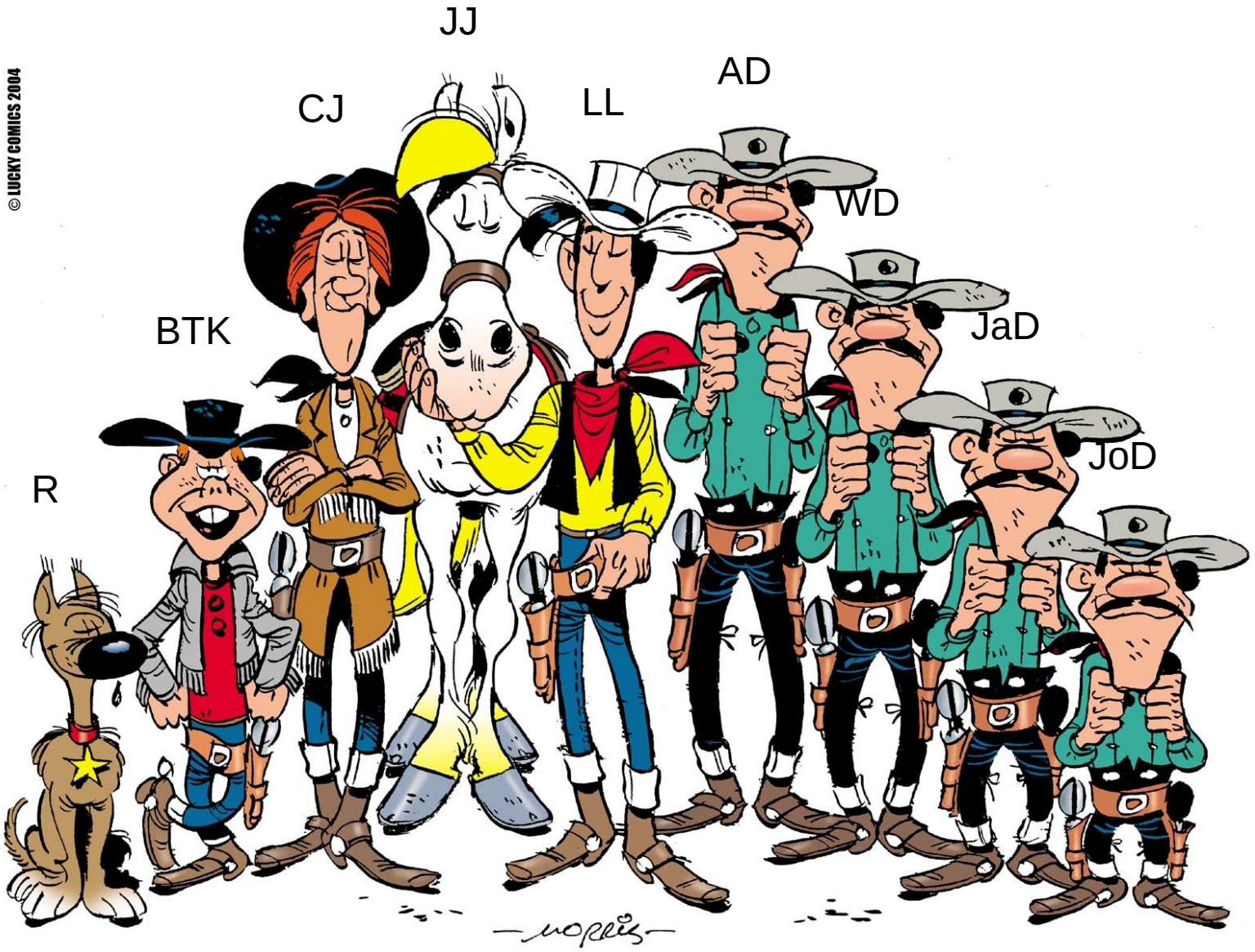
gi|31222580|ref|XP_317196.1| ENSANGP00000018305 [Anopheles...	199	5e-50
gi|31200011|ref|XP_308953.1| ENSANGP00000012782 [Anopheles...	147	2e-34
gi|31212829|ref|XP_315399.1| ENSANGP00000020876 [Anopheles...	147	3e-34
gi|31200017|ref|XP_308956.1| ENSANGP00000005793 [Anopheles...	147	3e-34

Homo sapiens (man) [mammals] taxid 9606

gi|11276067|ref|NP_003059.1| snail 2; neural crest transcr...	198	1e-49
gi|41150294|ref|XP_370995.1| snail homolog 3 [Homo sapiens]	184	2e-45
gi|27552822|gb|AAH41461.1| Similar to snail homolog 3 (Dro...	183	3e-45
gi|4325322|gb|AAD17332.1| zinc finger protein [Homo sapiens]	166	3e-40
gi|18765741|ref|NP_005976.2| snail 1 homolog; snail 1 zinc...	166	3e-40
gi|13775236|ref|NP_112599.1| scratch; scratch 1 [Homo sapi...	152	6e-36
gi|15928387|gb|AAH14675.1| Unknown (protein for IMAGE:4156...	150	2e-35
gi|12697478|emb|CAC00548.2| dJ850E9.1 (novel C2H2 type zin...	150	4e-35
gi|42662251|ref|XP_372858.2| similar to scratch; scratch 1...	145	6e-34
gi|42656568|ref|XP_376171.1| KIAA1843 protein [Homo sapiens]	139	6e-32
gi|37540450|ref|XP_098940.2| similar to zinc finger protei...	113	4e-24
gi|4508015|ref|NP_003447.1| zinc finger protein 205; zinc ...	110	3e-23

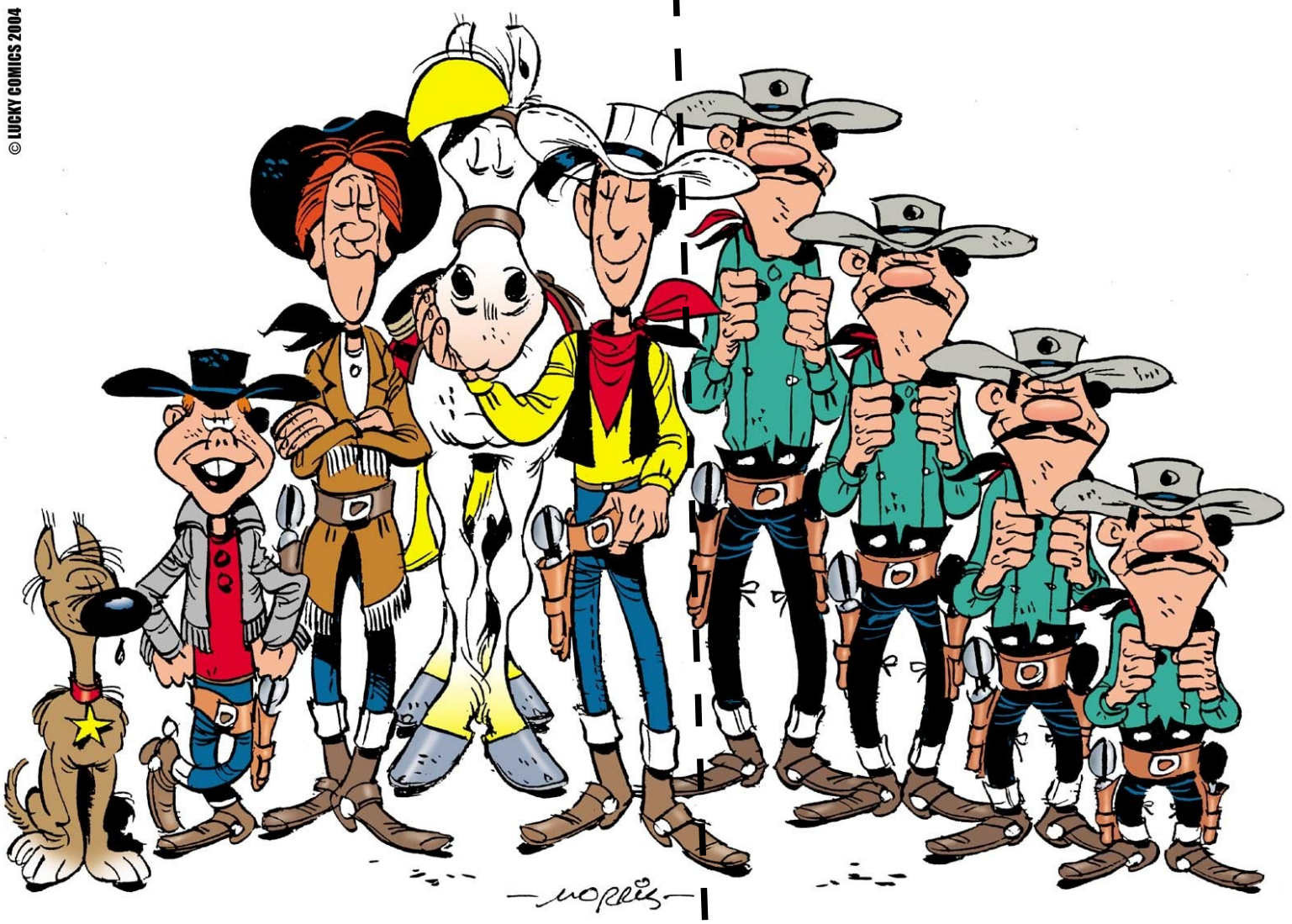
Taxons et espèces

Un taxon regroupe des entités possédant en commun certains caractères


JJ
AD
CJ
LL
WD
BTK
JaD
JoD
R
© LUCKY COMICS 2004
MORRIS

Les NON Daltons

Les Frères Dalton



Taxons et espèces

Un taxon regroupe des entités possédant en commun certains caractères

Une espèce biologique est groupe de population naturelle interféconde, génétiquement isolé de groupes similaires (Ernst Mayr).

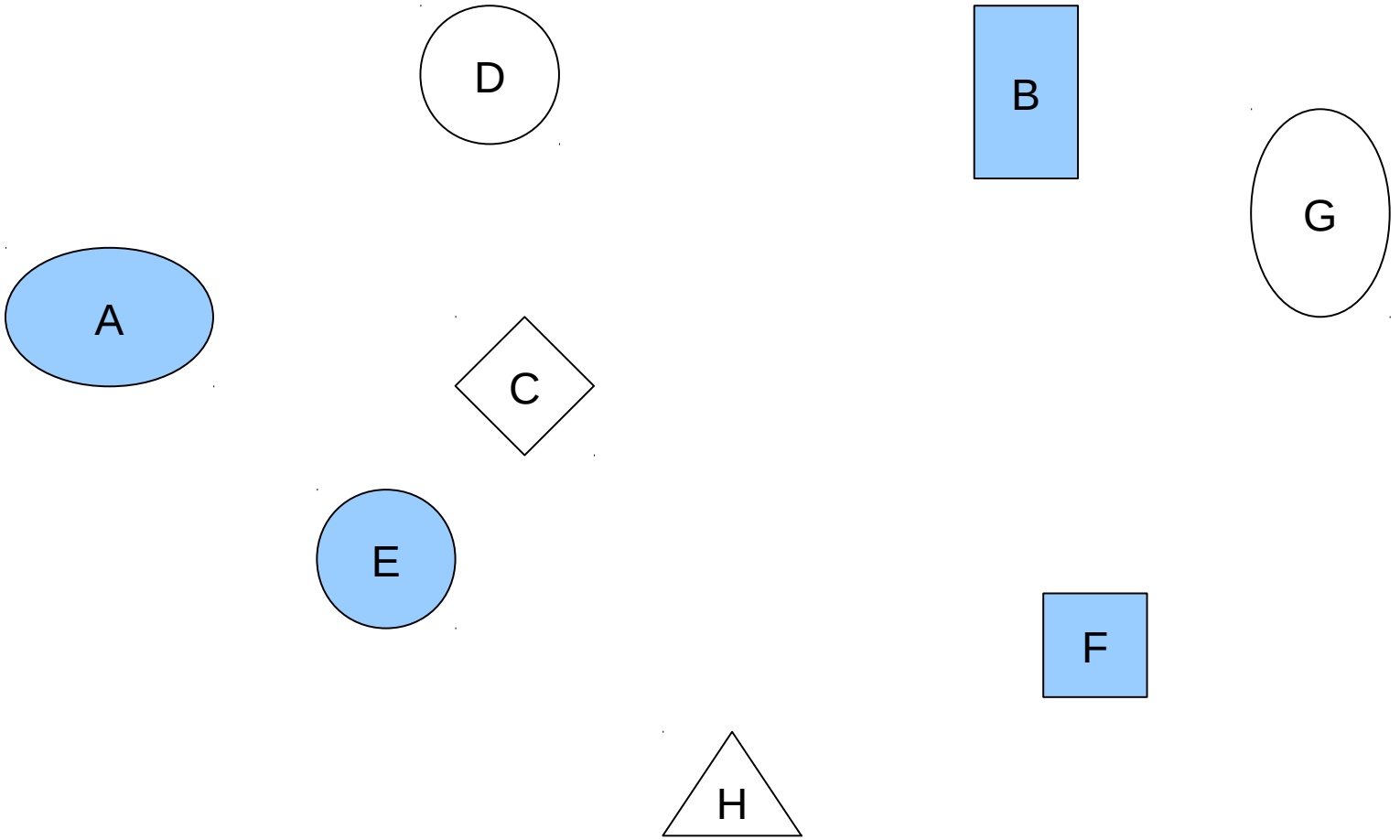
Taxons et espèces

Un taxon regroupe des entités possédant en commun certains caractères

Une espèce biologique est groupe de population naturelle interféconde, génétiquement isolé de groupes similaires (Ernst Mayr).

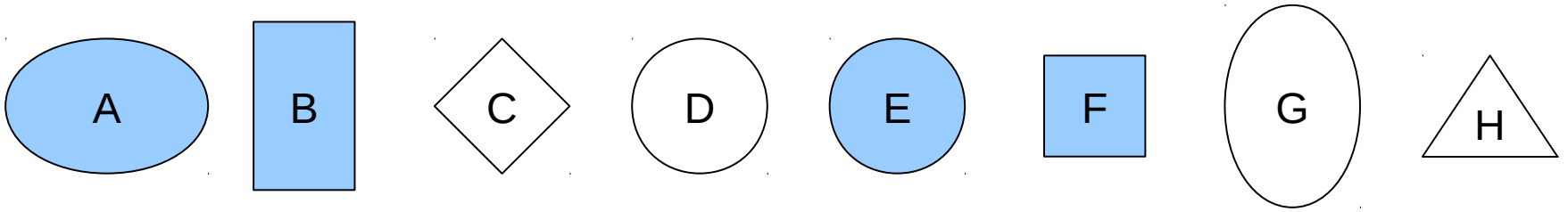
Une espèce phylogénétique est un groupe présentant une combinaison unique de caractères.

Comment estimer les relations évolutives entre plusieurs espèces ?

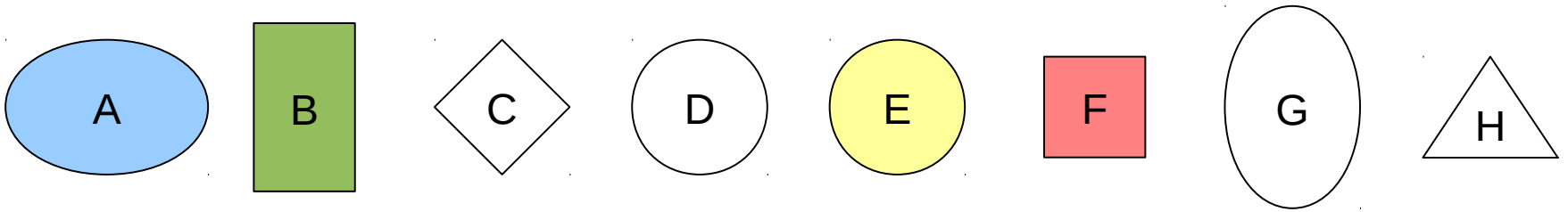


Quelles sont les deux espèces les plus proches ?

La notion de caractère et d'états possible pour un caractère



Couleur	1	1	0	0	1	1	0	0
Forme	0	1	1	0	0	1	0	1

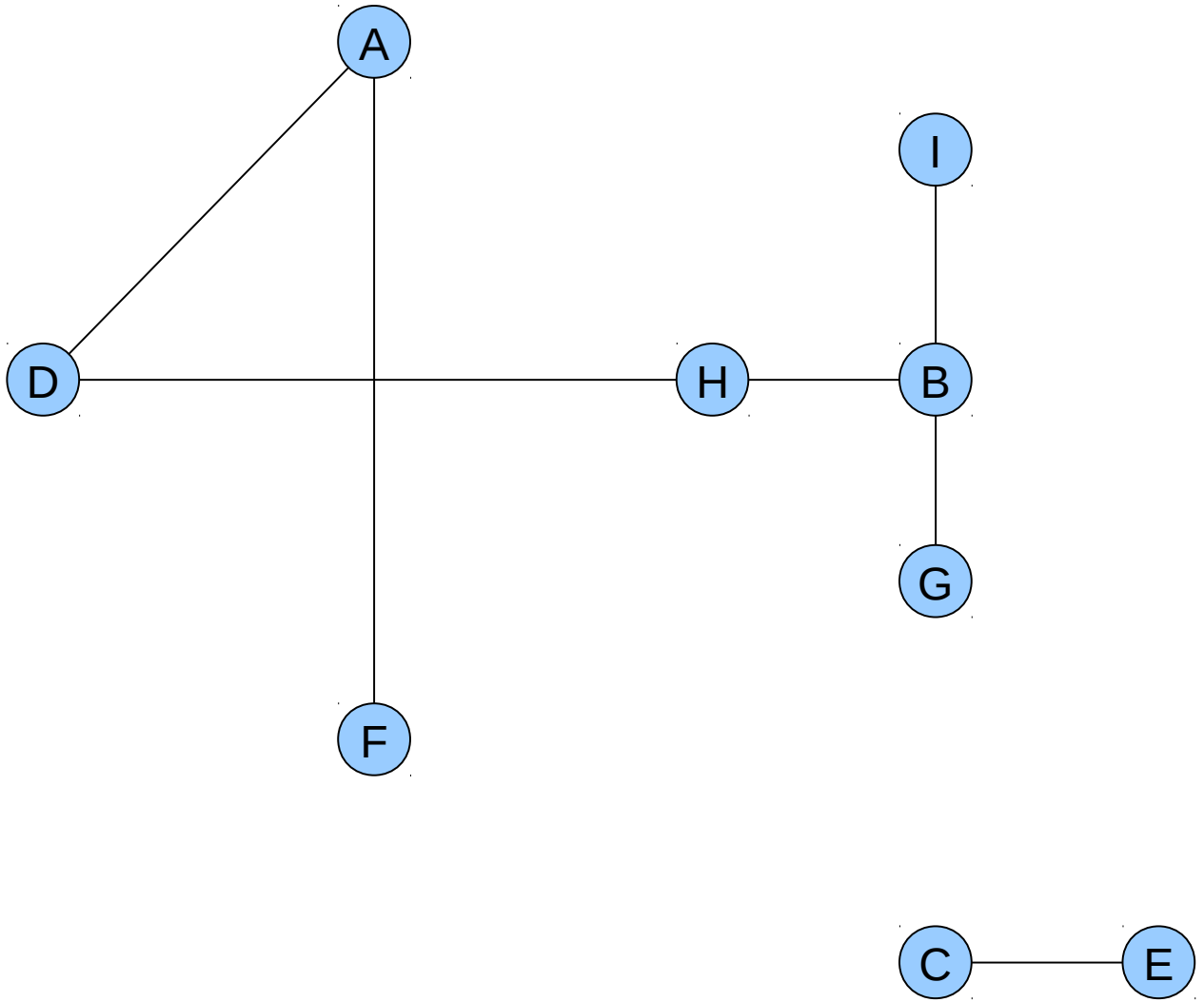


Couleur	1	2	0	0	3	4	0	0
Forme	1	1	1	0	0	1	0	1

Phylogenetic analyses

```
seq 1 MARSATTACKS          -MARS-AT-TACKS
seq 2 CMAAATRTSKS          CMA--AATRTS-KS
seq 3 MARTIANS             -MARTIAN-----S
```

Comment représenter les relations évolutives
entre plusieurs espèces ?


A
I
D
H
B
G
F
C
E

Jean-Baptiste Pierre Antoine de Monet
Chevalier de Lamarck
(Bazentin, 1744 – Paris, 1829)

PHILOSOPHIE
ZOOLOGIQUE,
OU
EXPOSITION

Des Considérations relatives à l'histoire naturelle des Animaux ; à la diversité de leur organisation et des facultés qu'ils en obtiennent ; aux causes physiques qui maintiennent en eux la vie et donnent lieu aux mouvemens qu'ils exécutent ; enfin, à celles qui produisent, les unes le sentiment, et les autres l'intelligence de ceux qui en sont doués ;

PAR J.-B.-P.-A. LAMARCK,

Professeur de Zoologie au Muséum d'Histoire Naturelle, Membre de l'Institut de France et de la Légion d'Honneur, de la Société Philomatique de Paris, de celle des Naturalistes de Moscou, Membre correspondant de l'Académie Royale des Sciences de Munich, de la Société des Amis de la Nature de Berlin, de la Société Médicale d'Emulation de Bordeaux, de celle d'Agriculture, Sciences et Arts de Strasbourg, de celle d'Agriculture du département de l'Oise, de celle d'Agriculture de Lyon, Associé libre de la Société des Pharmaciens de Paris, etc.

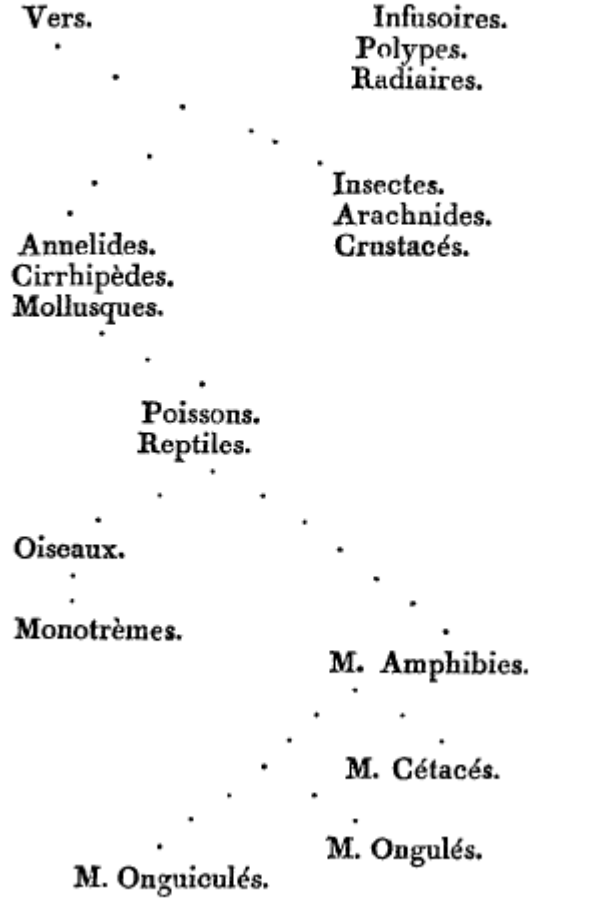
TOME PREMIER.

A PARIS,

Chez { DENTU, Libraire, rue du Pont de Lodi, N°. 3 ;
L'AUTEUR, au Muséum d'Histoire Naturelle (Jardin des Plantes).

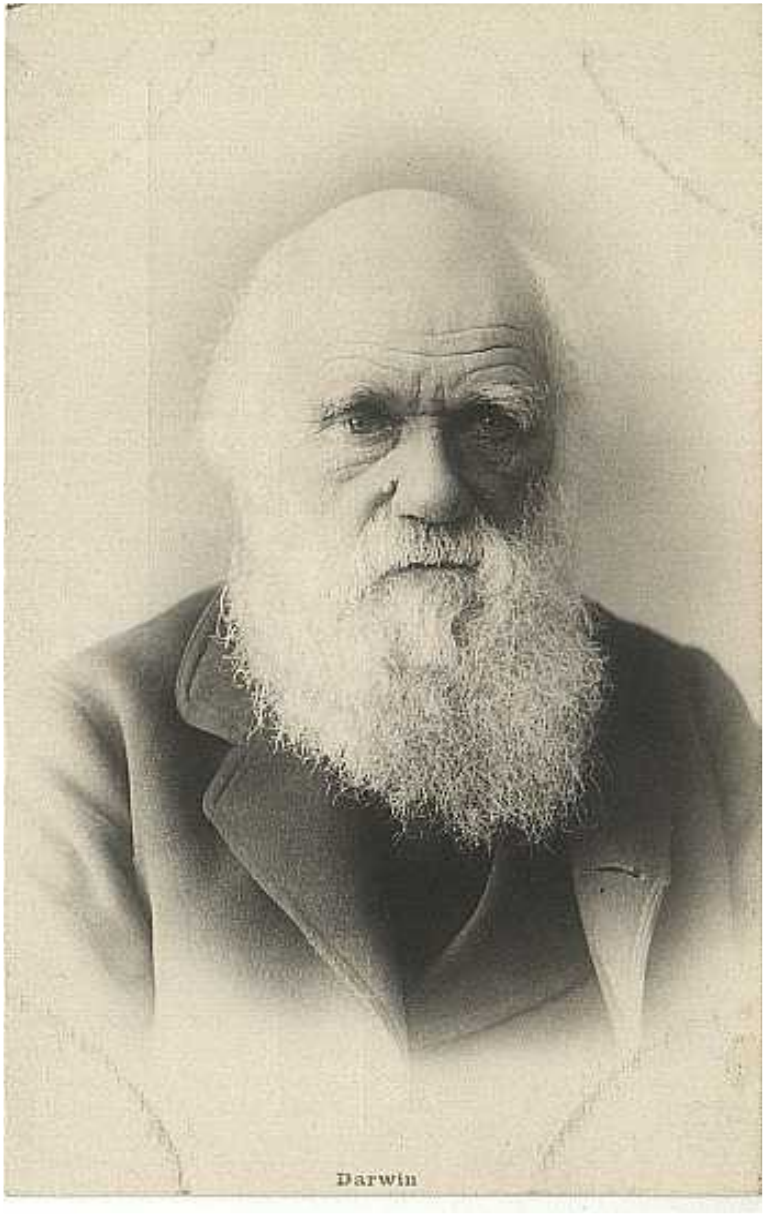
M. DCCC. IX.

TABLEAU
Servant à montrer l'origine des différens animaux.



Cette série d'animaux commençant par deux

Charles Robert Darwin
(Shrewsbury 1809- Downe 1882)



ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

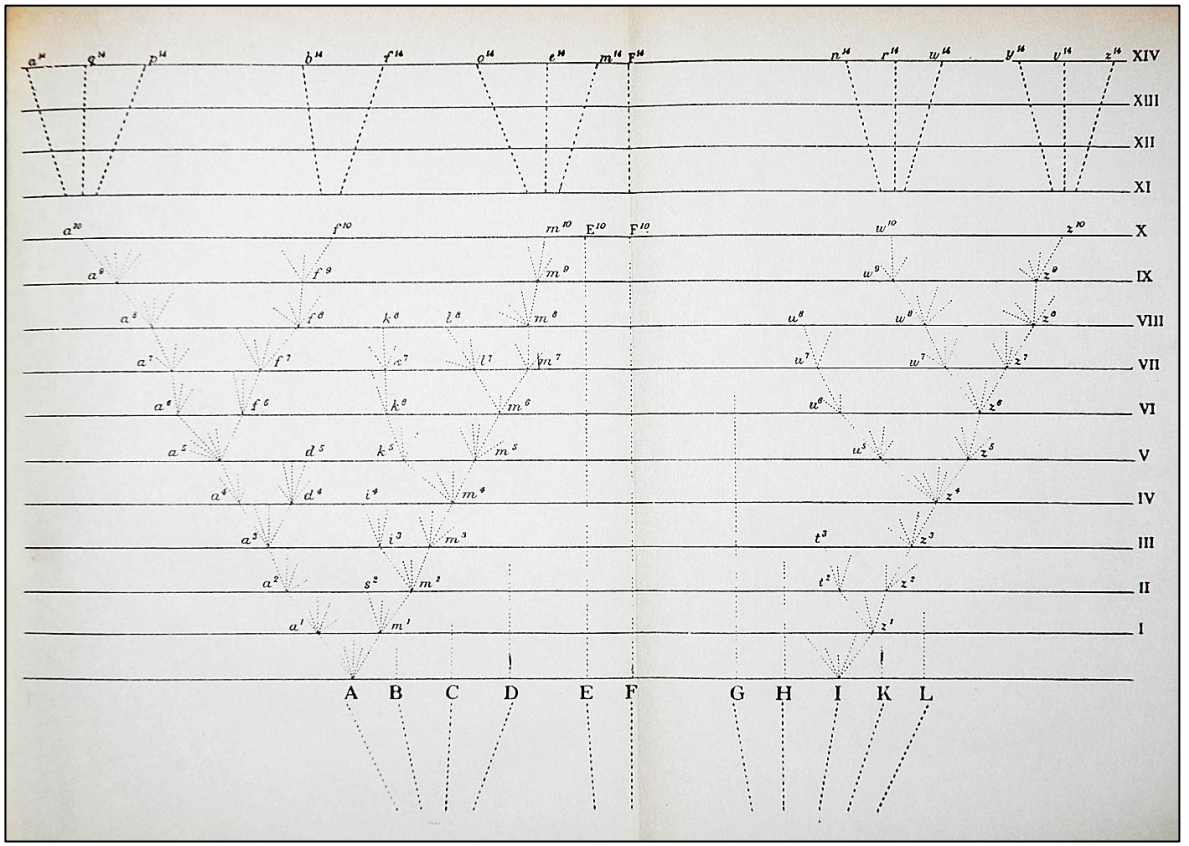
FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;

AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

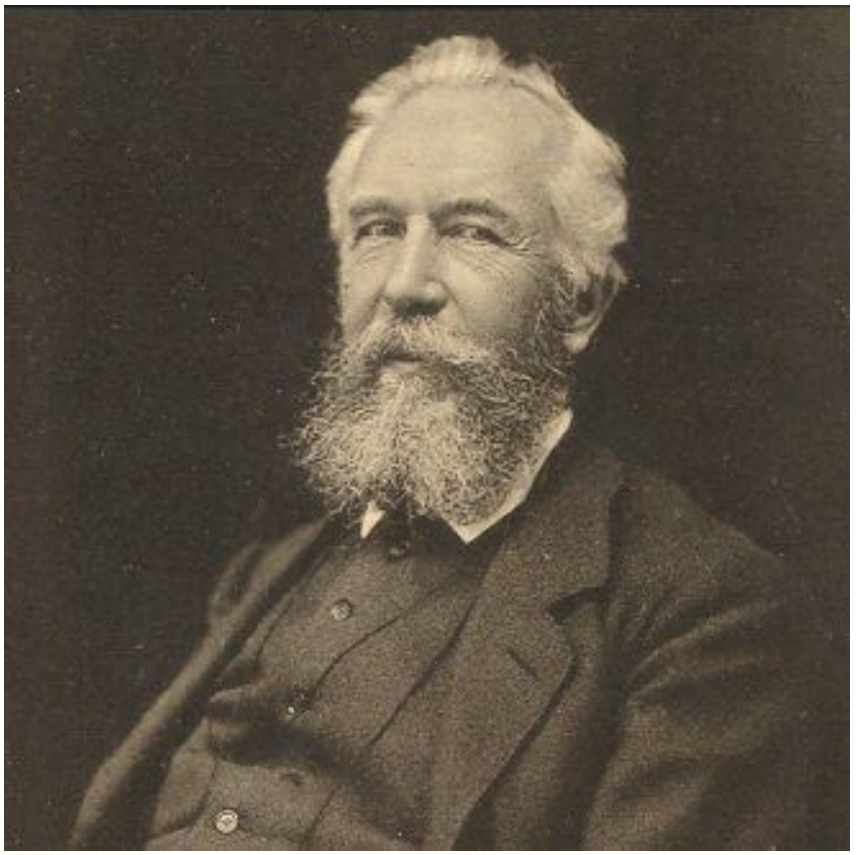
LONDON:

JOHN MURRAY, ALBEMARLE STREET.

1859.



Ernst Heinrich Philipp August Haeckel
(Postdam 1834 – Iena 1919)



GENERELLE MORPHOLOGIE
DER ORGANISMEN.

ALLGEMEINE GRUNDZÜGE

DER ORGANISCHEN FORMEN-WISSENSCHAFT,

MECHANISCH BEGRÜNDET DURCH DIE VON

CHARLES DARWIN

REFORMIRTE DESCENDENZ-THEORIE,

VON

ERNST HAECKEL.

ERSTER BAND:

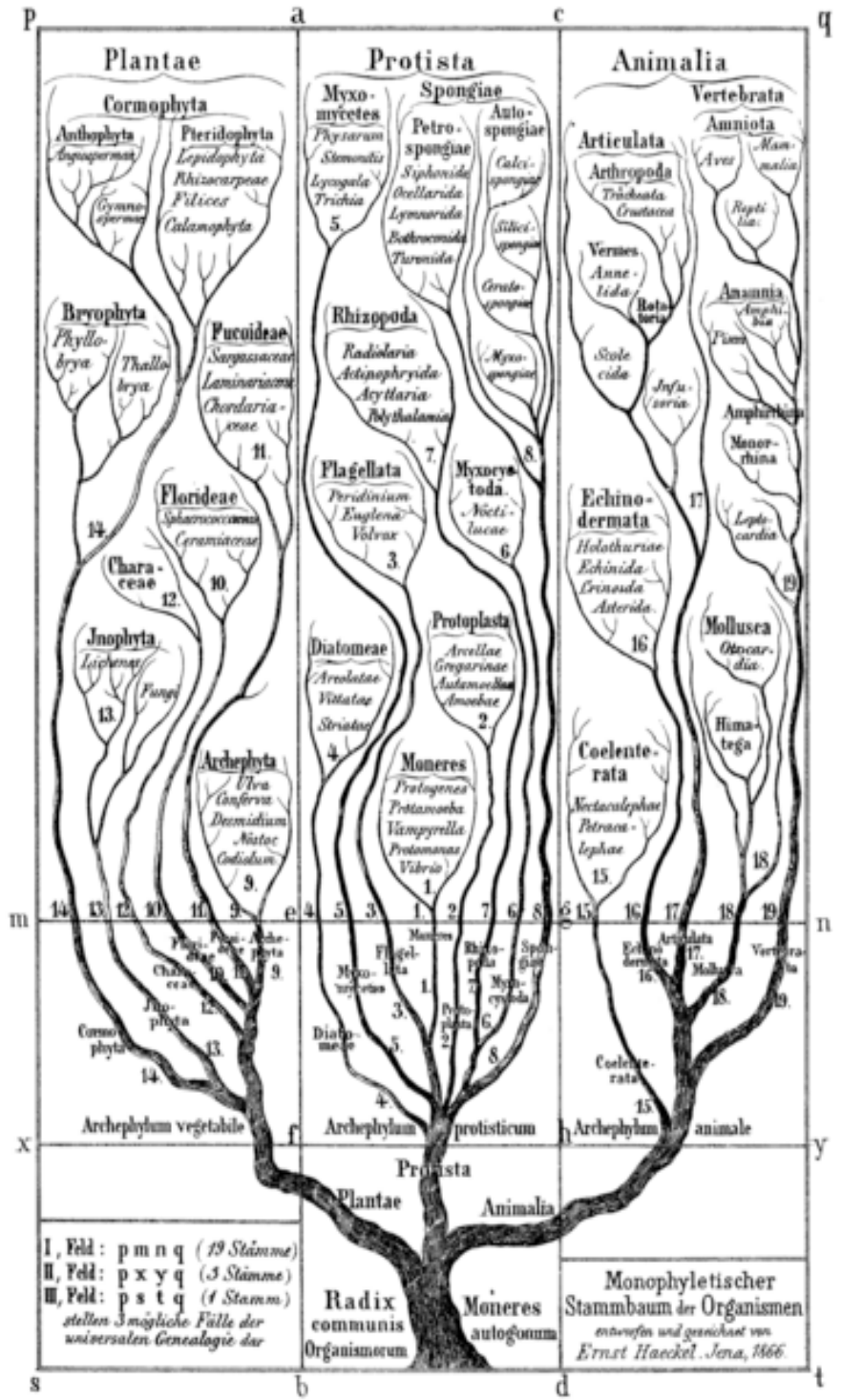
ALLGEMEINE ANATOMIE

DER ORGANISMEN.

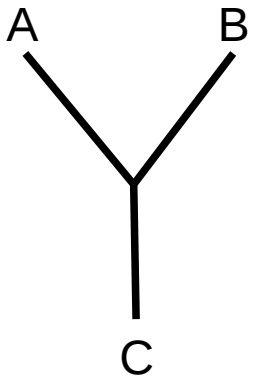
„E PUR SI MUOVE!"

MIT ZWEI PROMORPHOLOGISCHEN TAFELN.

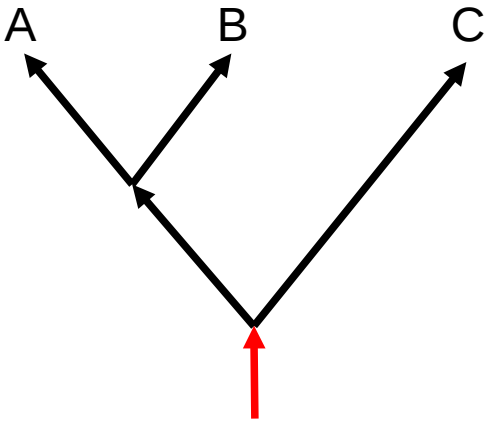
BERLIN.
DRUCK UND VERLAG VON GEORG REIMER.
1866.



Différents types d'arbres pour schématiser les relations évolutives entre différentes entités

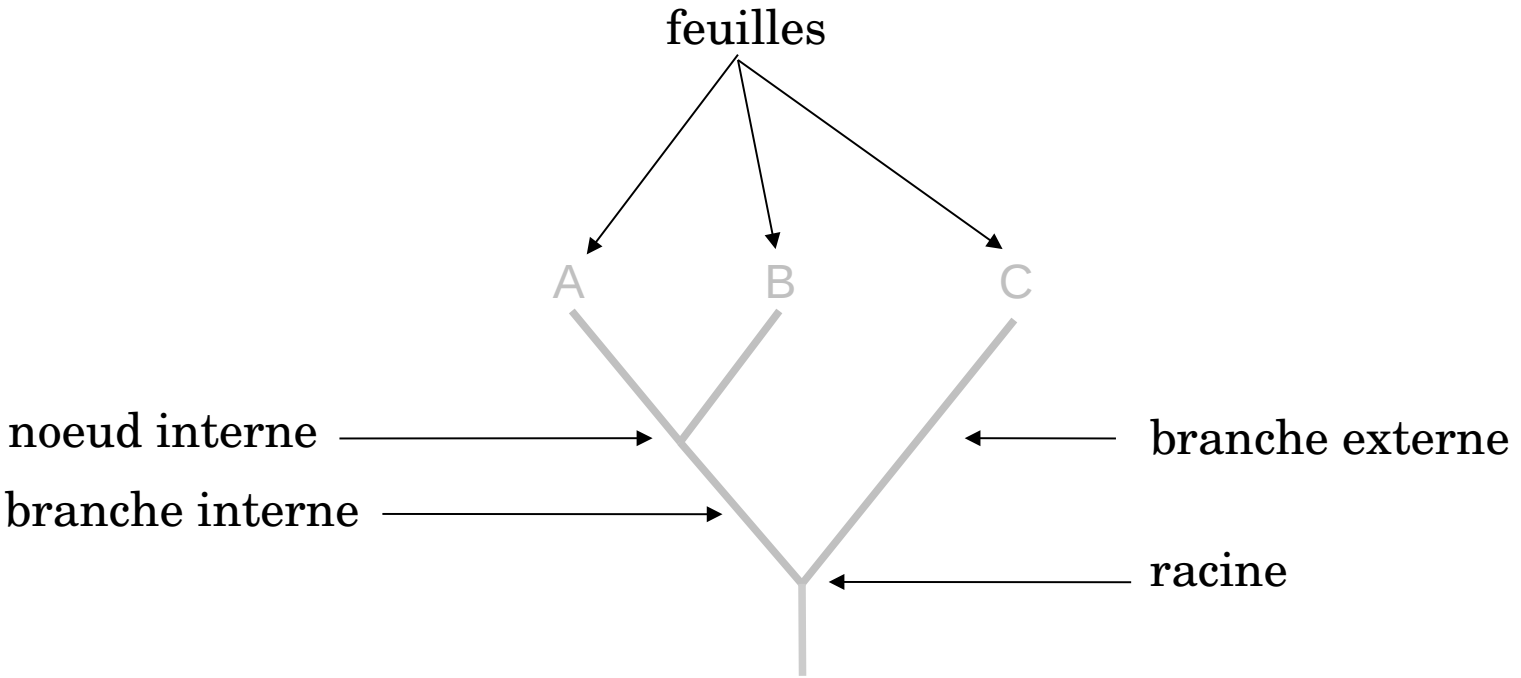


Arbre non raciné
(graphe connexe non cyclique)



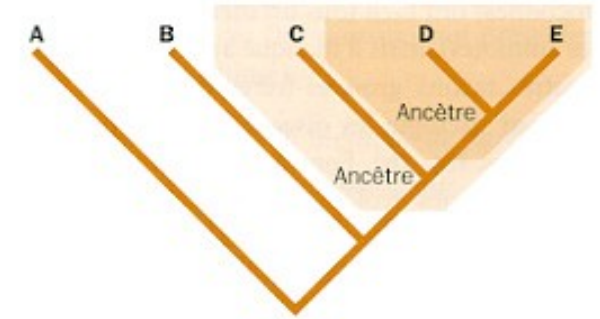
Arbre raciné
(graphe non cyclique et orienté)

Quelques définitions



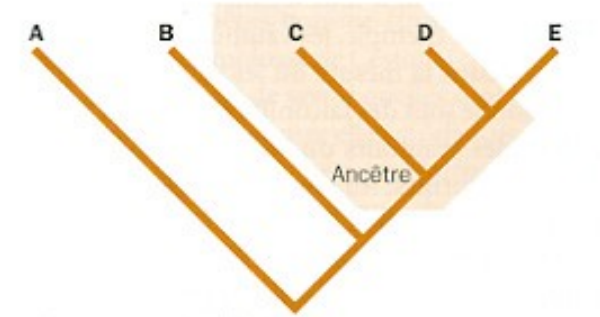
1 arbre = 1 topologie + des longueurs de branches

Monophylie, Paraphylie et Polyphylie



a. Groupes monophylétiques

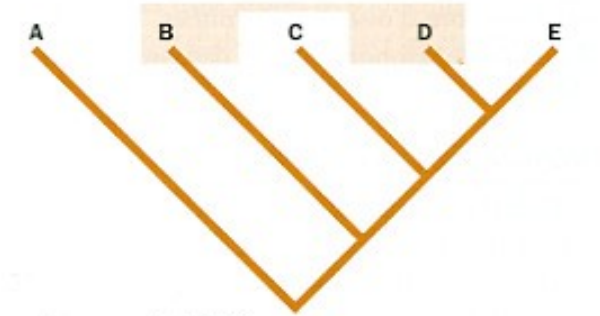
Un ancêtre et tous ses descendants



b. Groupe paraphylétique

Un ancêtre et une partie de ses descendants

ex. poissons, reptiles



c. Groupe polyphylétique

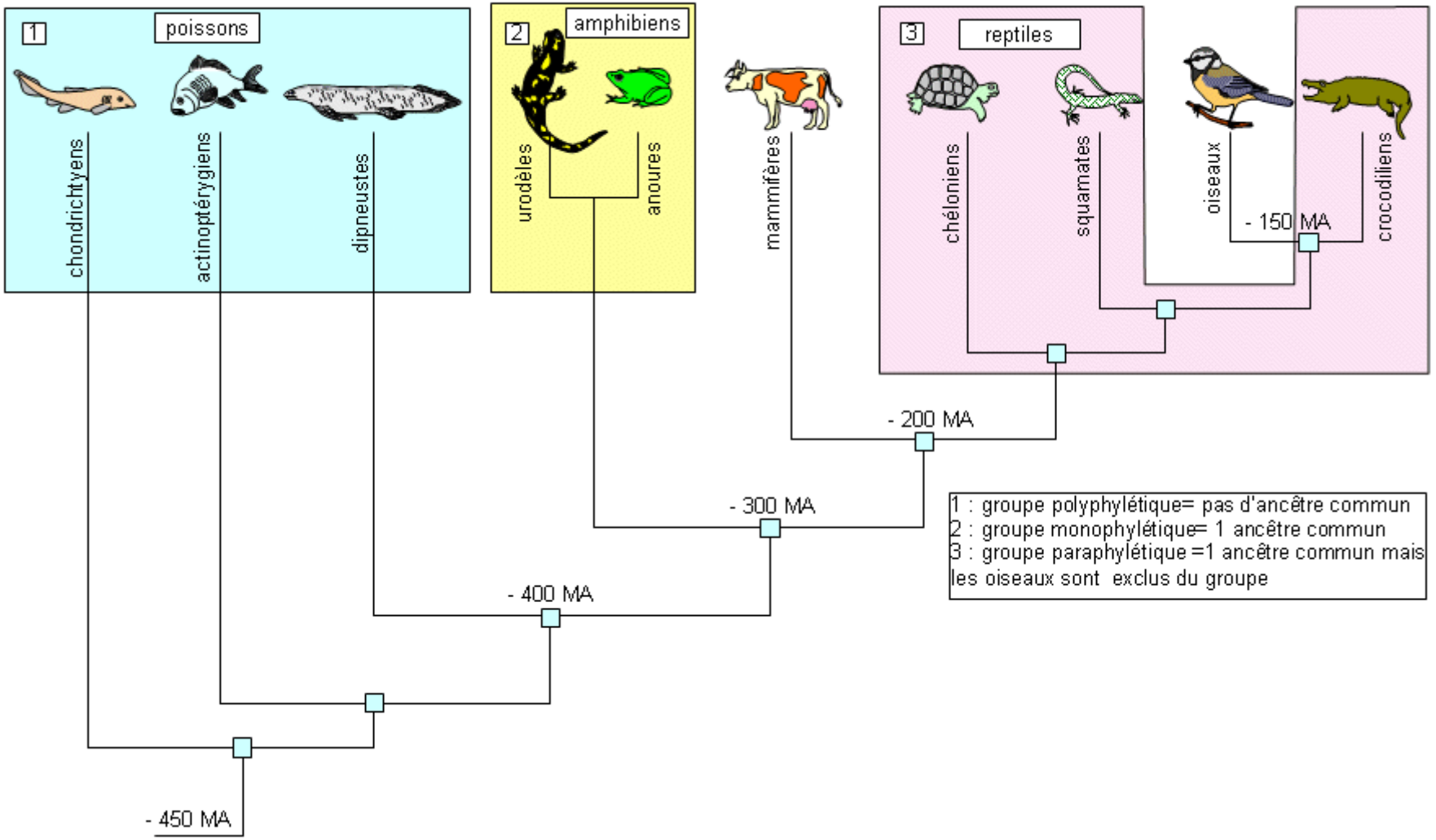
Des membres sans ancêtre

Monophylie, Paraphylie et Polyphylie

Symplésiomorphies
Caractères ancestraux

Synapomorphies
Caractères dérivés partagés

Symplésiomorphies
Caractères ancestraux



d'après Alain Gallien

Etant donnée une liste de caractères associés à un ensemble d'entités, comment construire un arbre retraçant les liens évolutifs entre toutes ces entités ?

Comment proposer un scénario évolutif à partir de l'observation des différences et ressemblances ?

1. Les méthodes de parcimonie
2. Les méthodes phénétiques (de distance)
3. Les méthodes probabilistes (maximum de vraisemblance et Bayesiennes)

Qu'est ce que la parcimonie ?

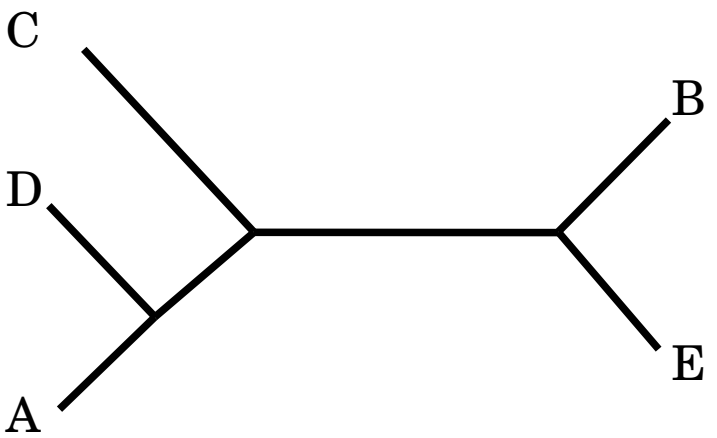
Edwards and Cavalli-Sforza (1963)

Le scénario évolutif proposé nécessite un nombre minimum d'hypothèses.

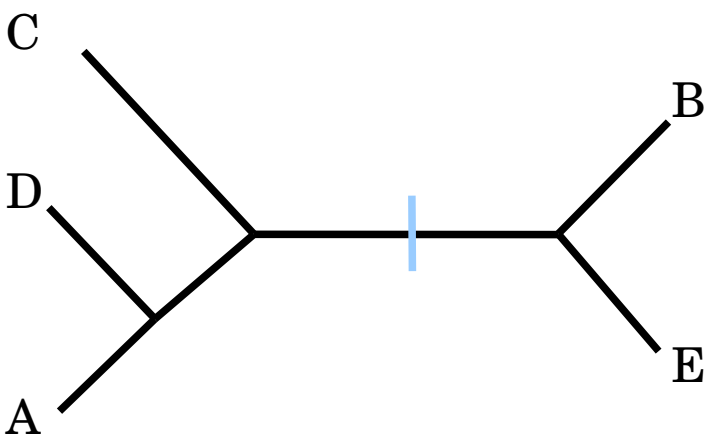
	1	2	3	4	5	6
A	1	0	0	1	1	0
B	0	0	1	0	0	0
C	1	1	0	0	0	0
D	1	1	0	1	1	1
E	0	0	1	1	1	0

Cela revient à proposer un scénario ou le nombre de changements évolutifs est minimal

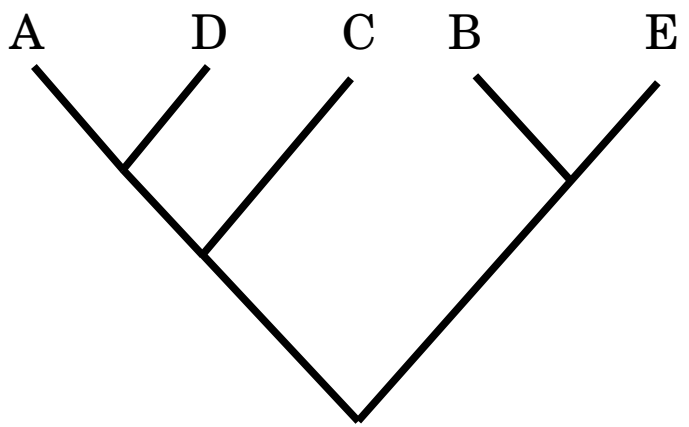
	1	2	3	4	5	6
A	1	0	0	1	1	0
B	0	0	1	0	0	0
C	1	1	0	0	0	0
D	1	1	0	1	1	1
E	0	0	1	1	1	0



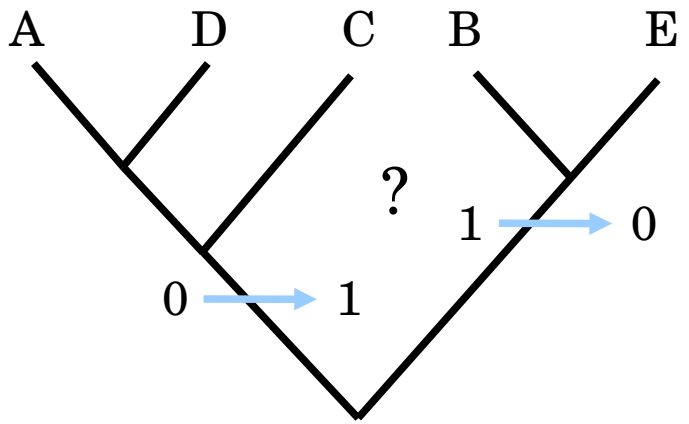
	1	2	3	4	5	6
A	1	0	0	1	1	0
B	0	0	1	0	0	0
C	1	1	0	0	0	0
D	1	1	0	1	1	1
E	0	0	1	1	1	0



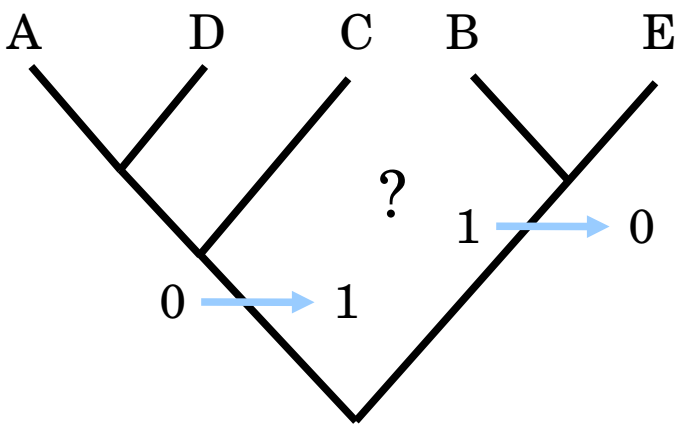
	1	2	3	4	5	6
A	1	0	0	1	1	0
B	0	0	1	0	0	0
C	1	1	0	0	0	0
D	1	1	0	1	1	1
E	0	0	1	1	1	0

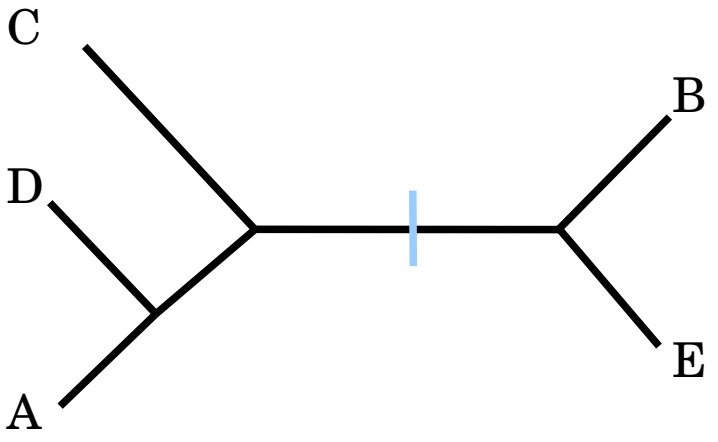


	1	2	3	4	5	6
A	1	0	0	1	1	0
B	0	0	1	0	0	0
C	1	1	0	0	0	0
D	1	1	0	1	1	1
E	0	0	1	1	1	0

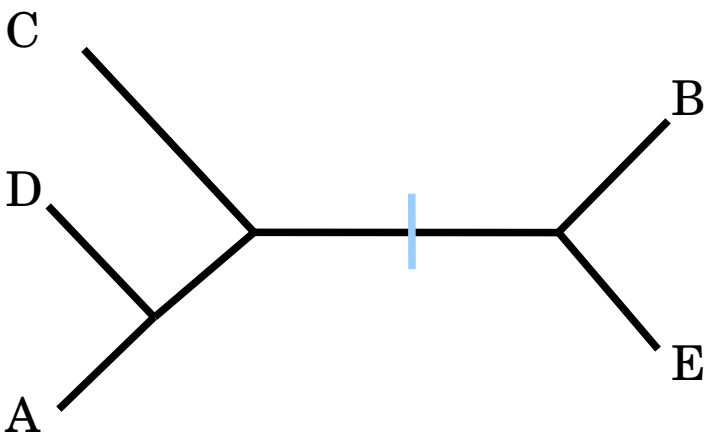


	1	2	3	4	5	6
A	1	0	0	1	1	0
B	0	0	1	0	0	0
C	1	1	0	0	0	0
D	1	1	0	1	1	1
E	0	0	1	1	1	0

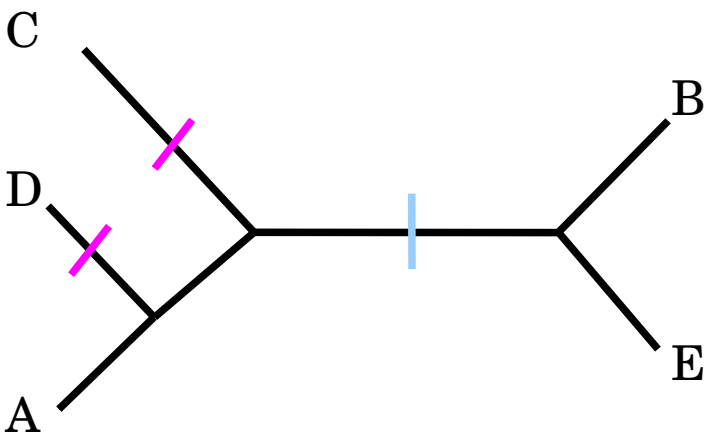




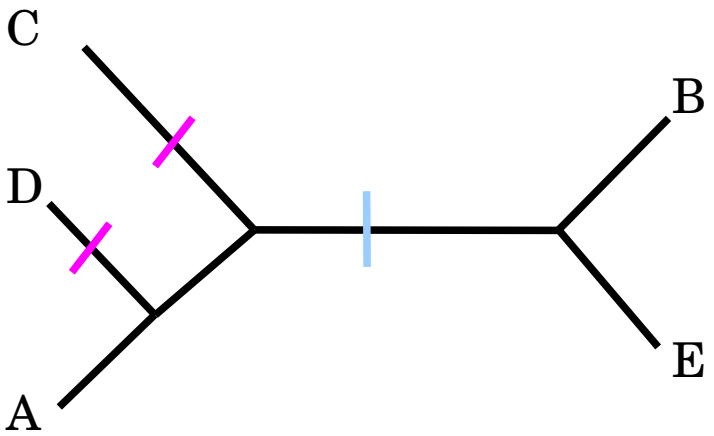
	1	2	3	4	5	6
A	1	0	0	1	1	0
B	0	0	1	0	0	0
C	1	1	0	0	0	0
D	1	1	0	1	1	1
E	0	0	1	1	1	0



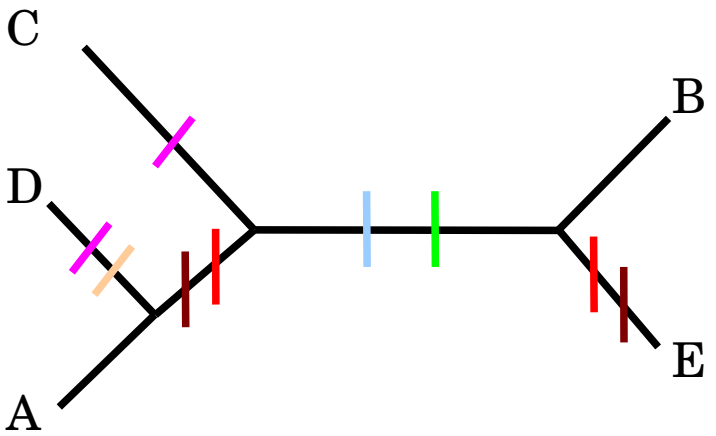
	1	2	3	4	5	6
A	1	0	0	1	1	0
B	0	0	1	0	0	0
C	1	1	0	0	0	0
D	1	1	0	1	1	1
E	0	0	1	1	1	0



	1	2	3	4	5	6
A	1	0	0	1	1	0
B	0	0	1	0	0	0
C	1	1	0	0	0	0
D	1	1	0	1	1	1
E	0	0	1	1	1	0



	1	2	3	4	5	6
A	1	0	0	1	1	0
B	0	0	1	0	0	0
C	1	1	0	0	0	0
D	1	1	0	1	1	1
E	0	0	1	1	1	0



S'agit-il de l'arbre le plus parcimonieux ?

Combien d'arbres non racinés ?

Combien d'arbres non racinés ?

Un arbre avec *n* taxons, possède :

n-3 branches internes
n branches externes

On peut ajouter un *n+1* ième taxon, sur les branches internes ou externes,

soit : (*n*-3)+*n* = 2*n*-3 possibilités

Le nombre total d'arbres avec *n* taxons, T_n est donc égal à :

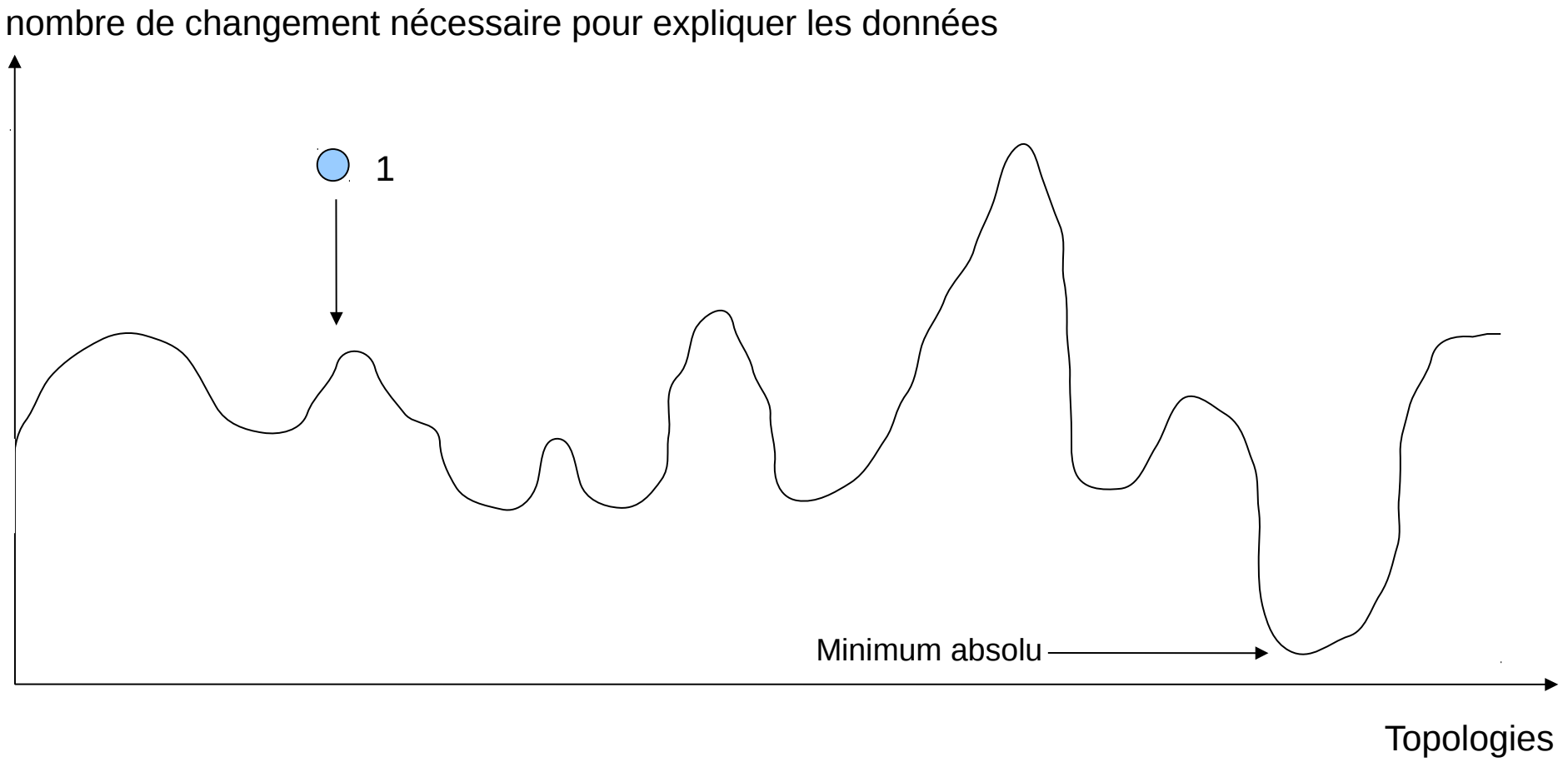
$T_{n-1} * (2(n-1)-3) = T_{n-1} * (2n-5)$

Donc $T_n = \prod_{k=3}^{n} (2n-5)$

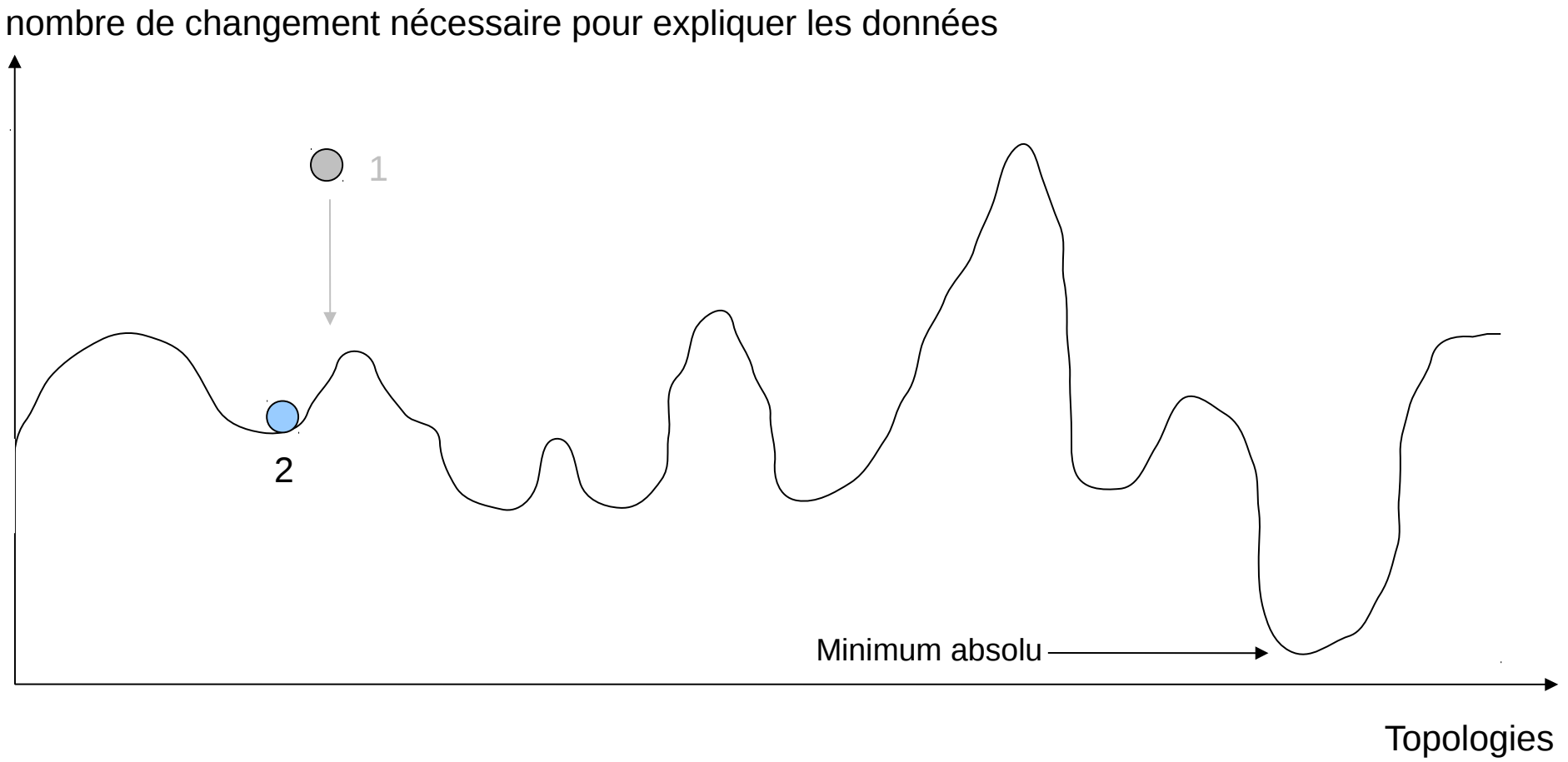
Combien d'arbres non racinés ?

Taxons	Nombre d'arbres	
3	1	
4	3	
5	15	
6	105	
7	945	
8	10395	
9	135135	
10	$2{,}03.10^{6}$	
20	$2{,}22.10^{20}$	← N_A
30	$8{,}69.10^{36}$	
40	$1{,}31.10^{55}$	
50	$2{,}84.10^{74}$	← N_E
100	$1{,}7.10^{182}$	

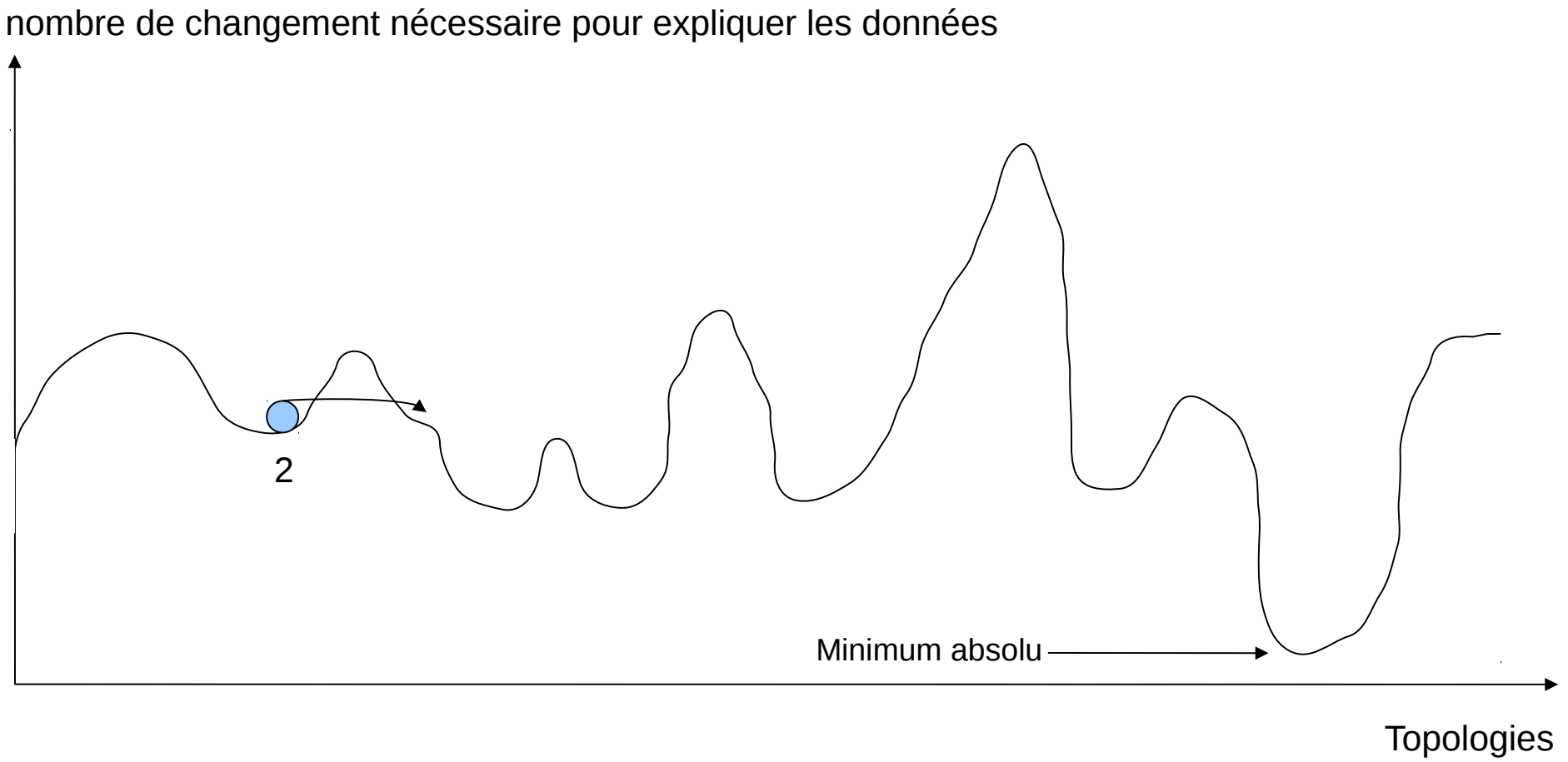
Rechercher le meilleur arbre par les méthodes heuristiques



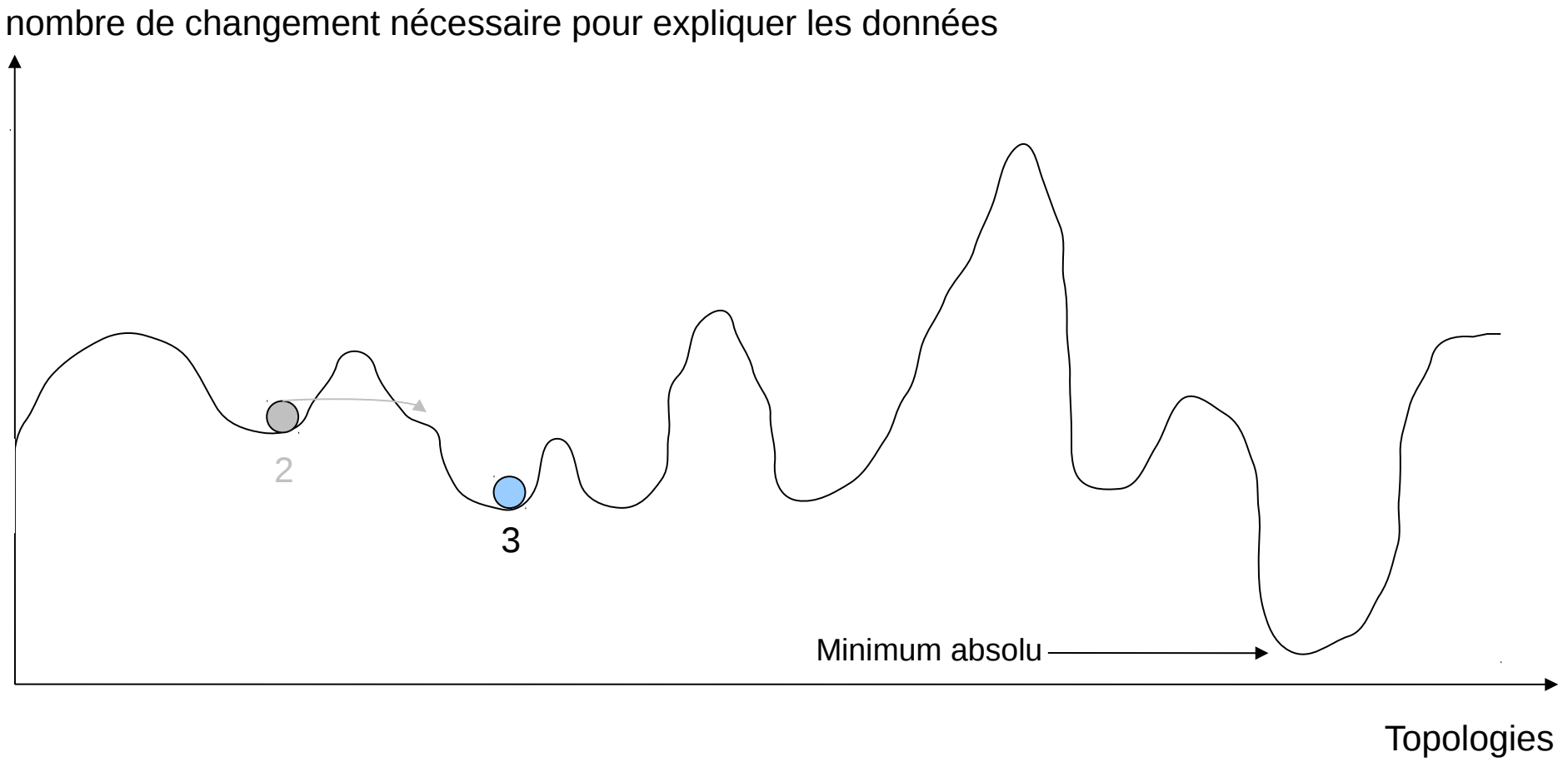
Rechercher le meilleur arbre par les méthodes heuristiques



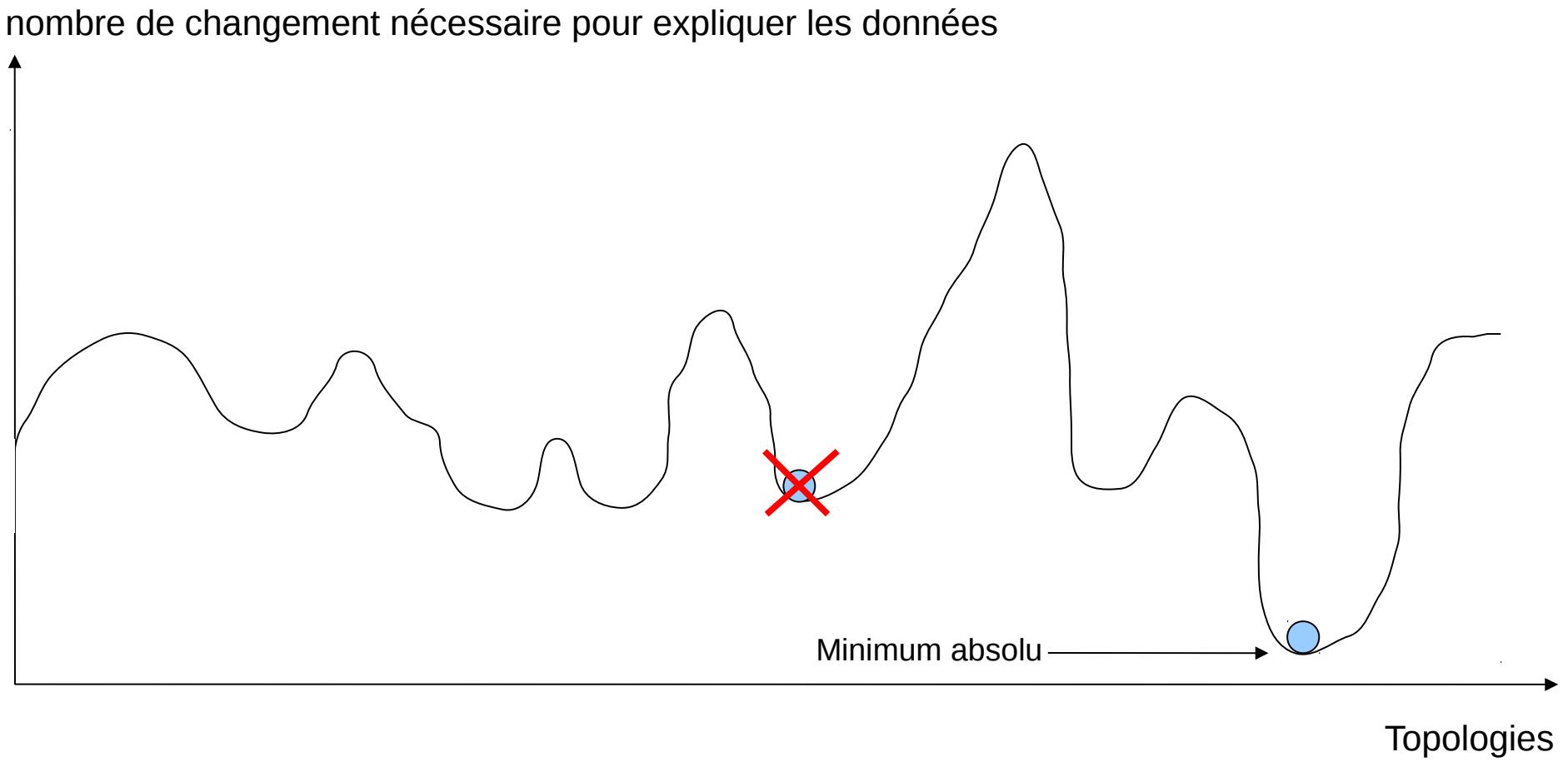
Rechercher le meilleur arbre par les méthodes heuristiques



Rechercher le meilleur arbre par les méthodes heuristiques

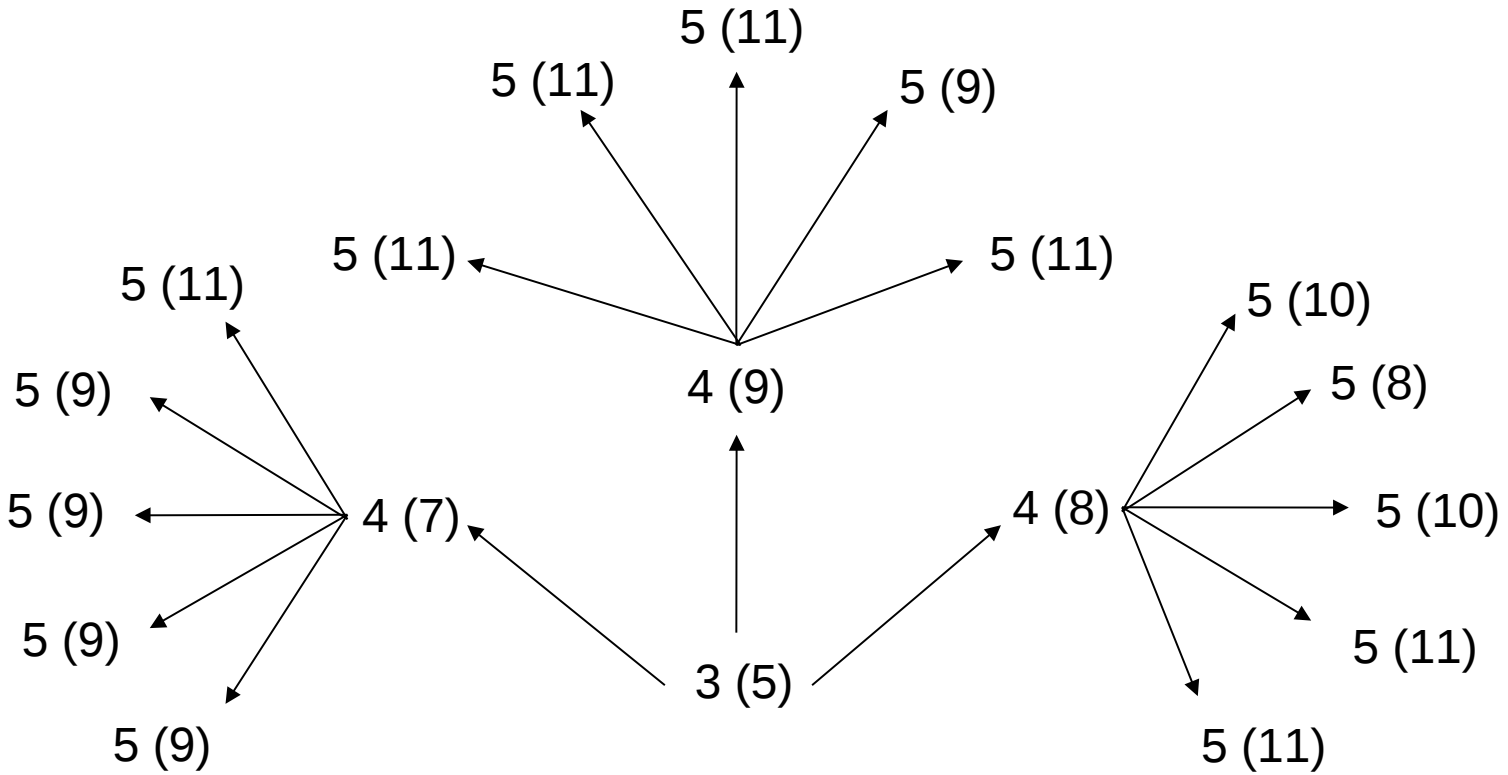


Rechercher le meilleur arbre par les méthodes heuristiques

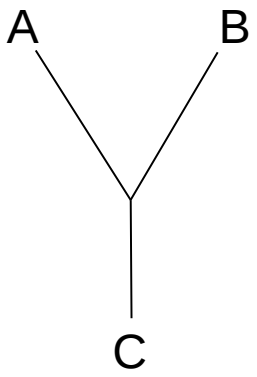


Rechercher le meilleur arbre par les méthodes heuristiques

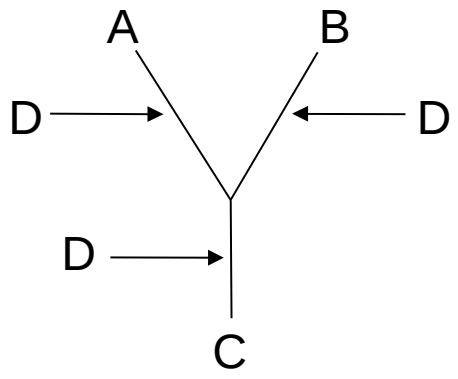
Accélerer la recherche par Branch and Bound



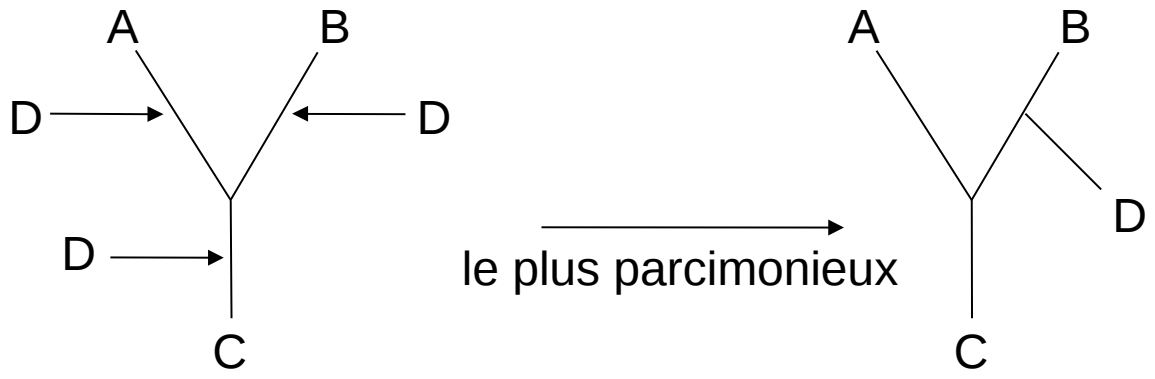
Construire l'arbre de départ par addition séquentielle



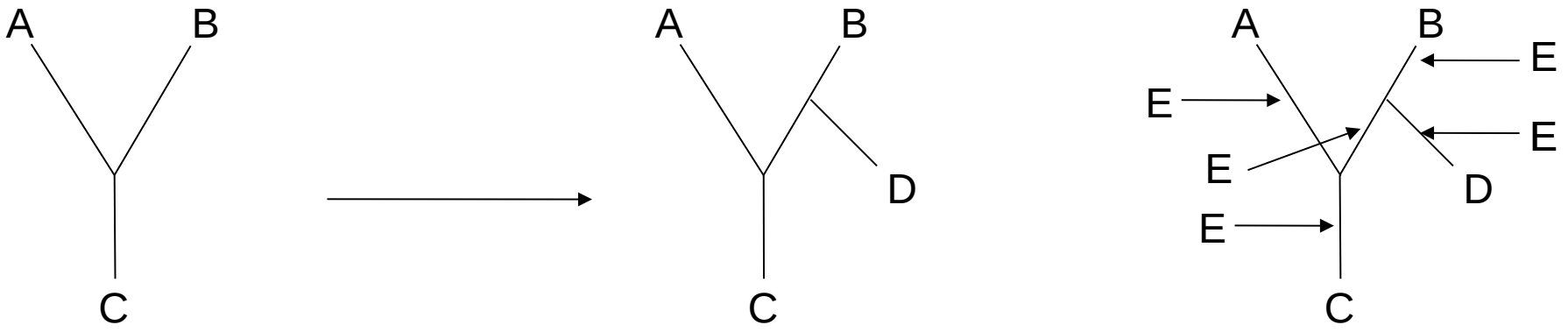
Construire l'arbre de départ par addition séquentielle



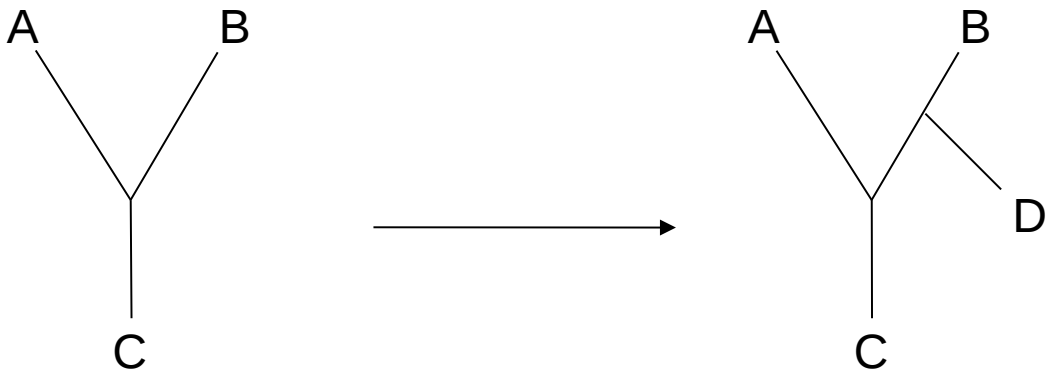
Construire l'arbre de départ par addition séquentielle

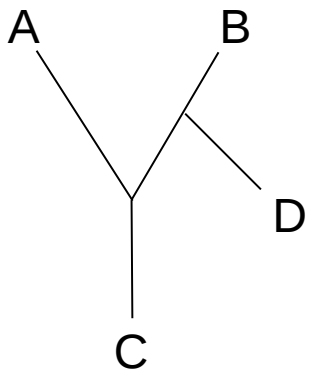


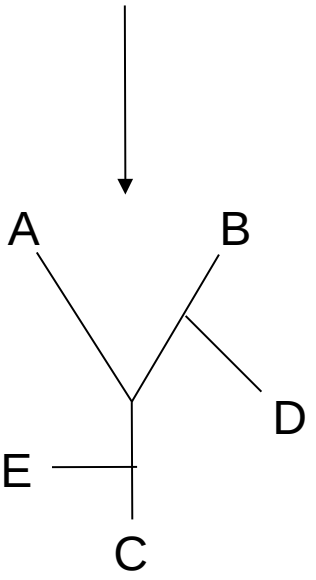
Construire l'arbre de départ par addition séquentielle



Construire l'arbre de départ par addition séquentielle



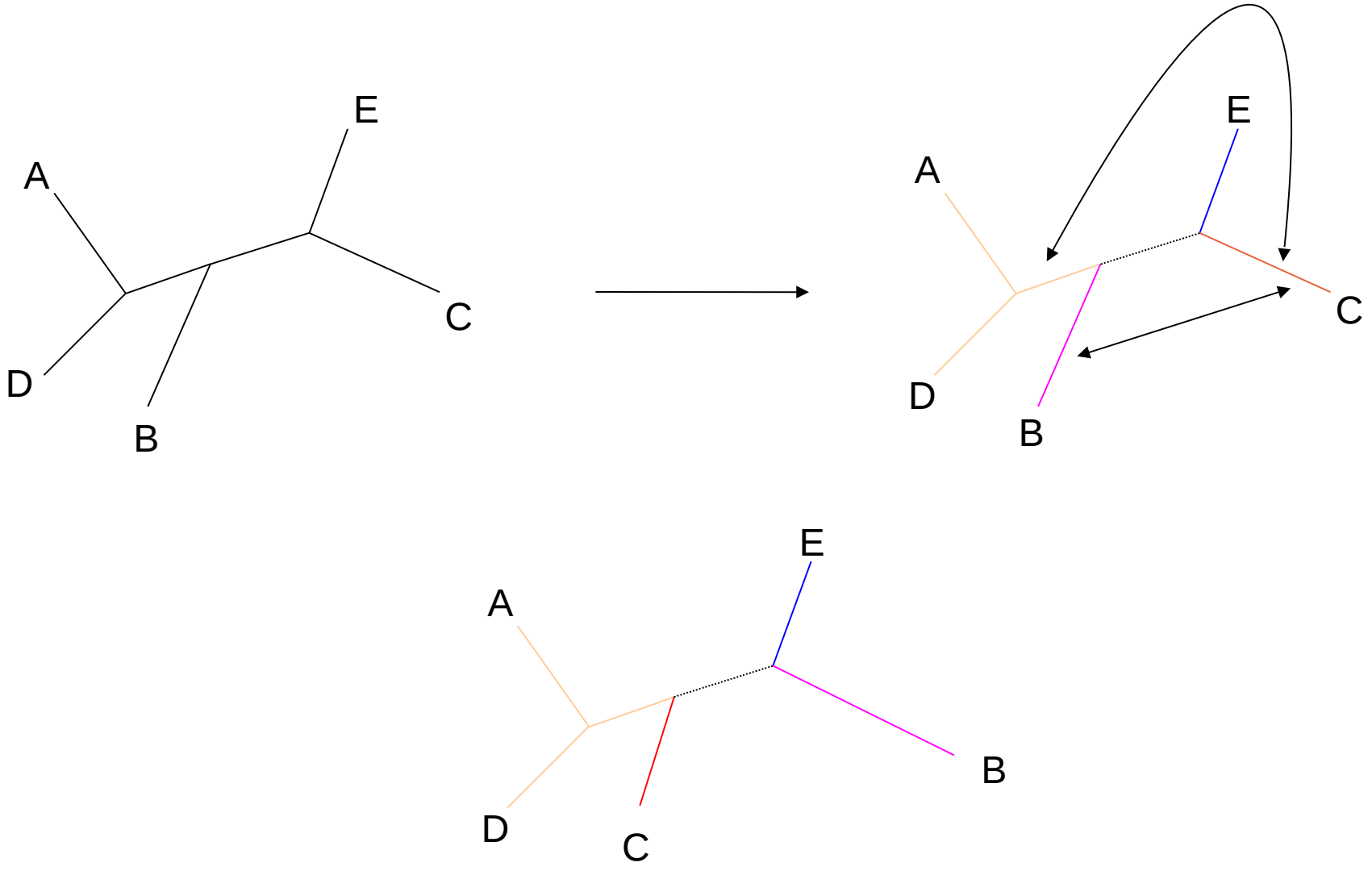




etc...

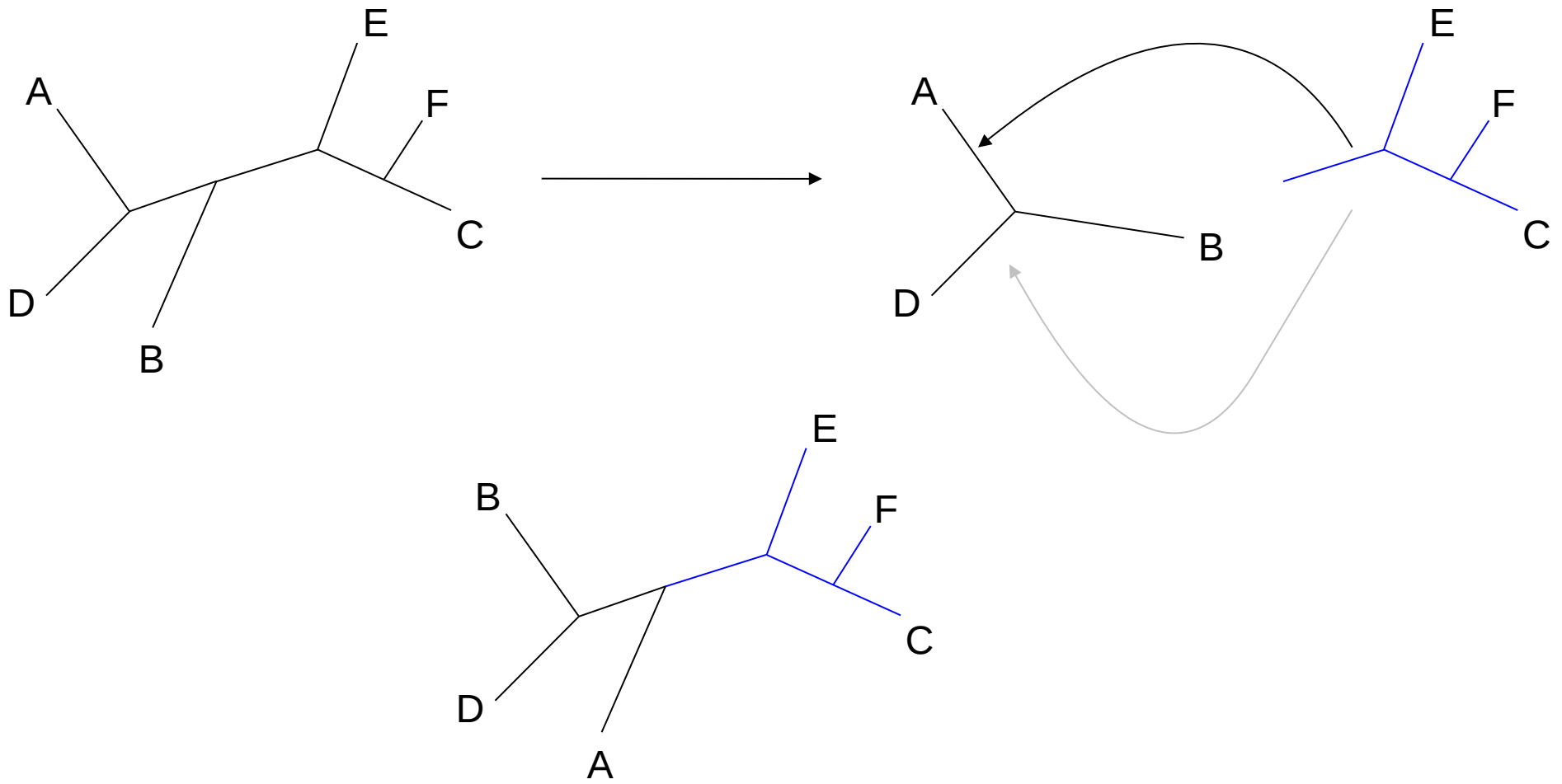
Rechercher le meilleur arbre par les méthodes heuristiques

NNI : Nearest Neighbour Interchange



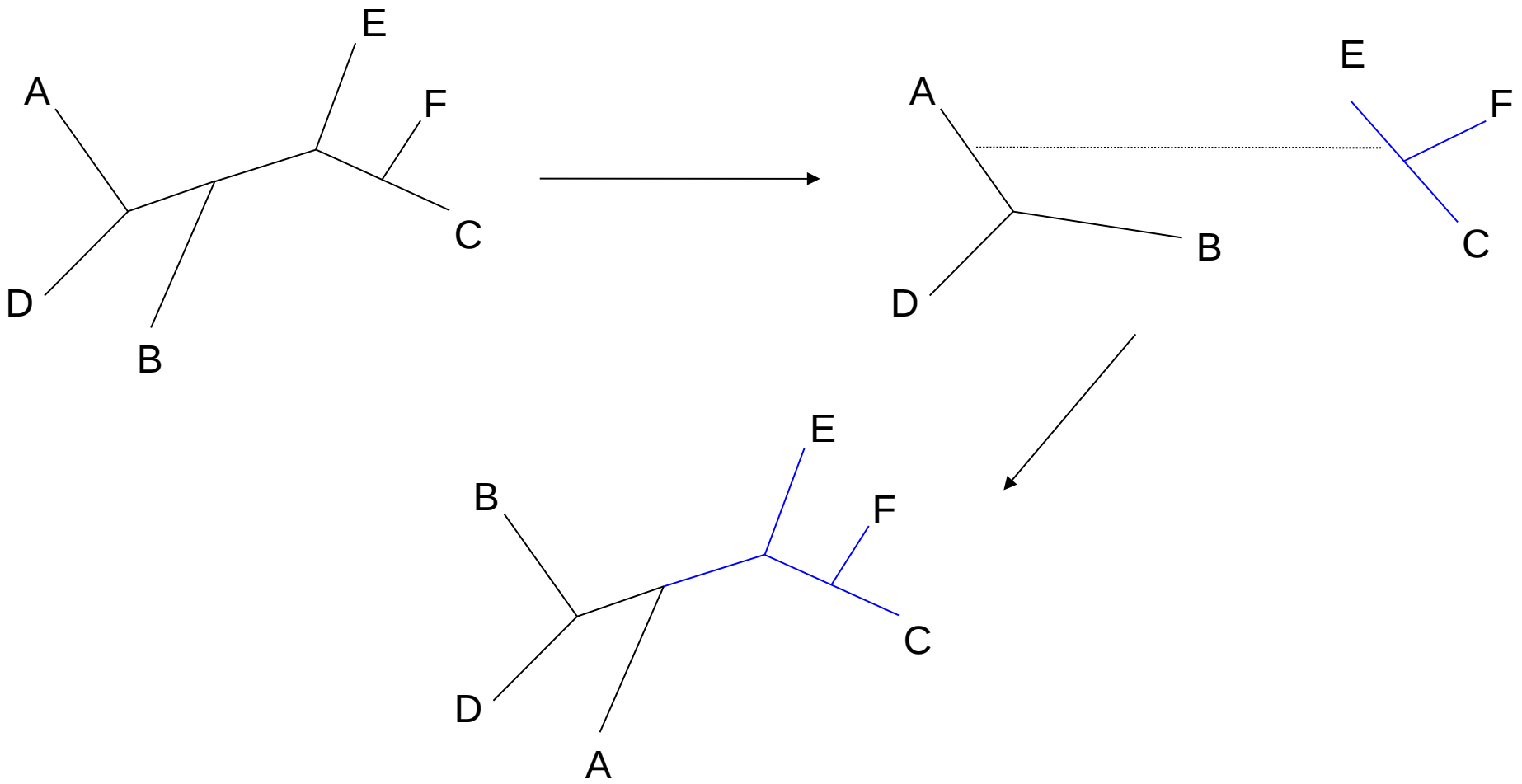
Rechercher le meilleur arbre par les méthodes heuristiques

SPR : Subtree Pruning and Regrafting

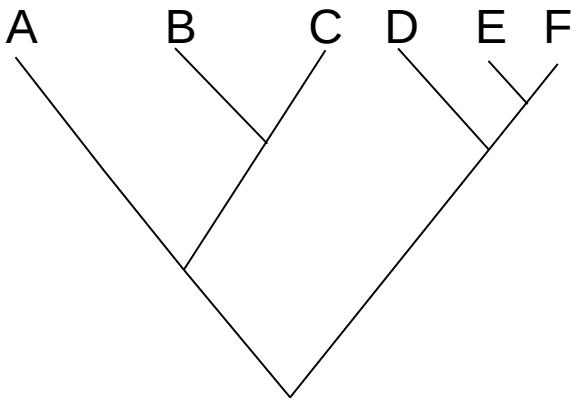


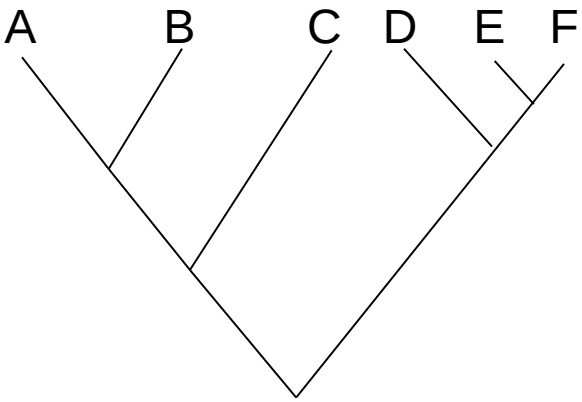
Rechercher le meilleur arbre par les méthodes heuristiques

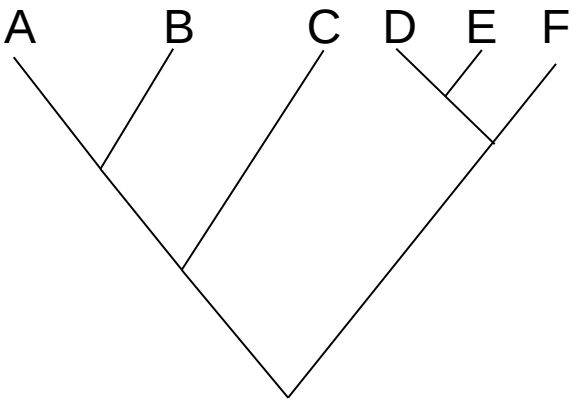
TBR : Tree bisection and reconnection

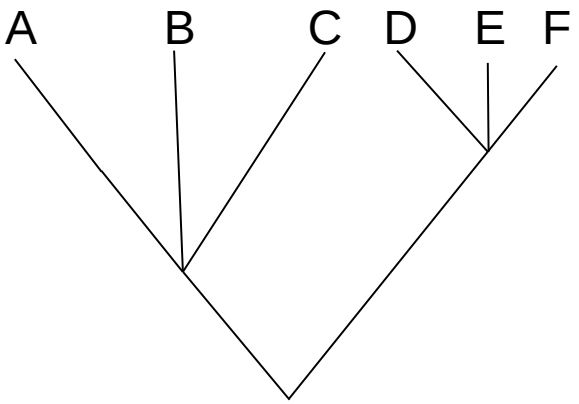


Arbres de consensus

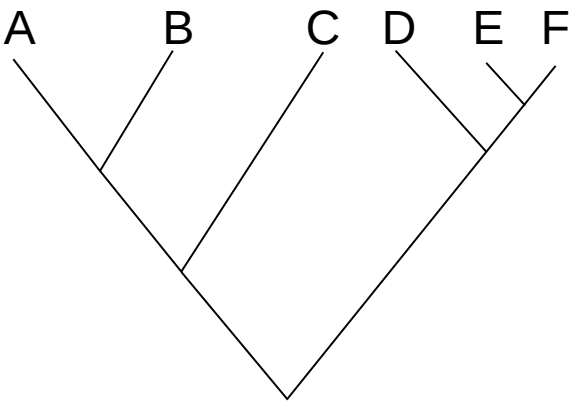








strict



50% majority rule

Etant donnée une liste de caractères associés à un ensemble d'entités, comment construire un arbre retraçant les liens évolutifs entre toutes ces entités ?

Comment proposer un scénario évolutif à partir de l'observation des différences et ressemblances ?

1. Les méthodes de parcimonie
2. Les méthodes phénétiques (de distance)
3. Les méthodes probabilistes (maximum de vraisemblance et Bayesiennes)

Similitude et distance

Plus la ressemblance globale entre deux entité est importante, plus la parenté à de chances d'être proche.

Plus la similitude entre deux entités *i* et *j* est forte et plus la distance entre elles d_{ij} est faible.

En outre les distances métriques doivent respecter les propriétés suivantes :

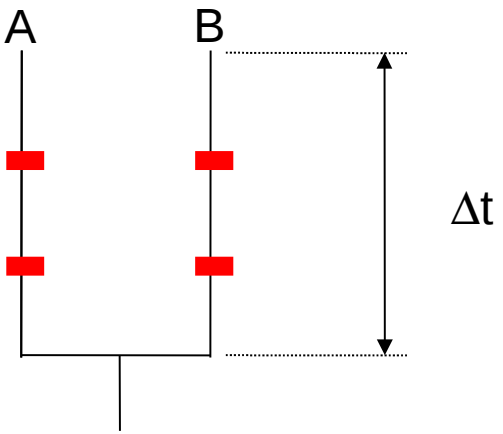
$d_{ij} > 0$ si $i <> j$

$d_{ij} = 0$ si $i = j$

$d_{ij} = d_{ji}$

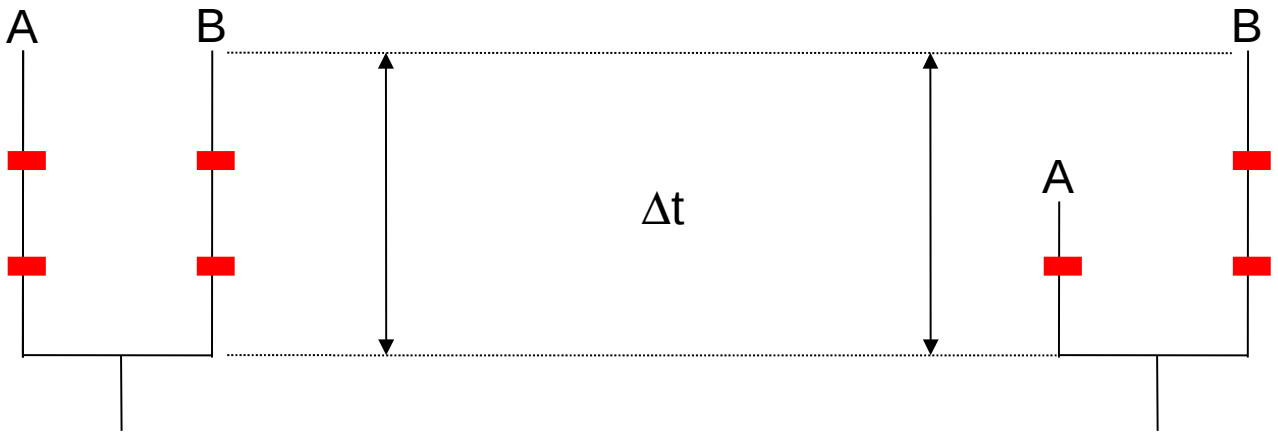
$d_{ik} + d_{kj} >= d_{ij}$

Relation entre temps et distance



d =f(t)

Relation entre distance et temps



$$d = r_i \cdot t$$

D

t

r_i est le taux d'évolution pour l'UE *i*

Calcul de la distance entre deux séquences nucléotidiques

Seq 1 A T T G T A T G T C C T G T A T G C A A

Seq 2 A T T A T A T T T C G T G A A T G C A T

$$d_{(seq1,seq2)} = \frac{\text{\# de substitutions}}{\text{\# de résidus}} = \frac{5}{20} = 0{,}25$$

Calcul de la distance entre deux séquences nucléotidiques

Seq 1 A T T G T A T G T C C T G T A T G C A A

Seq 2 A T T A T A T T T C G T G A A T G C A T

$$d_{(seq1,seq2)} = \frac{\#\text{ de substitutions}}{\#\text{ de résidus}} = \frac{5}{20} = 0{,}25$$

Seq 1 A T T G T A T G T C C T G T A T G C A A

Seq 3 A T T A T A T T T C G T G T A T G C A T

$$d_{(seq1,seq3)} = \frac{\#\text{ de substitutions}}{\#\text{ de résidus}} = \frac{4}{20} = 0{,}20$$

Calcul de la distance entre deux séquences nucléotidiques

Seq 1 A T T G T A T G T C C T G T A T G C A A

Seq 3 A T T A T A T T T C G T G T A T G C A T

$$d_{(seq1,seq3)} = \frac{\#\text{ de substitutions}}{\#\text{ de résidus}} = \frac{4}{20} = 0{,}20$$

ou

Seq 1 A T T G T A T G T C C T G T A T G C A A

A T T A T A T T T C G T G A A T G C A T

Seq 3 A T T A T A T T T C G T G T A T G C A T

$$d_{(seq1,seq3)} = \frac{\#\text{ de substitutions}}{\#\text{ de résidus}} = \frac{6}{20} = 0{,}30$$

Seq 1 A T T G T A T G T C C T G T A T G C A A

Seq 2 A T T A T A T T T C G T G A A T G C A T

$$d_{(seq1,seq2)} = \frac{\text{\# de substitutions}}{\text{\# de résidus}} = \frac{5}{20} = 0,25$$

Seq 1 A T T G T A T G T C C T G T A T G C A A

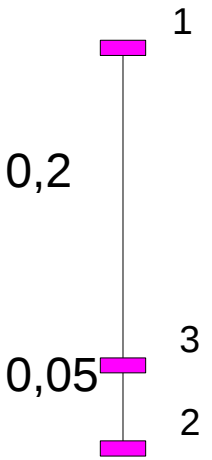
Seq 3 A T T A T A T T T C G T G T A T G C A T

$$d_{(seq1,seq3)} = \frac{\text{\# de substitutions}}{\text{\# de résidus}} = \frac{4}{20} = 0,20$$

Seq 2 A T T A T A T T T C G T G A A T G C A T

Seq 3 A T T A T A T T T C G T G T A T G C A T

$$d_{(seq1,seq3)} = \frac{\text{\# de substitutions}}{\text{\# de résidus}} = \frac{1}{20} = 0,05$$

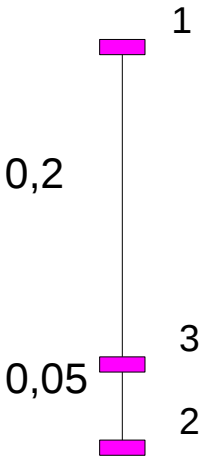


Calcul de la distance moléculaire entre deux séquences nucléotidiques

Seq 1 A T T G T A T G T C C T G T A T G C A A

Seq 3 A T T A T A T T T C G T G T A T G C A T

$$d_{(seq1,seq3)} = \frac{\text{\# de substitutions}}{\text{\# de résidus}} = \frac{4}{20} = 0{,}20$$

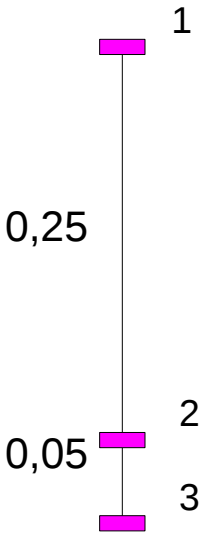


Seq 1 A T T G T A T G T C C T G T A T G C A A

A T T A T A T T T C G T G A A T G C A T

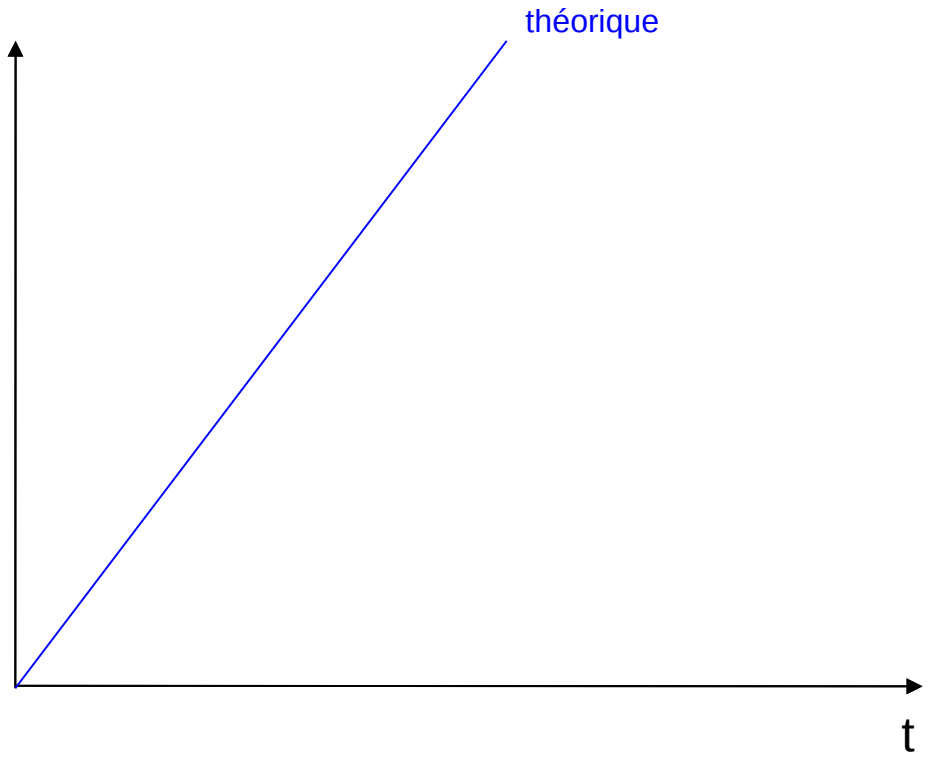
Seq 3 A T T A T A T T T C G T G T A T G C A T

$$d_{(seq1,seq3)} = \frac{\text{\# de substitutions}}{\text{\# de résidus}} = \frac{6}{20} = 0{,}30$$



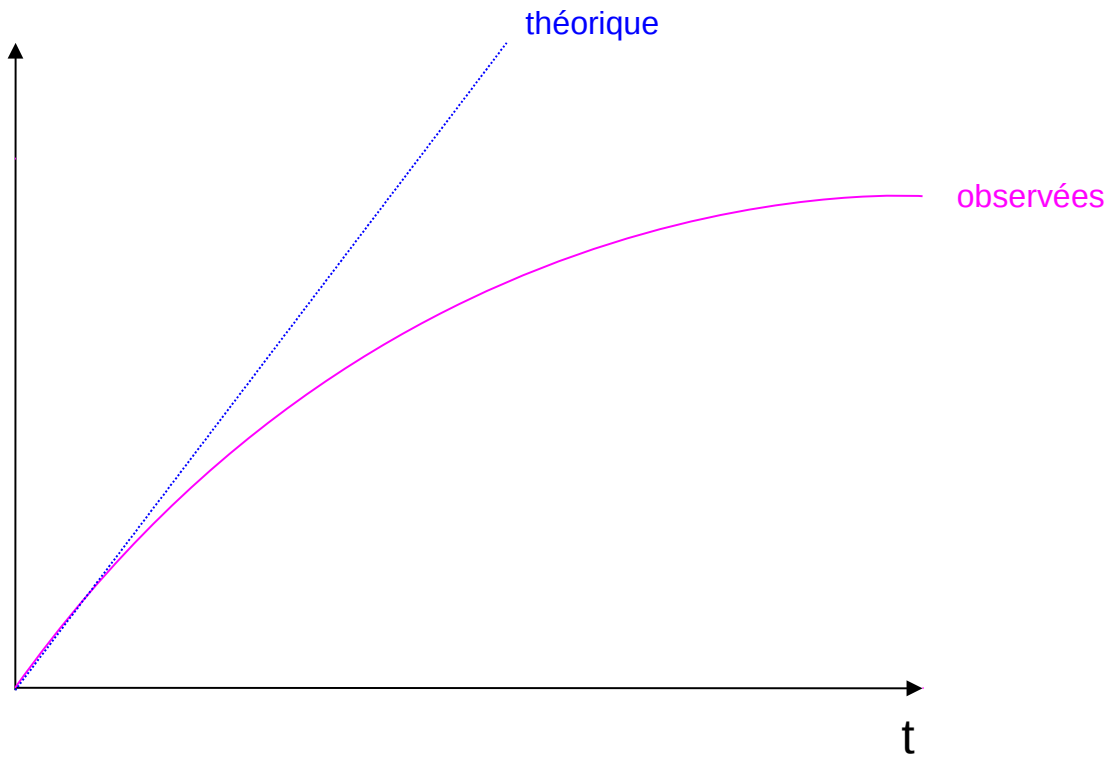
Calcul de la distance entre deux séquences nucléotidiques

nombre de substitutions
entre deux séquences



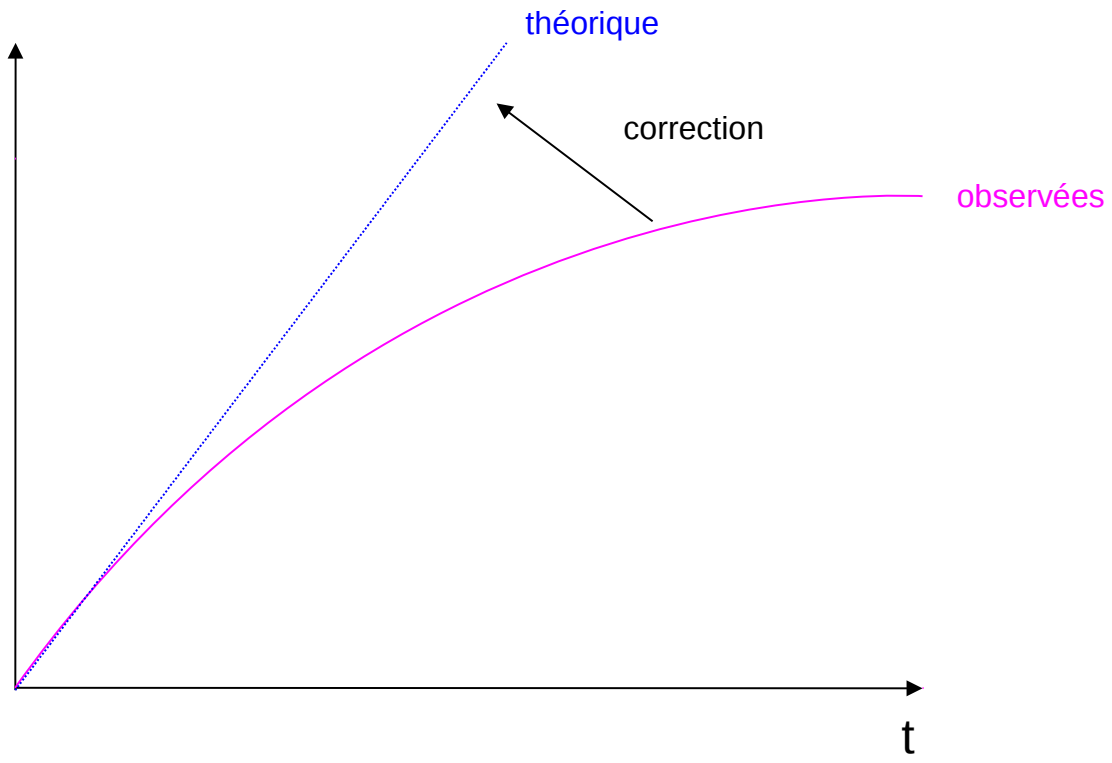
Calcul de la distance entre deux séquences nucléotidiques

nombre de substitutions
entre deux séquences



Calcul de la distance entre deux séquences nucléotidiques

nombre de substitutions
entre deux séquences

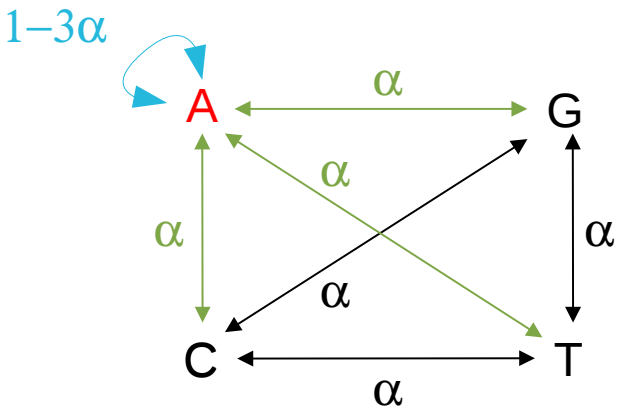


Calcul de la distance entre deux séquences nucléotidiques

Le modèle à un paramètre de Jukes-Cantor (1969) : JC69

Toutes les substitutions sont équiprobables

Les proportions de chacune des 4 bases sont identiques

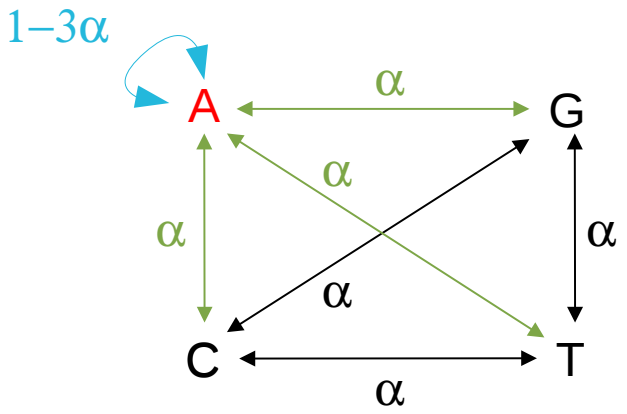


Calcul de la distance entre deux séquences nucléotidiques

Le modèle à un paramètre de Jukes-Cantor (1969) : JC69

Toutes les substitutions sont équiprobables

Les proportions de chacune des 4 bases sont identiques



A → A C/G/T → A

$$p_A(t+1) = p_A(t)(1-3\alpha) + [1-p_A(t)]\alpha = -4\alpha p_A(t) + p_A(t) + \alpha$$

(t discret)

$$\Delta p_A = p_A(t+1) - p_A(t) = -4\alpha p_A(t) + \alpha = \frac{dp_A(t)}{dt}$$

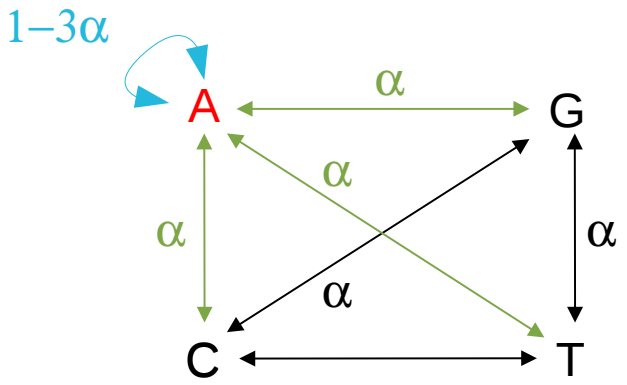
(t continu)

Calcul de la distance entre deux séquences nucléotidiques

Le modèle à un paramètre de Jukes-Cantor (1969) : JC69

Toutes les substitutions sont équiprobables

Les proportions de chacune des 4 bases sont identiques



(t continu) $$\frac{dp_A(t)}{dt} = -4\alpha p_A(t) + \alpha$$

La solution de cette équation différentielle est de la forme :

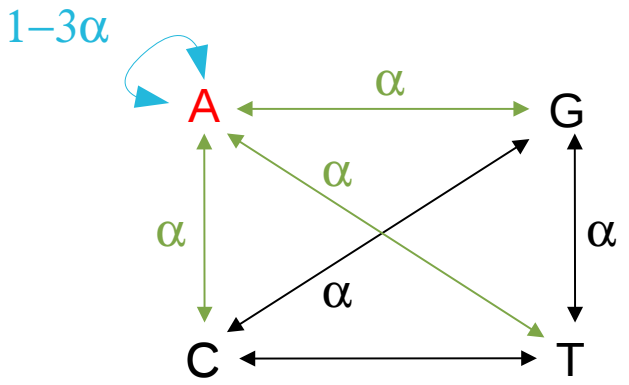
$$P_A(t) = \frac{1}{4} + \left[P_A(0) - \frac{1}{4} \right] e^{-4\alpha t}$$

Calcul de la distance entre deux séquences nucléotidiques

Le modèle à un paramètre de Jukes-Cantor (1969) : JC69

Toutes les substitutions sont équiprobables

Les proportions de chacune des 4 bases sont identiques



$$P_A(t) = \frac{1}{4} + \left[P_A(0) - \frac{1}{4} \right] e^{-4\alpha t}$$

Si $P_A(0) = 1$ alors $P_A(t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} = p_{AA}(t) = p_{ii}(t)$

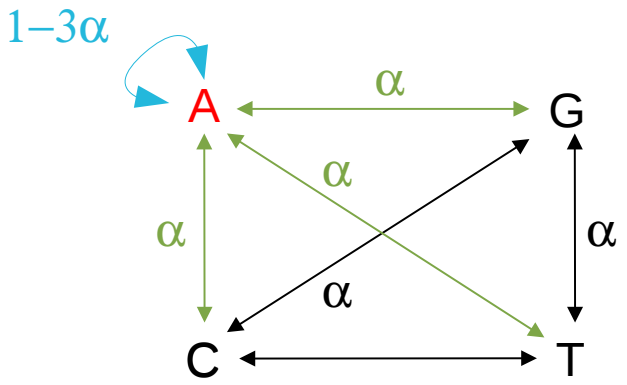
Si $P_A(0) = 0$ alors $P_A(t) = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} = p_{C/G/T \to A}(t) = p_{ij}(t)$

Calcul de la distance entre deux séquences nucléotidiques

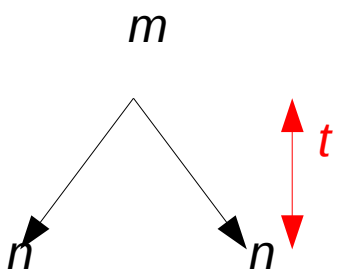
Le modèle à un paramètre de Jukes-Cantor (1969) : JC69

Toutes les substitutions sont équiprobables

Les proportions de chacune des 4 bases sont identiques



Entre deux séquences la probabilité qu'il y ait le même nucléotide à la même position :



$$p_{id} = p_{mA}^{\,2} + p_{mT}^{\,2} + p_{mC}^{\,2} + p_{mG}^{\,2} = P_{ii}^{\,2} + 3.p_{ij}^{\,2} \qquad (m = \text{A/C/G/T})$$

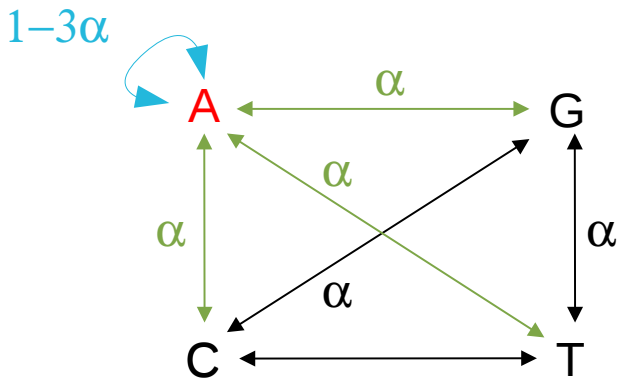
$$p_{id} = \left[\frac{1}{4} + \frac{3}{4} e^{-4\alpha t}\right]^2 + 3.\left[\frac{1}{4} - \frac{1}{4} e^{-4\alpha t}\right]^2 = \frac{1}{4} + \frac{3}{4} e^{-8\alpha t}$$

Calcul de la distance entre deux séquences nucléotidiques

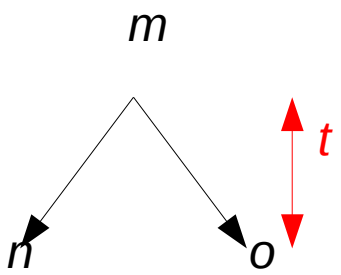
Le modèle à un paramètre de Jukes-Cantor (1969) : JC69

Toutes les substitutions sont équiprobables

Les proportions de chacune des 4 bases sont identiques



Entre deux séquences la probabilité qu'il y ait des nucléotides différents à la même position :



$$P_{nid} = 1 - p_{id} = 1 - \left[\frac{1}{4} + \frac{3}{4} e^{-8\alpha t}\right] = \frac{3}{4}(1 - e^{-8\alpha t})$$

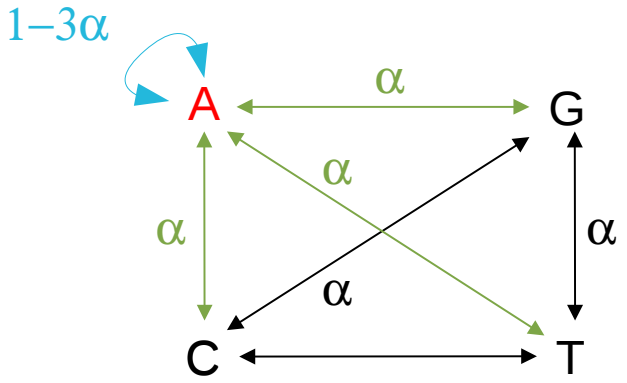
Ou encore, $8\alpha t = -\mathrm{Ln}\left(1 - \frac{4}{3} p_{nid}\right)$ où α et t sont inconnus.

Calcul de la distance entre deux séquences nucléotidiques

Le modèle à un paramètre de Jukes-Cantor (1969) : JC69

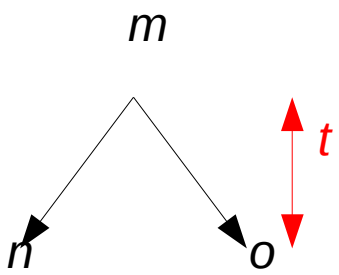
Toutes les substitutions sont équiprobables

Les proportions de chacune des 4 bases sont identiques



Entre deux séquences la probabilité qu'il y ait des nucléotides différents à la même position :

$$8\alpha t = -\ln\left(1 - \frac{4}{3} p_{nid}\right)$$



Pour une séquence, le nombre de substitution par site est : 3α

Pour deux séquences séparées depuis un temps t,

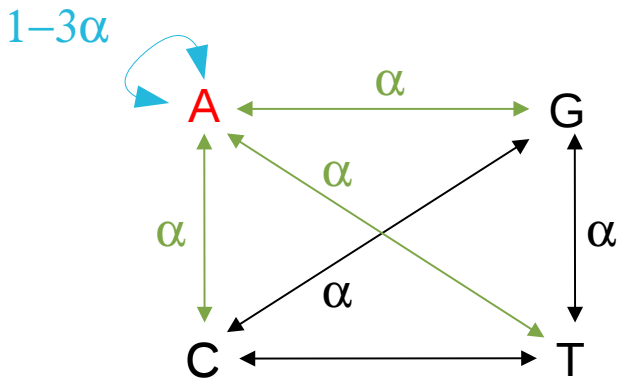
ce nombre est égal à : $2.3\alpha t = 6\alpha t = K$

Calcul de la distance entre deux séquences nucléotidiques

Le modèle à un paramètre de Jukes-Cantor (1969) : JC69

Toutes les substitutions sont équiprobables

Les proportions de chacune des 4 bases sont identiques



Entre deux séquences la probabilité qu'il y ait des nucléotides différents à la même position :

$$d' où, \quad \frac{4}{3} K = - \mathrm{Ln} \left(1 - \frac{4}{3} p_{nid}\right)$$

$$et \quad K = - \frac{3}{4} \mathrm{Ln} \left(1 - \frac{4}{3} p_{nid}\right)$$

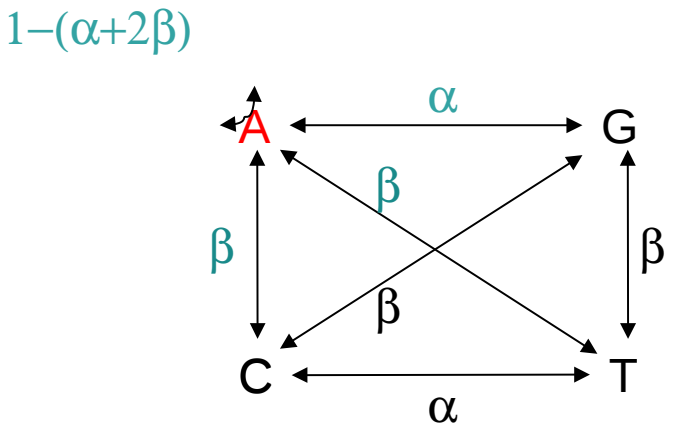
Nombre de substitutions estimées

p-distance

Calcul de la distance entre deux séquences nucléotidiques

Le modèle à deux paramètres de Kimura (1980) : K2P

On distingue les probabilités de transition (AG et CT) et de transversion (AC,AT,GC et GT).
Les proportions de chacune des 4 bases sont identiques



Soit p la proportion de transitions observées, q la proportion de transversions observées et k le nombre de substitutions estimées entre les deux séquences, alors :

$$k = -\frac{1}{2}\ln(1- 2p -q) - \frac{1}{4}\ln(1 -2q)$$

Calcul de la distance entre deux séquences nucléotidiques

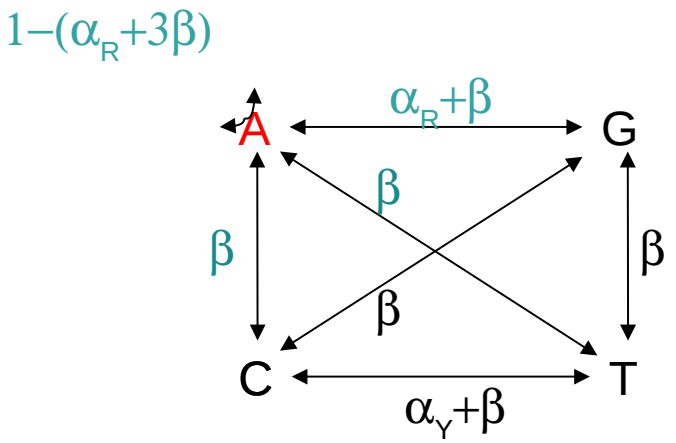
Le modèle à six paramètres de Tamura et Nei (1993) : TN93

On distingue : la probabilité de transition (AG): α_R

la probabilité de transition (CT) : α_Y

la probabilité de substitution : β

Les proportions de chacune des 4 bases ne sont pas forcément identiques (π_A, π_C, π_G et π_T).



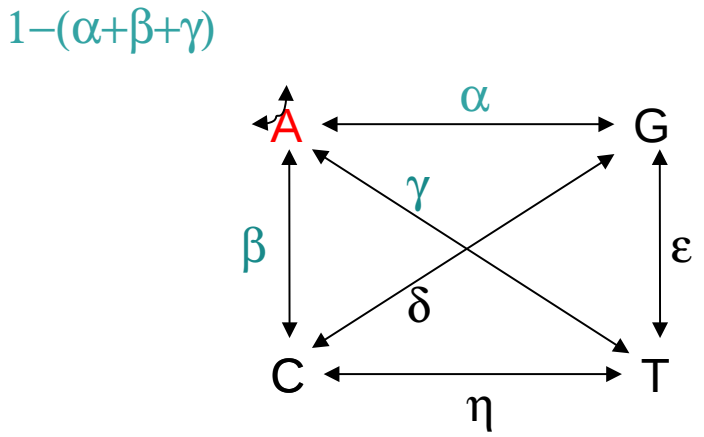
Si $\alpha_R=\alpha_Y$ alors nous somme dans le modèle de Felsenstein (1984) : F84

Si $\alpha_r/\alpha_Y = \pi_r/\pi_y$ alors nous sommes dans le modèle Hasegawa, Kishino and Yano (1985) : HKY

Calcul de la distance entre deux séquences nucléotidiques

Le modèle General Time Reversible : GTR

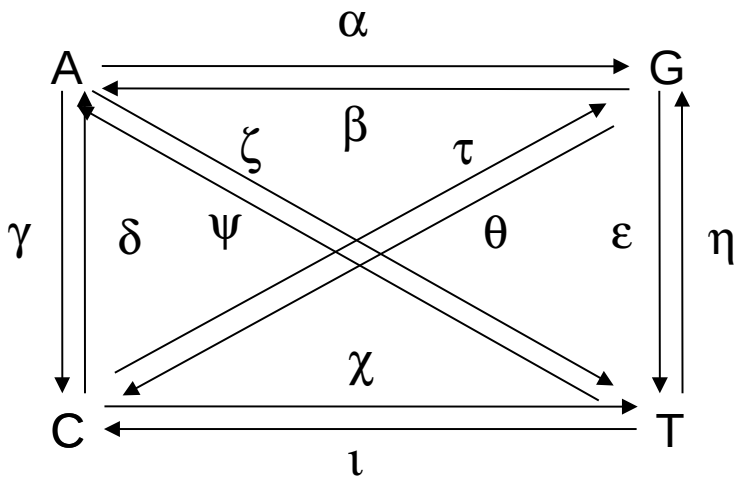
On distingue 6 probabilités de substitution.
Les proportions de chacune des 4 bases ne sont pas forcément identiques (π_A, π_C, π_G et π_T).



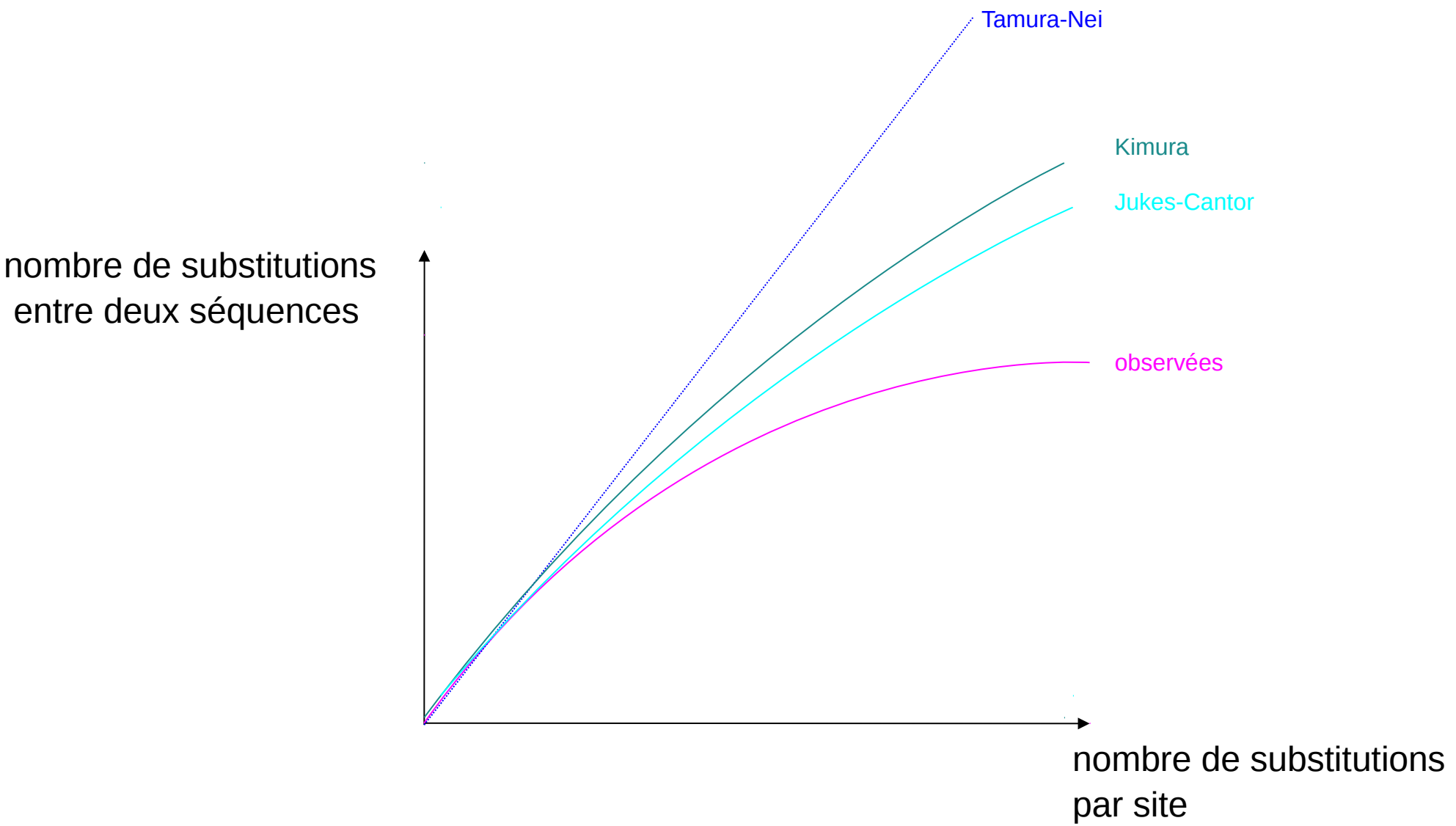
Calcul de la distance entre deux séquences nucléotidiques

Le modèle General à 12 paramètres :

On distingue 12 probabilités de substitution.
Les proportions de chacune des 4 bases ne sont pas forcément identiques (π_A, π_C, π_G et π_T).



Calcul de la distance entre deux séquences nucléotidiques



Unweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

Alignement multiple

A	GCTTGTCCGTTACGAT
B	ACTTGTCTGTTACGAT
C	ACTTGTCCGAAACGAT
D	ACTTGACCGTTTCCTT
E	AGATGACCGTTTCGAT
F	ACTACACCCTTATGAG

Matrice de distance

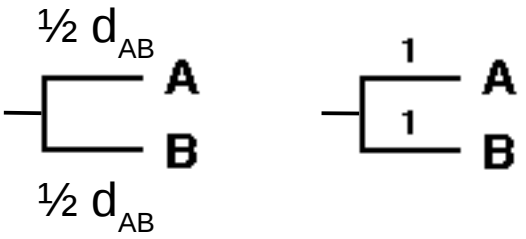
	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

Unweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

$$d_{AX} = d_{BX} = \frac{d_{AB}}{2}$$

Matrice de distance

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8



On regroupe les OTUs les plus proches et on calcule une nouvelle matrice.

Unweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

Matrice de distance M_i

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

$$d_{AB,X} = \frac{d_{A,X} + d_{B,X}}{2}$$

Matrice de distance M_{i+1}

	A B	C	D	E
C				
D		6		
E		6	4	
F		8	8	8

Unweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

Matrice de distance M_i

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

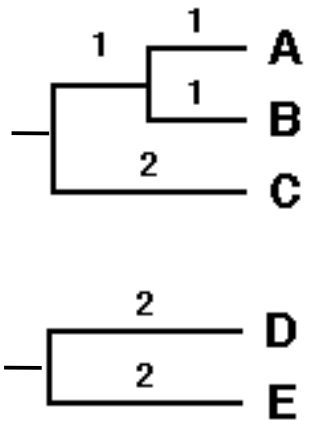
Matrice de distance M_{i+1}

	A,B	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

Unweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

Matrice de distance M_{i+1}

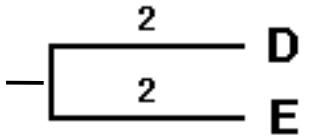
	A,B	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8



Unweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

Matrice de distance M_{i+1}

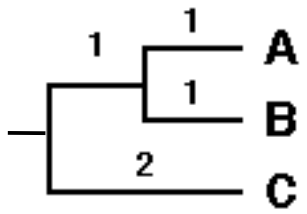
	A,B	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8



Unweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

Matrice de distance M_{i+2}

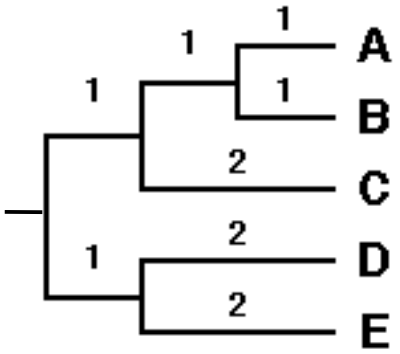
	A,B	C	D,E
C	4		
D,E	6	6	
F	8	8	8



Unweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

Matrice de distance M_{i+3}

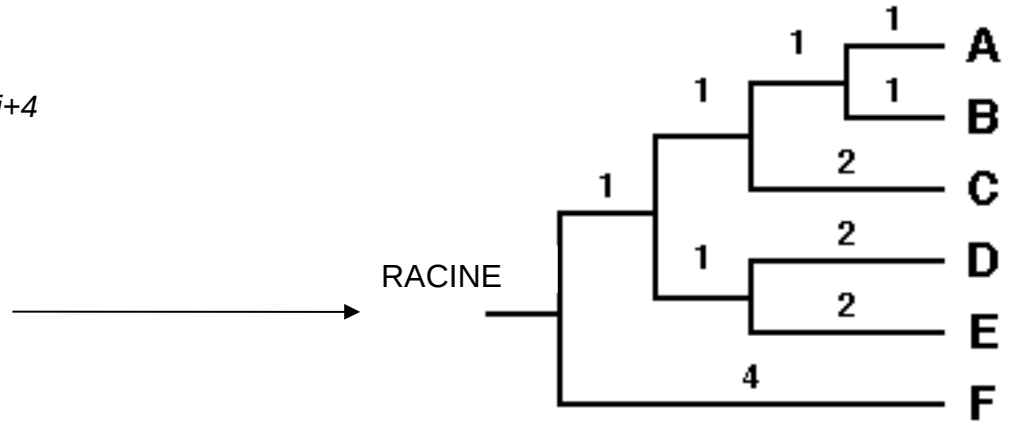
	(A,B),C	D,E
D,E	6	
F	8	8



Unweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

Matrice de distance M_{i+4}

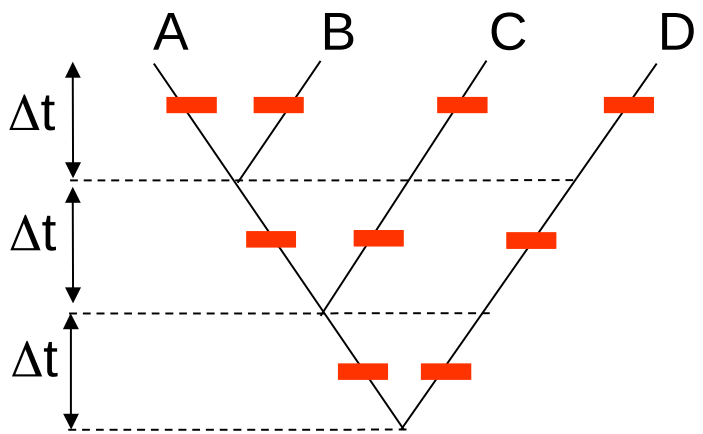
	((A,B),C),(D,E)
F	8

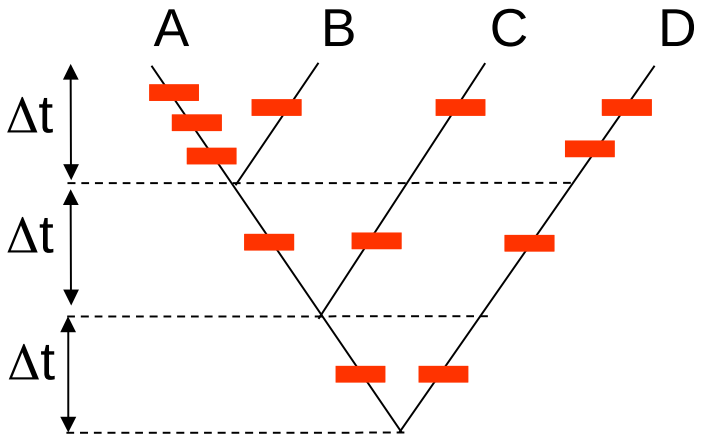


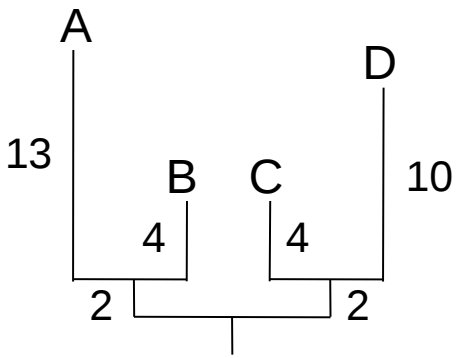
La méthode UPGMA :

- est rapide et simple

- mais sensible à l'horloge moléculaire (MCH, Zuckerkandl and Pauling, 1962)



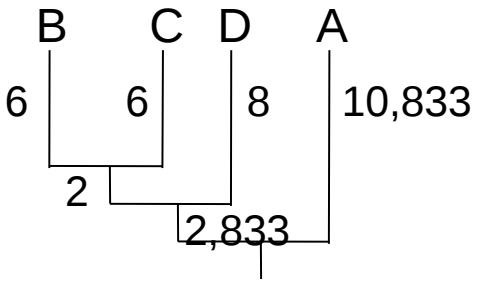




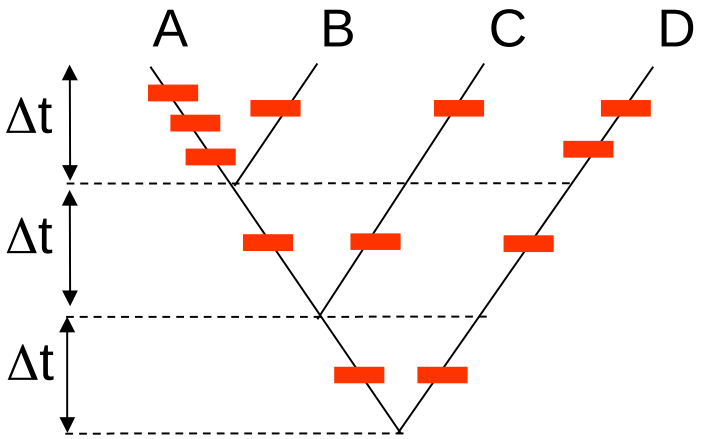
Arbre vrai

	A	B	C
B	17		
C	21	12	
D	27	18	14

Matrice de distance
(arborée)



Arbre UPGMA
(LBA)



Etant donnée une liste de caractères associés à un ensemble d'entités, comment construire un arbre retraçant les liens évolutifs entre toutes ces entités ?

Comment proposer un scénario évolutif à partir de l'observation des différences et ressemblances ?

1. Les méthodes de parcimonie
2. Les méthodes phénétiques (de distance)
3. Les méthodes probabilistes (maximum de vraisemblance et Bayesiennes)

Méthode de maximum de vraisemblance

$$P(D \text{ et } H) = P(D|H)\,P(H) = P(H|D)\,P(D)$$

Probabilité conditionnelle

Méthode de maximum de vraisemblance

$$P(D \text{ et } H) = P(D|H)\,P(H) = P(H|D)\,P(D)$$

Probabilité conditionnelle

fonction de vraisemblance de H

probabilité marginale (a priori) de H

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

Théorème de Bayes

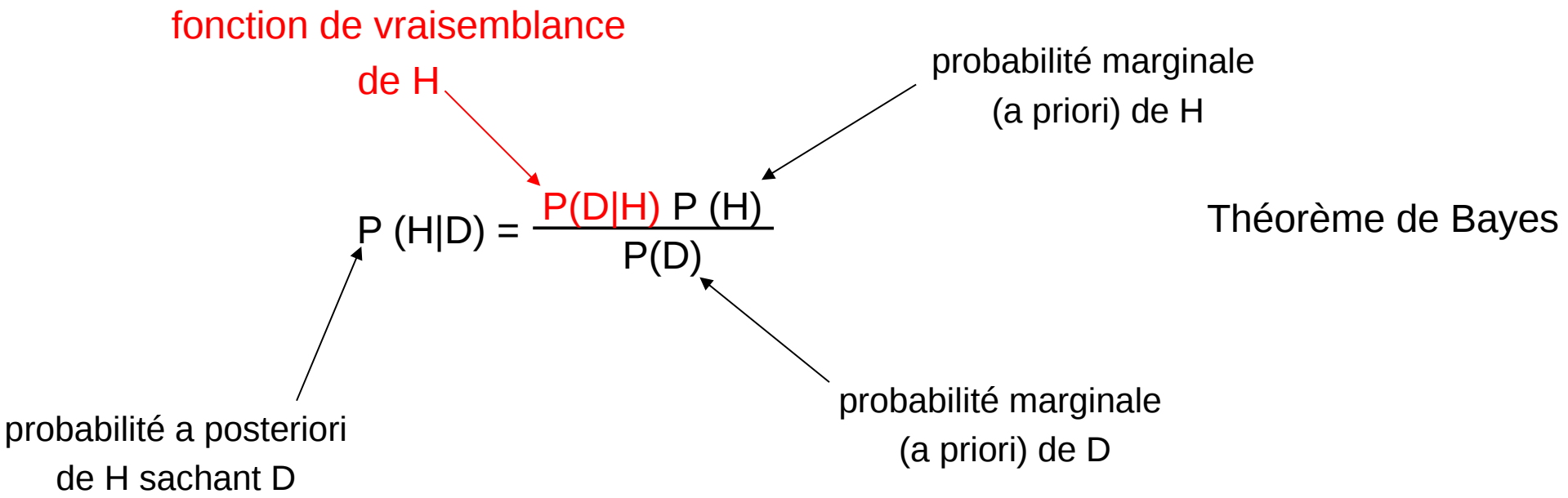
probabilité a posteriori de H sachant D

probabilité marginale (a priori) de D

Méthode de maximum de vraisemblance

$$P(D \text{ et } H) = P(D|H)\,P(H) = P(H|D)\,P(D)$$

Probabilité conditionnelle



Méthode de maximum de vraisemblance

La vraisemblance est donc la probabilité d'observer un jeu de données (D) sachant une hypothèse H :

$$L = P(D|H)$$

On considère que l'hypothèse pour laquelle cette probabilité est maximale est celle qui explique le mieux les données.

Méthode de maximum de vraisemblance

Exemple : Soit p la probabilité d'obtenir pile, quelle est la vraisemblance d'observer le résultat suivant si on réalise 6 lancés ?

D : Pile Face Face Pile Pile Pile

$L = P(D|H) = ?$

Méthode de maximum de vraisemblance

Exemple : Soit p la probabilité d'obtenir pile, quelle est la vraisemblance d'observer le résultat suivant si on réalise 6 lancés ?

D : Pile Face Face Pile Pile Pile

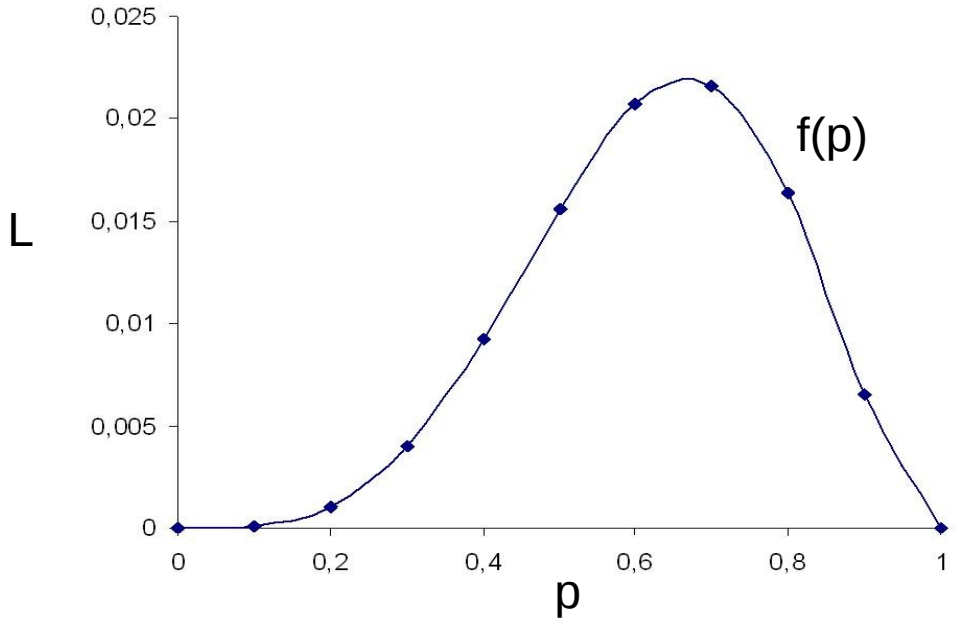
$L = P(D|H) = P(D|p) = p\,(1-p)\,(1-p)\,p\,p\,p = p^4(1-p)^2$

Méthode de maximum de vraisemblance

Exemple : Soit p la probabilité d'obtenir pile, quelle est la vraisemblance d'observer le résultat suivant si on réalise 6 lancés ?

D : Pile Face Face Pile Pile Pile

$L = P(D|H) = P(D|p) = p\,(1-p)\,(1-p)\,p\,p\,p = p^4(1-p)^2$

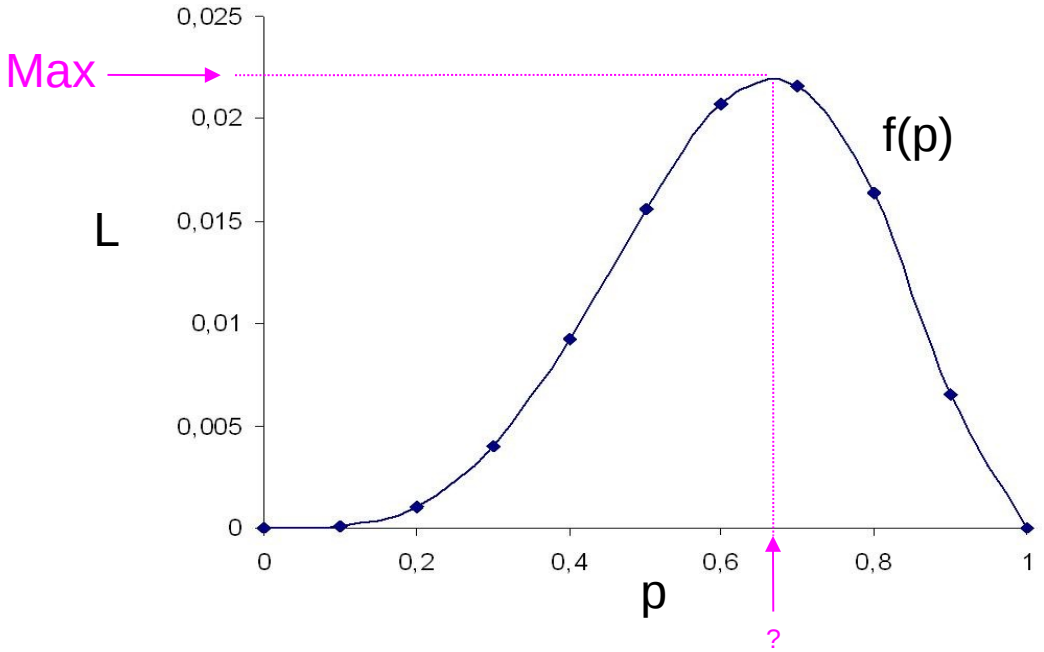


Méthode de maximum de vraisemblance

Exemple : Soit p la probabilité d'obtenir pile, quelle est la vraisemblance d'observer le résultat suivant si on réalise 6 lancés ?

D : Pile Face Face Pile Pile Pile

$L = P(D|H) = P(D|p) = p\,(1-p)\,(1-p)\,p\,p\,p = p^4(1-p)^2$

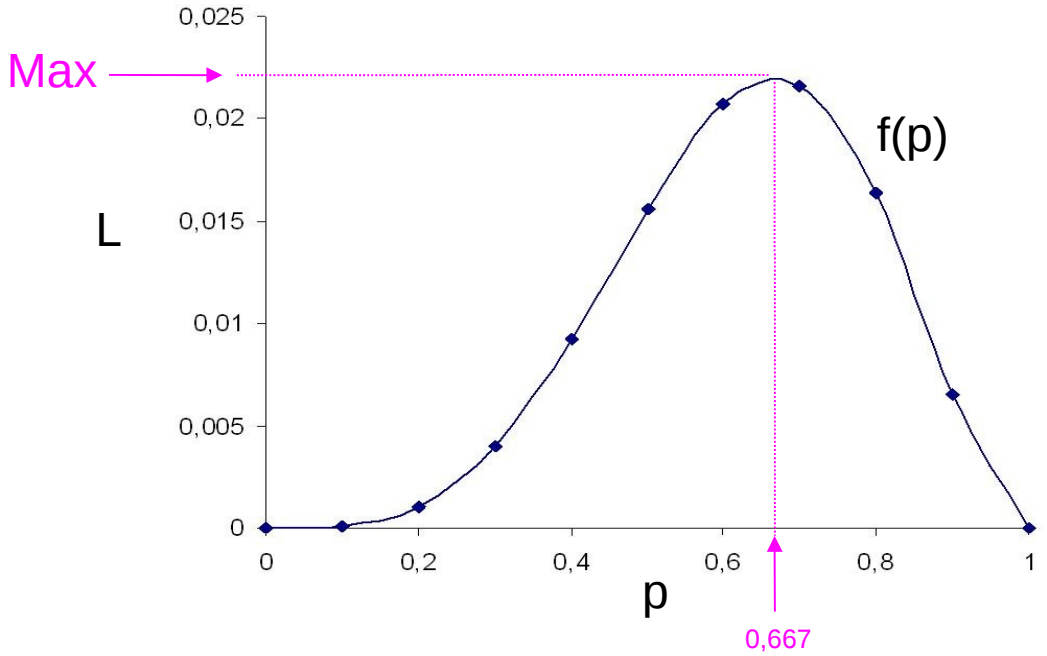


Méthode de maximum de vraisemblance

Exemple : Soit p la probabilité d'obtenir pile, quelle est la vraisemblance d'observer le résultat suivant si on réalise 6 lancés ?

D : Pile Face Face Pile Pile Pile

$L = P(D|H) = P(D|p) = p\,(1-p)\,(1-p)\,p\,p\,p = p^4(1-p)^2$



$\ln L = \ln [\, p^4 (1-p)^2 \,] = 4 \ln p + 2 \ln (1-p)$

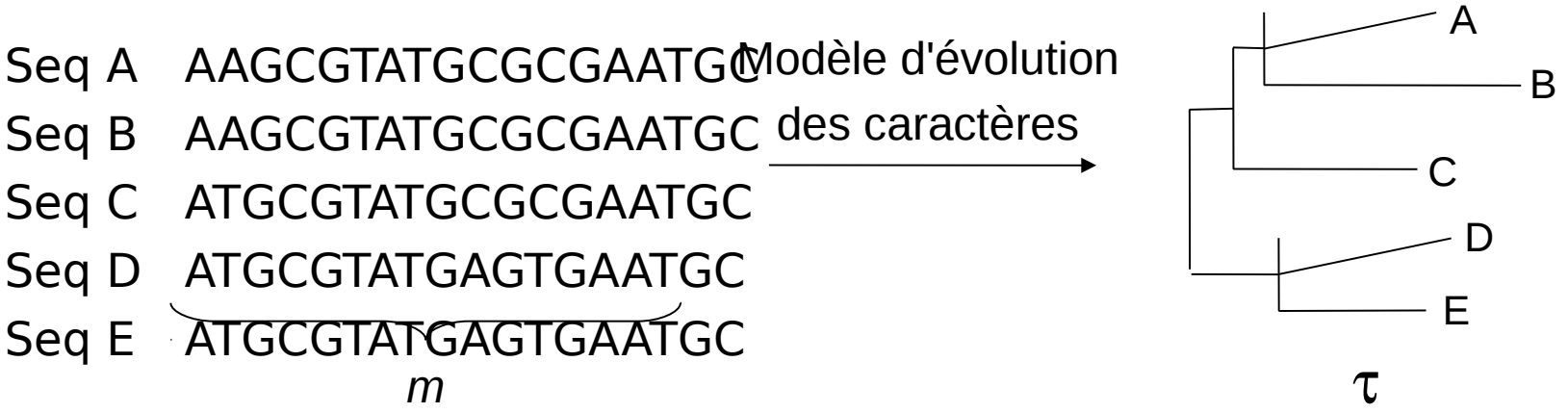
$$\frac{d(\ln L)}{d\,p} = \frac{4}{p} - \frac{2}{(1-p)} = 0$$

$$p = \frac{4}{6} = \frac{2}{3} = 0{,}666666667$$

La vraisemblance est maximale quand p, le paramètre du modèle est égal à 2/3.

Méthode de maximum de vraisemblance

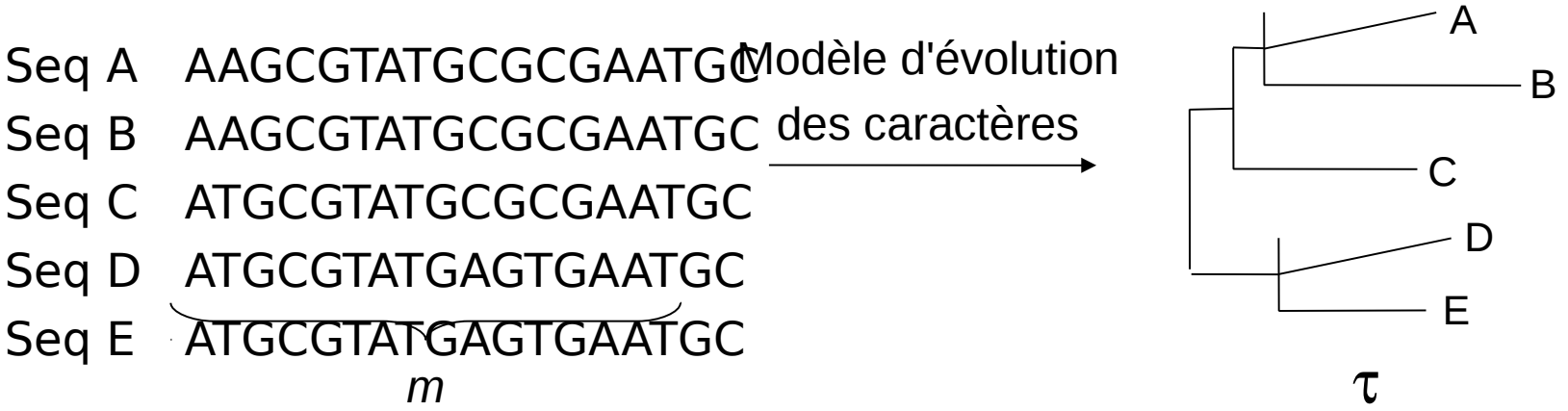
En phylogénie moléculaire, les données seront composées d'un ensemble de caractères (des séquences) et l'hypothèse sera constituée par une topologie et un modèle d'évolution.



$$L = P(D|\tau,M)$$

Méthode de maximum de vraisemblance

En phylogénie moléculaire, les données seront composées d'un ensemble de caractères (des séquences) et l'hypothèse sera constituée par une topologie et un modèle d'évolution.



Les *m* caractères étant considérés comme indépendants alors

$$L = P(D|\tau,M) = \prod_{i=1}^{m} P(D^{(i)}|\tau,M) = \prod_{i=1}^{m} L^{(i)}$$

$$\ln L = \ln\left(\prod_{i=1}^{m} L^{(i)}\right) = \sum_{i=1}^{m} \ln L^{(i)}$$

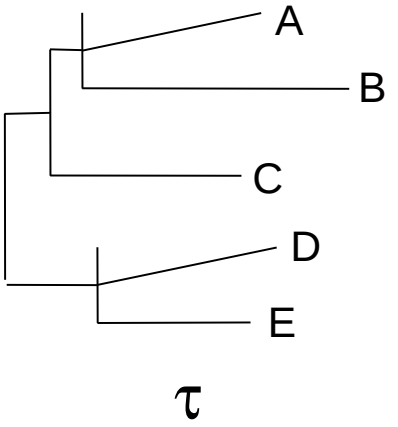
Méthode de maximum de vraisemblance

En phylogénie moléculaire, les données seront composées d'un ensemble de caractères (des séquences) et l'hypothèse sera constituée par une topologie et un modèle d'évolution.

Seq A AAGCGTATGCGCGAATGC
Seq B AAGCGTATGCGCGAATGC
Seq C ATGCGTATGCGCGAATGC
Seq D ATGCGTATGAGTGAATGC
Seq E ATGCGTATGAGTGAATGC

m

Modèle d'évolution des caractères



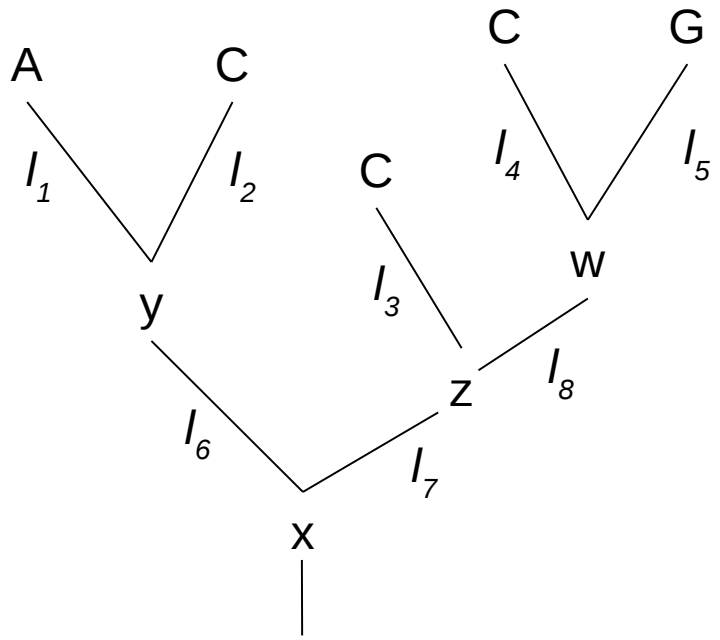
τ

Les *m* caractères étant considérés comme indépendants alors

$$L = P(D|\tau,M) = \prod_{i=1}^{m} P(D^{(i)}|\tau,M) = \prod_{i=1}^{m} L^{(i)}$$

$$\ln L = \ln\left(\prod_{i=1}^{m} L^{(i)}\right) = \sum_{i=1}^{m} \ln L^{(i)}$$

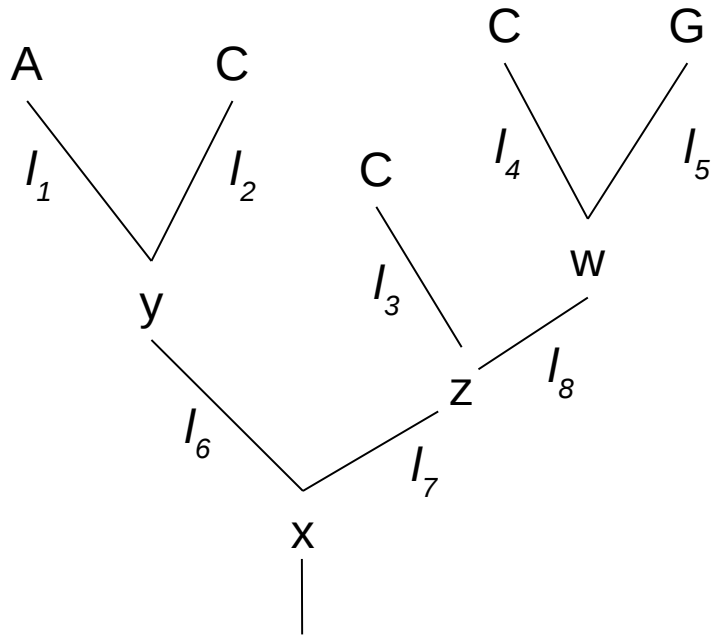
Méthode de maximum de vraisemblance



$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z \sum_w P(A,C,C,C,G,x,y,z,w|\tau,M)$$

$$P(A,C,C,C,G,x,y,w,z|\tau,M) = P(x)$$
$$P(y|x,l_6)\, P(A|y,l_1)\, P(C|y,l_2)$$
$$P(z|x,l_7)\, P(C|z,l_3)\, P(w|z,l_8)\, P(C|w,l_4)\, P(G|w,l_5)$$

Méthode de maximum de vraisemblance



$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z \sum_w P(A,C,C,C,G,x,y,z,w|\tau,M)$$

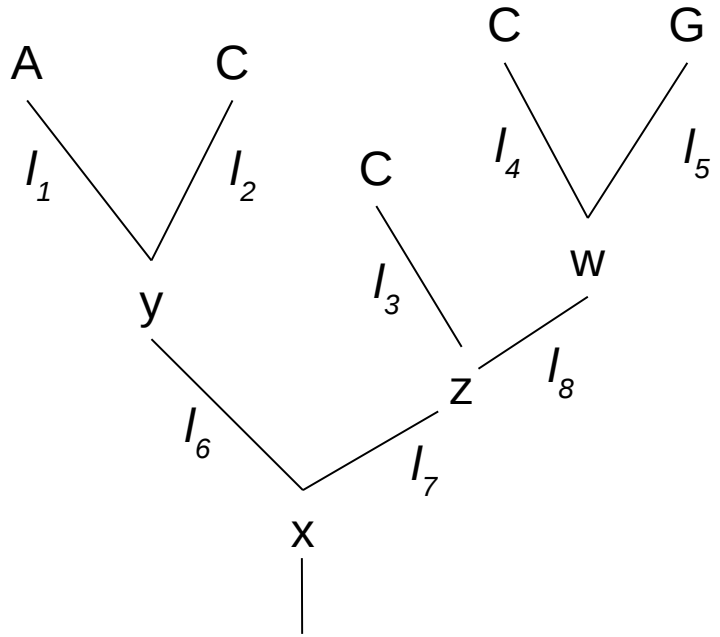
$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z \sum_w$$

$P(x)$

$P(y|x,l_6)\ P(A|y,l_1)\ P(C|y,l_2)$

$P(z|x,l_7)\ P(C|z,l_3)\ P(w|z,l_8)\ P(C|w,l_4)\ P(G|w,l_5)$

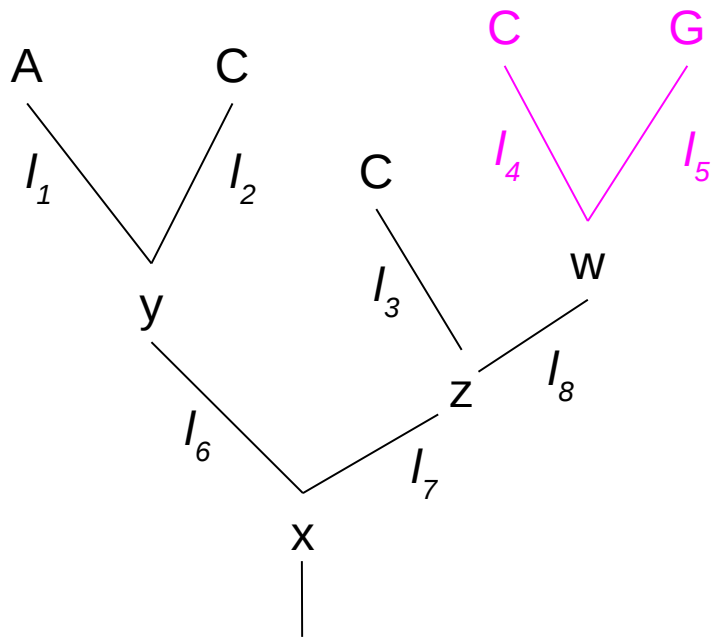
Méthode de maximum de vraisemblance



$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z \sum_w P(A,C,C,C,G,x,y,z,w|\tau,M)$$

$$P(D^{(i)}|\tau,M) = \sum_x P(x) \left(\sum_y P(y|x,l_6)\, P(A|y,l_1)\, P(C|y,l_2) \left(\sum_z P(z|x,l_7)\, P(C|z,l_3) \left(\sum_w P(w|z,l_8)\, P(C|w,l_4)\, P(G|w,l_5)\right)\right)\right)$$

Méthode de maximum de vraisemblance



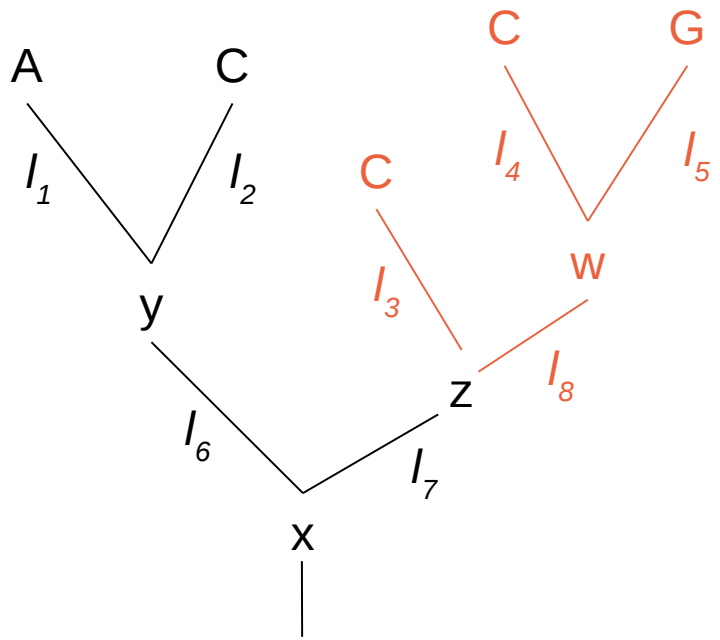
$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z \sum_w P(A,C,C,C,G,x,y,z,w|\tau,M)$$

vraisemblance conditionnelle

$L_8^{(i)}(w)$

$$P(D^{(i)}|\tau,M) = \sum_x P(x) \left(\sum_y P(y|x,l_6)\, P(A|y,l_1)\, P(C|y,l_2)\right) \left(\sum_z P(z|x,l_7)\, P(C|z,l_3) \left(\sum_w P(w|z,l_8)\, P(C|w,l_4)\, P(G|w,l_5)\right)\right)$$

Méthode de maximum de vraisemblance

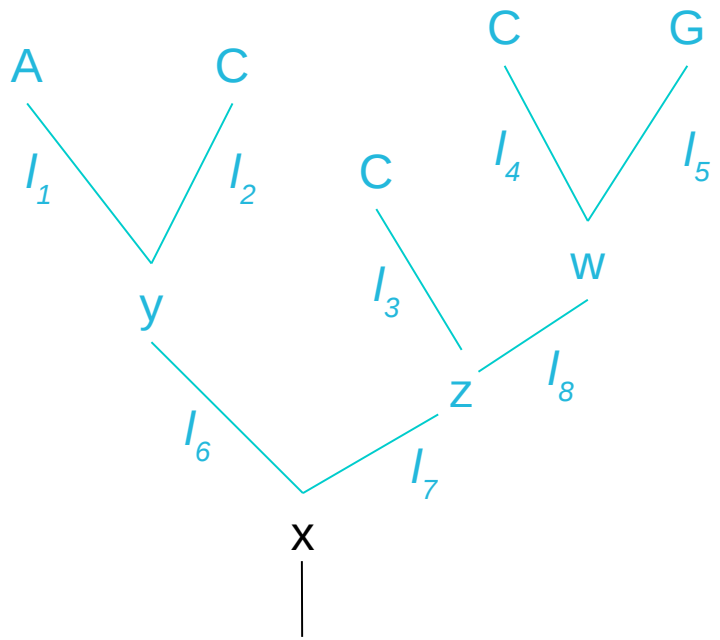


$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z \sum_w P(A,C,C,C,G,x,y,z,w|\tau,M)$$

$$L_7^{(i)}(z) = P(C|z,l_3) \sum_w P(w|z,l_8)\, L_8^{(i)}(w)$$

$$P(D^{(i)}|\tau,M) = \sum_x P(x) \left(\sum_y P(y|x,l_6)\, (A|y,l_1)\, P(C|y,l_2)\right) \left(\sum_z P(z|x,l_7)\, P(C|z,l_3) \left(\sum_w P(w|z,l_8)\, P(C|w,l_4)\, P(G|w,l_5)\right)\right)$$

Méthode de maximum de vraisemblance



$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z \sum_w P(A,C,C,C,G,x,y,z,w|\tau,M)$$

$$L_{root}^{(i)}(x) = \sum_y P(y|x,l_6)\, L_6^{(i)}(y) \sum_z P(z|x,l_7)\, L_7^{(i)}(z)$$

$$P(D^{(i)}|\tau,M) = \sum_x P(x)$$

$$\left(\sum_y P(y|x,l_6)\, (A|y,l_1)\, P(C|y,l_2)\right.$$

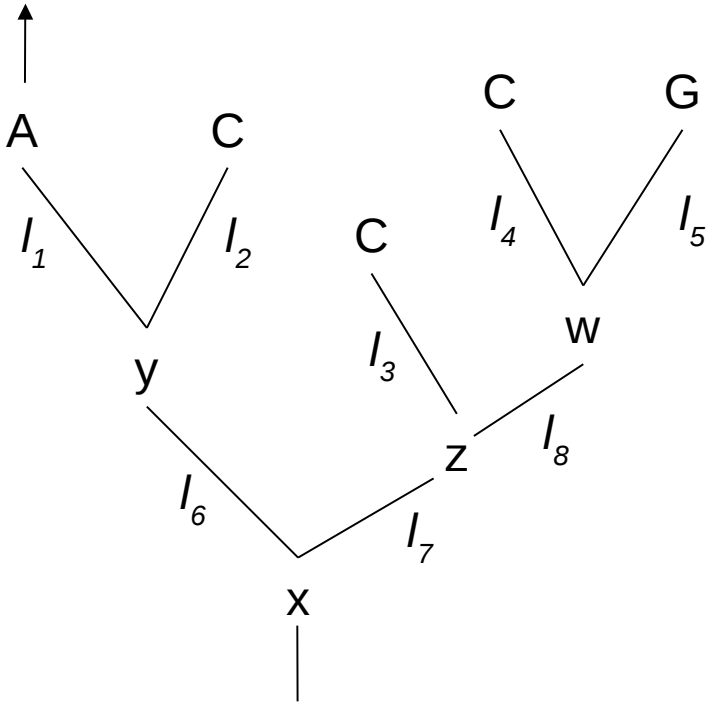
$$\left.\left(\sum_z P(z|x,l_7)\, P(C|z,l_3)\, \left(\sum_w P(w|z,l_8)\, P(C|w,l_4)\, P(G|w,l_5)\right)\right)\right)$$

Méthode de maximum de vraisemblance

$$P(D^{(i)}|\tau,M) = L^{(i)} = \sum_x \pi_x \, L_{root}^{(i)}(x)$$

probabilité à priori d'après le modèle d'évolution choisi

$(L^{(i)}(A), L^{(i)}(C), L^{(i)}(G), L^{(i)}(T)) = (1,0,0,0)$



$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z P(x)\, P(y|x,l_6)\, L_6^{(i)}(y)\, P(z|x,l_7)\, L_7^{(i)}(z)$$

or $P(x)\, P(y|x,l_6) = P(y)\, P(x|y,l_6)$

$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z P(y)\, P(x|y,l_6)\, L_6^{(i)}(y)\, P(z|x,l_7)\, L_7^{(i)}(z)$$

Méthode de maximum de vraisemblance

En résumé :

1. On explore l'univers des topologies

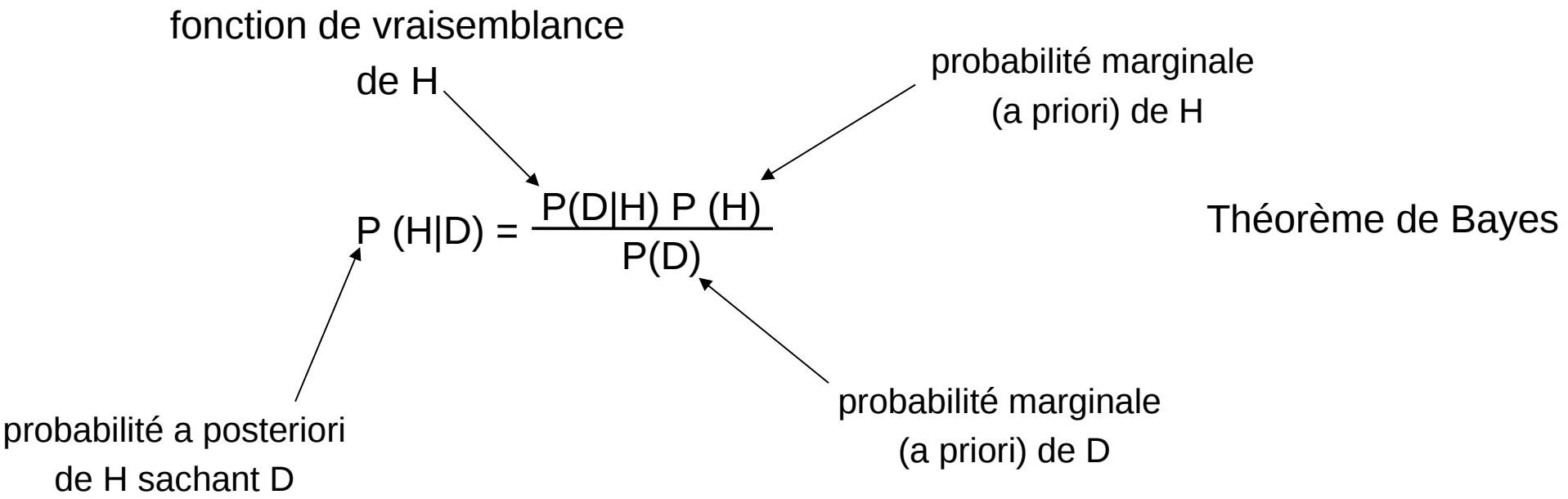
2. Pour chaque topologie on cherche les longueurs de branches
   et les paramètre de modèle pour lesquels la vraisemblance est maximale

3. On retient l'arbre pour lequel la topologie, les longueurs de branches et les paramètres de modèles présentent la vraisemblance maximale. Après un temps variable on doit théoriquement converger vers cette valeur maximale

Méthodes Bayesiennes

$$P(D \text{ et } H) = P(D|H)\,P(H) = P(H|D)\,P(D)$$

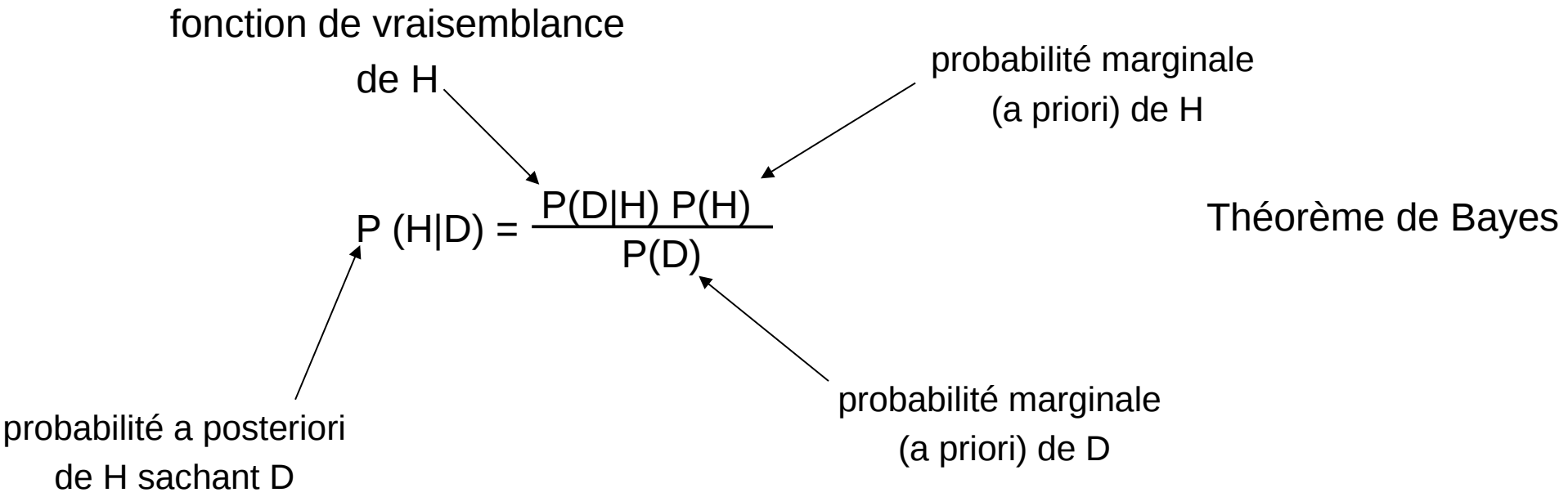
Probabilité conditionnelle



Méthodes Bayesiennes

$$P(D \text{ et } H) = P(D|H)\,P(H) = P(H|D)\,P(D)$$

Probabilité conditionnelle



Théorème de Bayes

D'après la loi des probabilités totales (alternatives) $P(D) = \sum_{H} P(D \text{ et } H)$

Méthodes Bayesiennes

$$P(D \text{ et } H) = P(D|H)\,P(H) = P(H|D)\,P(D)$$

Probabilité conditionnelle

fonction de vraisemblance de H

probabilité marginale (a priori) de H

$$P(H|D) = \frac{P(D|H)\,P(H)}{\sum_{H} P(H)\,P(D|H)}$$

Théorème de Bayes

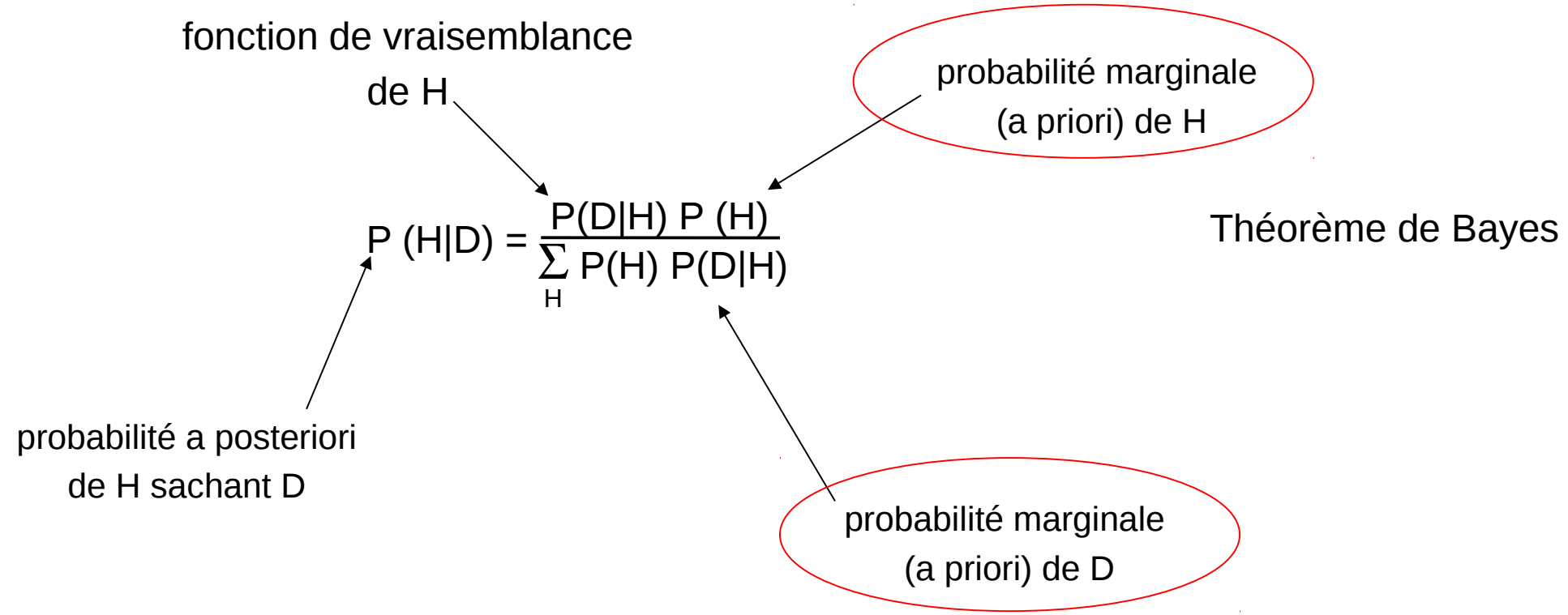
probabilité a posteriori de H sachant D

probabilité marginale (a priori) de D

Méthodes Bayesiennes

$$P(D \text{ et } H) = P(D|H)\,P(H) = P(H|D)\,P(D)$$

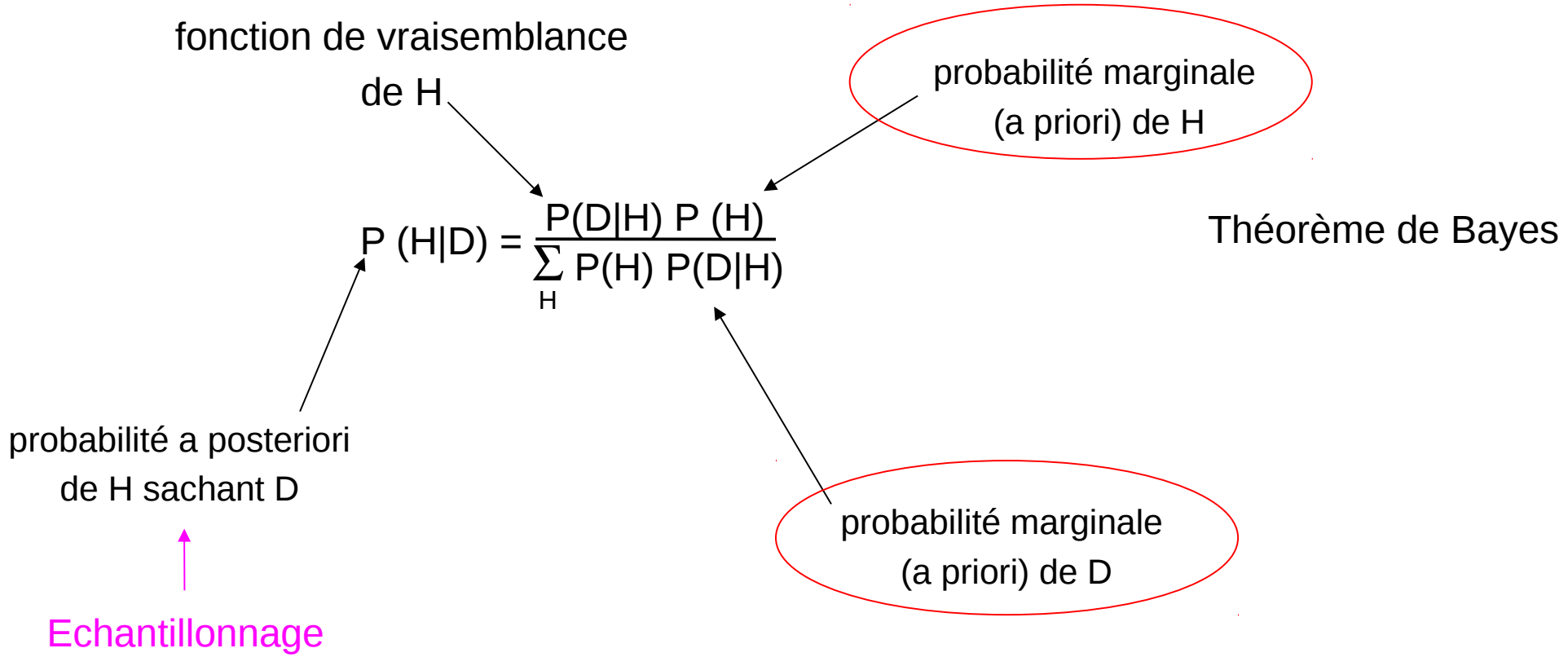
Probabilité conditionnelle



Méthodes Bayesiennes

$$P(D \text{ et } H) = P(D|H)\,P(H) = P(H|D)\,P(D)$$

Probabilité conditionnelle



Méthodes Bayesiennes

Utilisation du rapport des vraisemblances

$$P(H_i|D) = \frac{P(D|H_i)\,P(H_i)}{\sum_H P(H)\,P(D|H)} \qquad P(H_j|D) = \frac{P(D|H_j)\,P(H_j)}{\sum_H P(H)\,P(D|H)}$$

$$R = \frac{P(D|H_i)\,P(H_i)}{P(D|H_j)\,P(H_j)} = \frac{P(D|H_i)}{P(D|H_j)}\,\frac{P(H_i)}{P(H_j)}$$

Si les hypothèses H_i et H_j sont équiprobables alors :

$$R = \frac{P(D|H_i)}{P(D|H_j)} = \frac{P(D|\tau_i)}{P(D|\tau_j)}$$

Méthodes Bayesiennes

Utilisation du rapport des vraisemblances

$$P(H_i|D) = \frac{P(D|H_i)\,P(H_i)}{\sum_H P(H)\,P(D|H)} \qquad P(H_j|D) = \frac{P(D|H_j)\,P(H_j)}{\sum_H P(H)\,P(D|H)}$$

$$R = \frac{P(D|H_i)\,P(H_i)}{P(D|H_j)\,P(H_j)} = \frac{P(D|H_i)}{P(D|H_j)}\,\frac{P(H_i)}{P(H_j)}$$

Si les hypothèses H_i et H_j sont équiprobables alors :

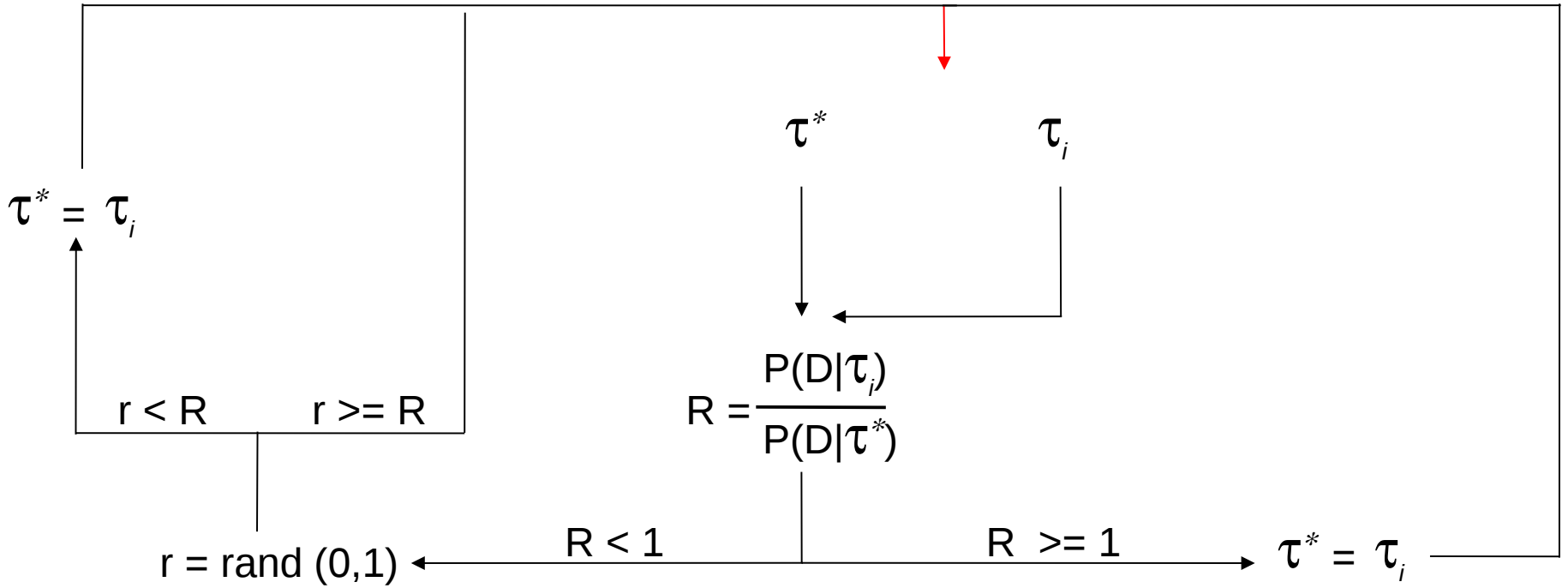
$$R = \frac{P(D|H_i)}{P(D|H_j)} = \frac{P(D|\tau_i)}{P(D|\tau_j)}$$

$L(\tau_i)$

$L(\tau_j)$

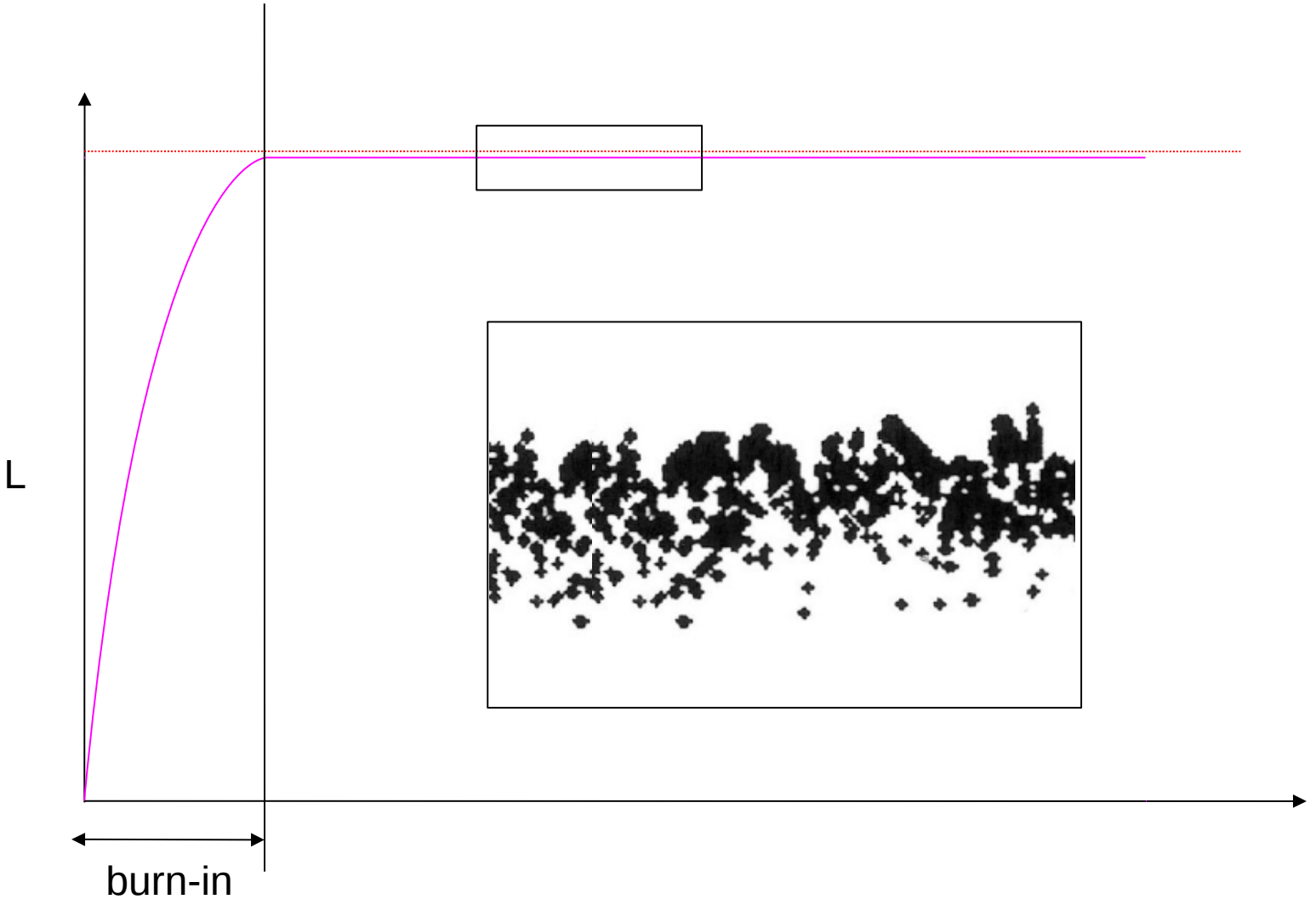
Méthodes Bayesiennes

Algorithme de Metropolis-Hastings (Markov Chain Monte Carlo, MCMC)



Méthodes Bayesiennes

Algorithme de Metropolis-Hastings (Markov Chain Monte Carlo, MCMC)



Méthodes Bayesiennes

Résumé :

1. Construction de CM reliant les arbres entre eux

2. obtention d'un échantillonnage reflétant la distribution des probabilités postérieures

3. Sommation des données :

    - Recherche de l'arbre de plus grande probabilité postérieure

    - Construction d'un consensus à partir des arbres de plus grande probabilités postérieures

    - Probabilité d'un clade : Σ des probabilités postérieures des arbres qui possèdent ce clade