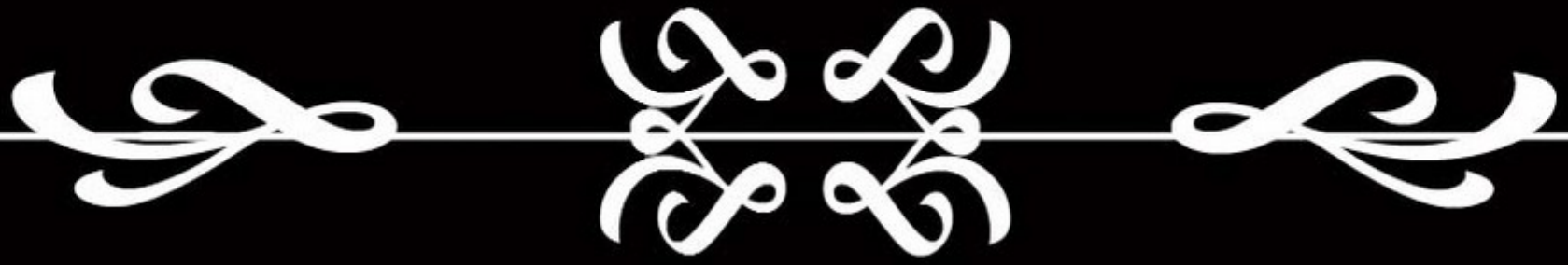
Une brève histoire

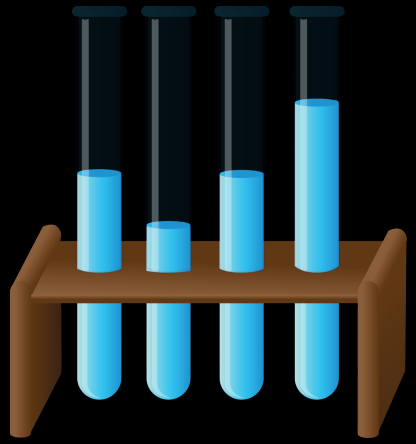
– de la –

BIOINFORMATIQUE



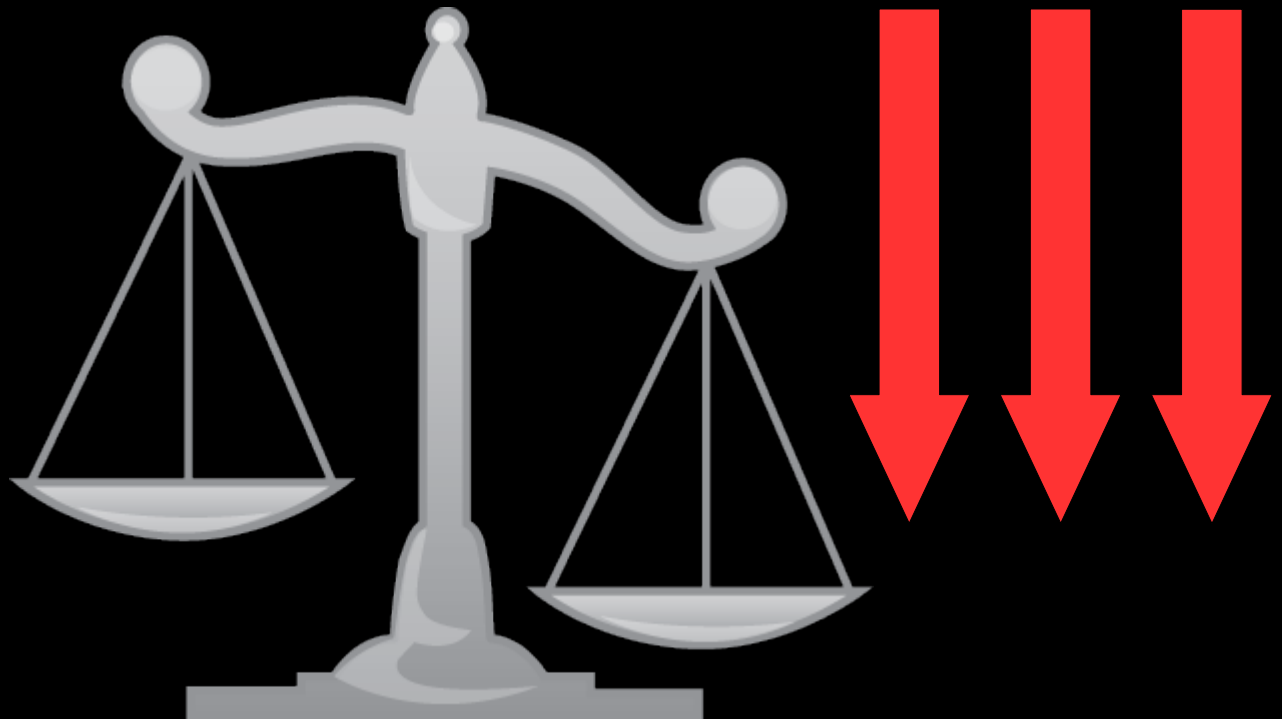
PARTIE I
(1950-1970)

FACTEURS LIMITANTS EN RECHERCHE...
...en 2015

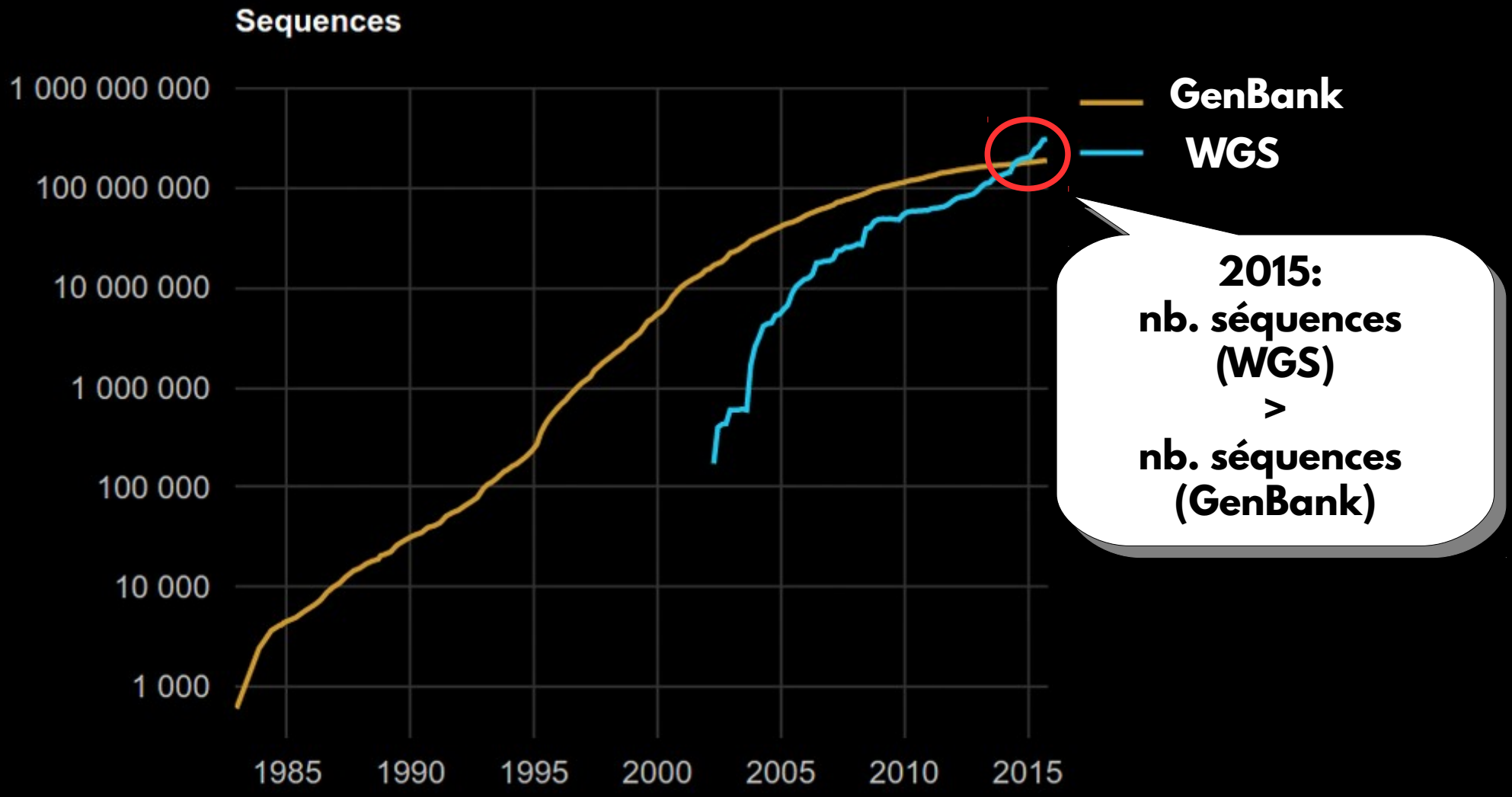


"Générer les données"

"Analyser les données"



CROISSANCE EXPONENTIELLE DE LA QUANTITÉ DE SÉQUENCES D'ADN PUBLIÉES !

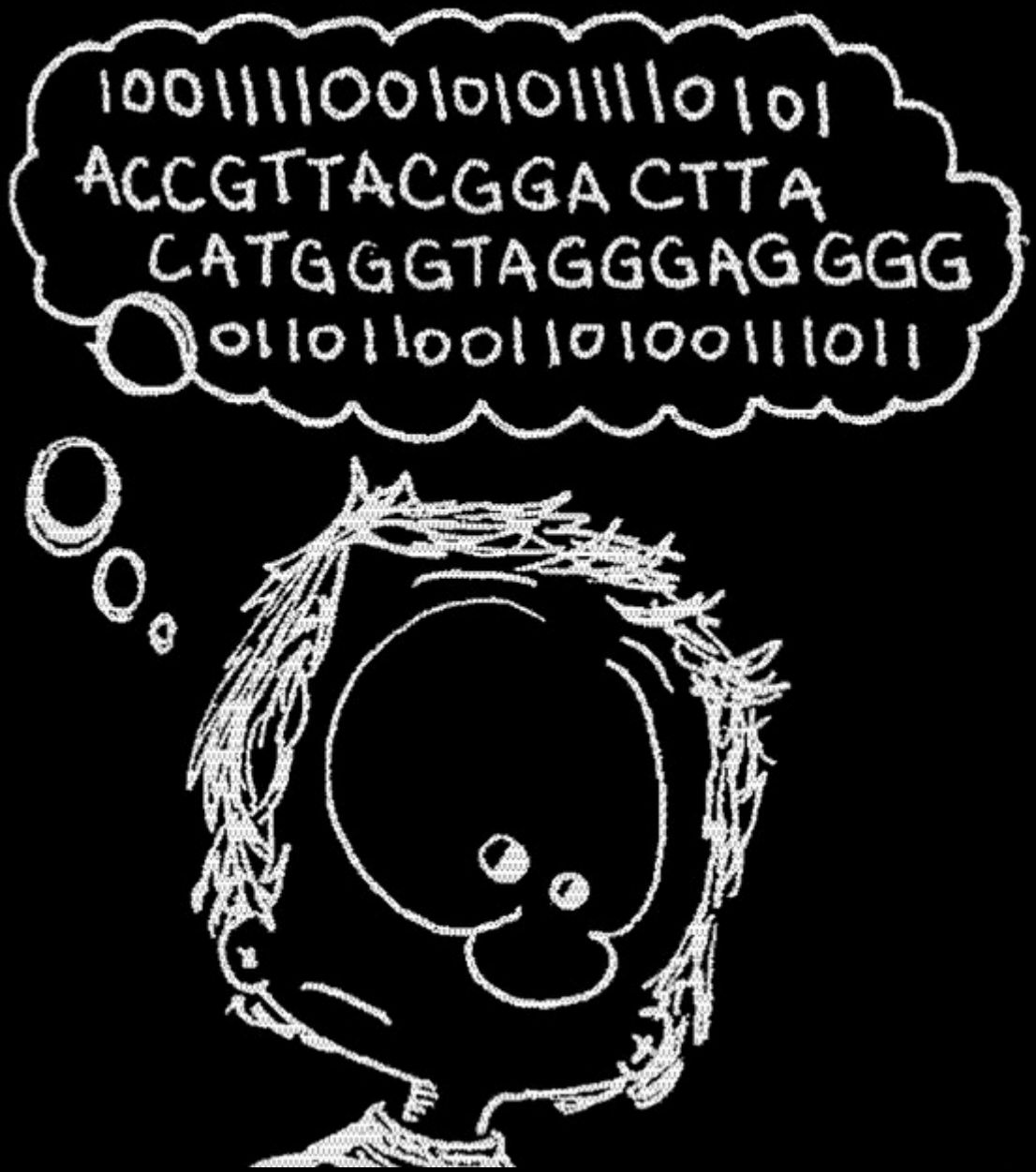


NCBI 2015

PROBLÈMES:
1) Notre capacité de calcul

PROBLÈMES:
2) Surcharge de données brutes

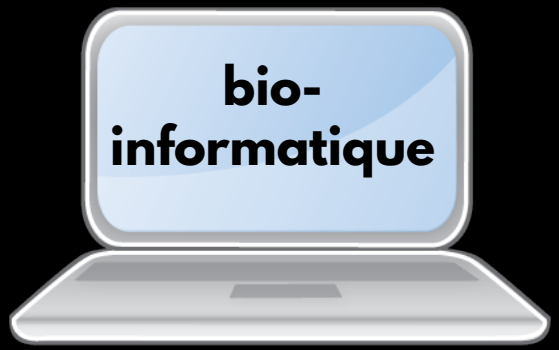
…le cerveau humain ne suffit pas à la tâche !



(lots of) DNA sequences

Human-readable DATA

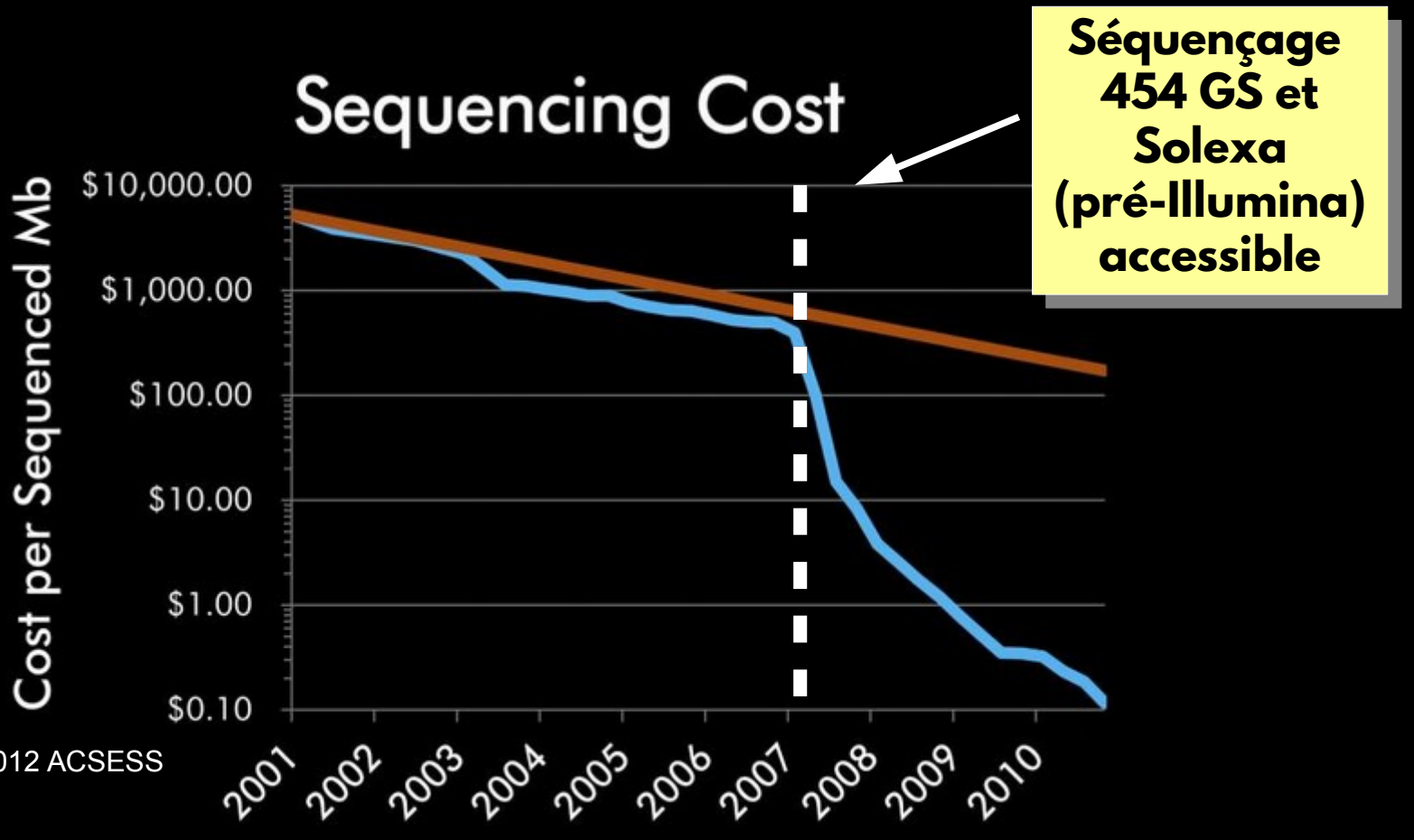
$v\,(Rx) \longrightarrow \infty$



(lots of) DNA sequences → Human-readable DATA

v (Rx) → Heures, jours, semaines…

TOUT PORTE À CROIRE QUE...



Poland et al. 2012 ACSESS

...la bioinformatique serait une discipline récente, venue prêter main forte à l'analyse de milliers de séquences d'ADN.

Est-ce vraiment le cas ? Pour le savoir...


RETOURNONS EN 1950...

1970
1965
1960
1955
1950

1950

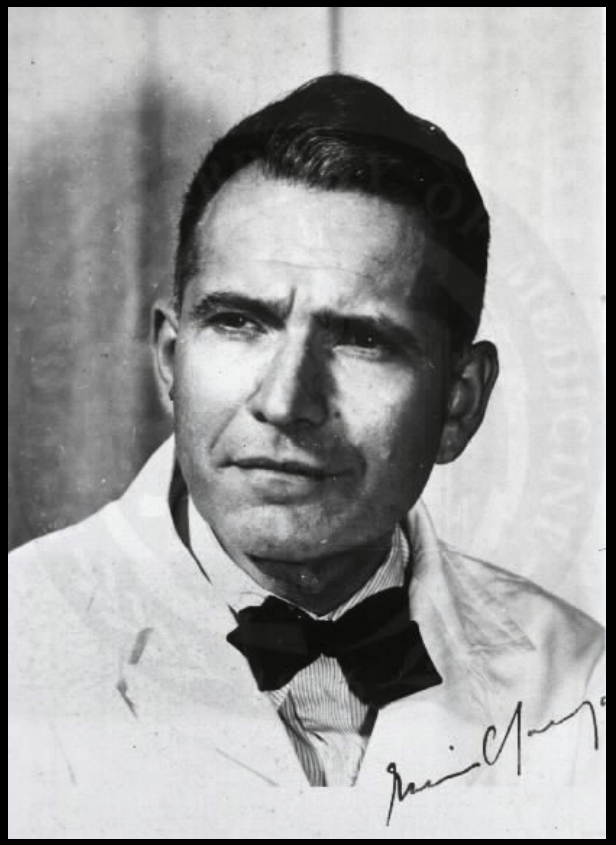
ON SAIT QUE... PEU DE CHOSES SUR LA STRUCTURE DE L'ADN.

- DÉSOXYRIBOSE
- PHOSPHATE
- QUATRE BASES AZOTÉES (A-T-C-G)

% A = % T

% C = % G

...et c'est presque tout.



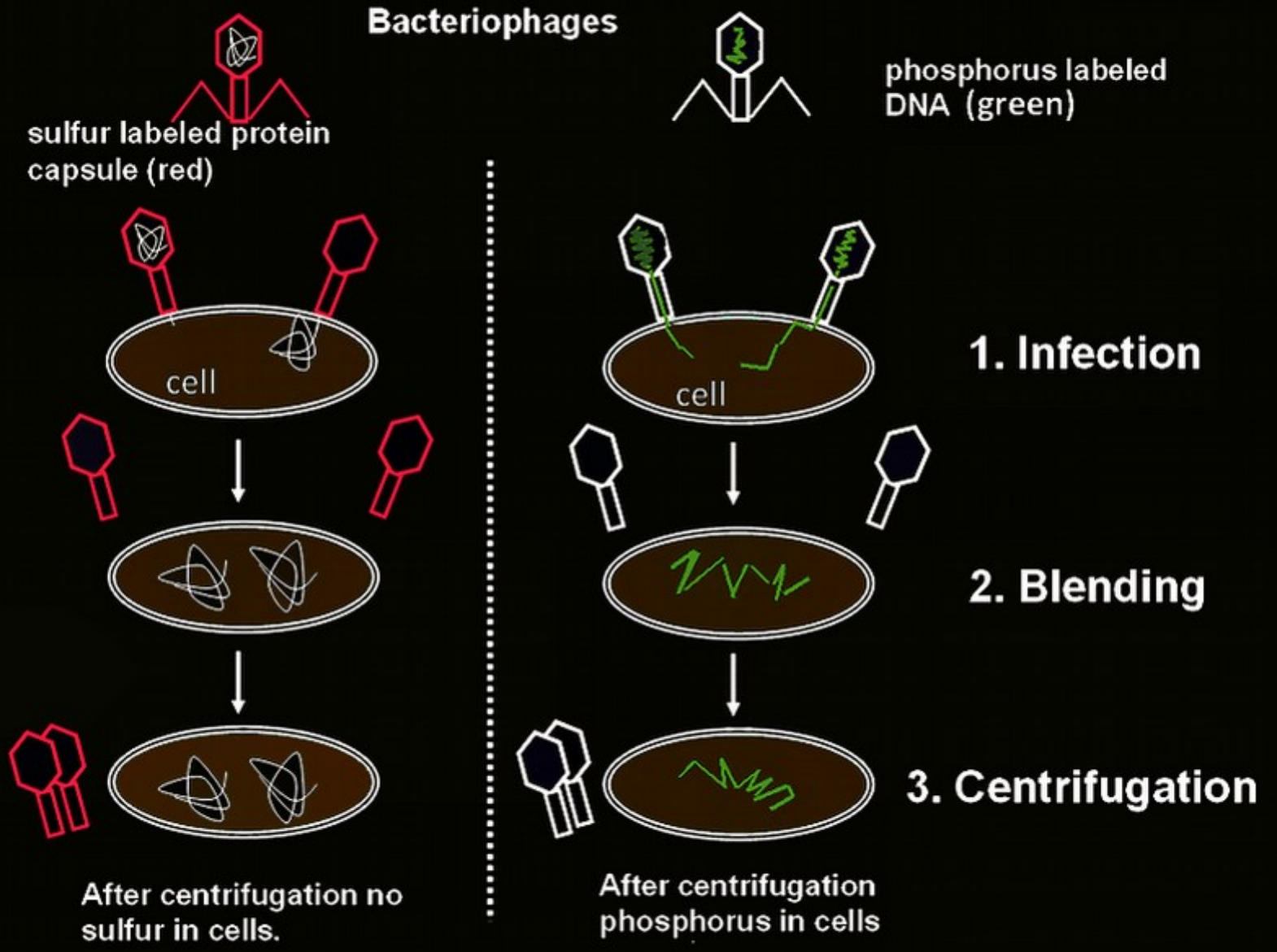
CHARGAFF (1950)

1970 - - - - 1965 - - - - - 1960 - - - - - 1955 - - - - 1950

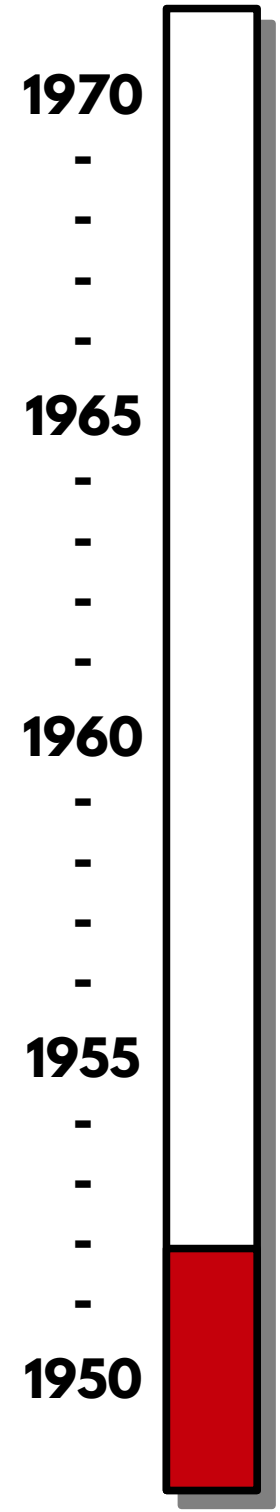
1952

ON SAIT QUE... L'ADN PORTE L'INFORMATION GÉNÉTIQUE.

DÉBAT CLOS PAR HERSHEY ET CHASE (1952)



Thomasione - Modified from German Wikipedia

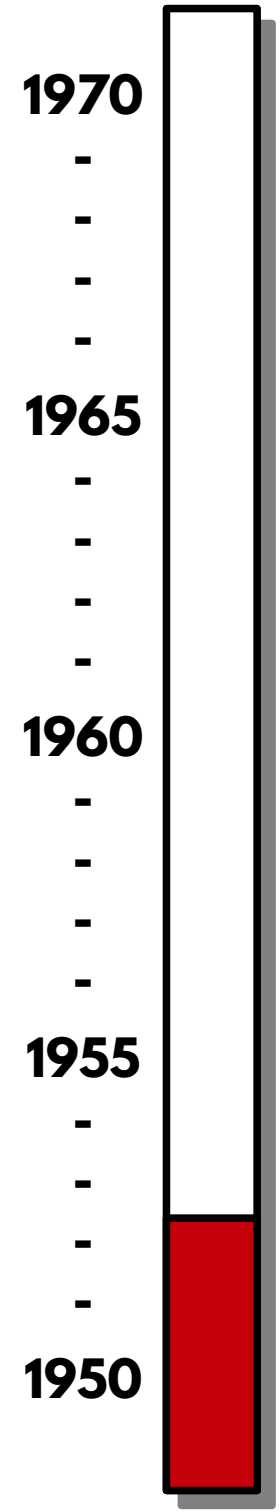


1952

ON SAIT QUE... PEU DE CHOSES SUR LA STRUCTURE DE L'ADN.

IL FAUDRA ATTENDRE :

- **1 an avant de connaître la structure de l'ADN;**
  (Watson et Crick 1953)
- **13 ans avant d'en déchiffrer le code;**
  (Nirenberg, Leder et al. 1965)
- **25 ans pour séquencer une molécule d'ADN.**
  (Gilbert et al. 1976 ; Sanger et al. 1977)



1955

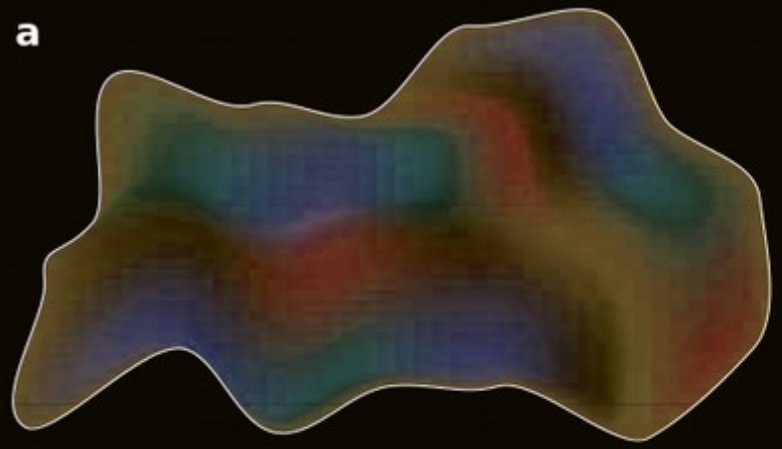
MAIS NOUS EN SAVIONS BEAUCOUP PLUS SUR... LES PROTÉINES.

1970
1965
1960
1955
1950

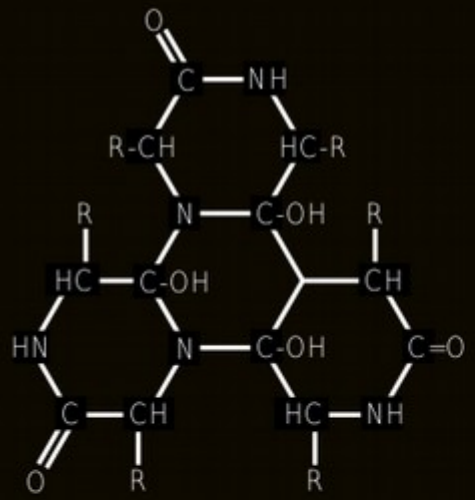
1955

MAIS NOUS EN SAVIONS BEAUCOUP PLUS SUR... LES PROTÉINES.

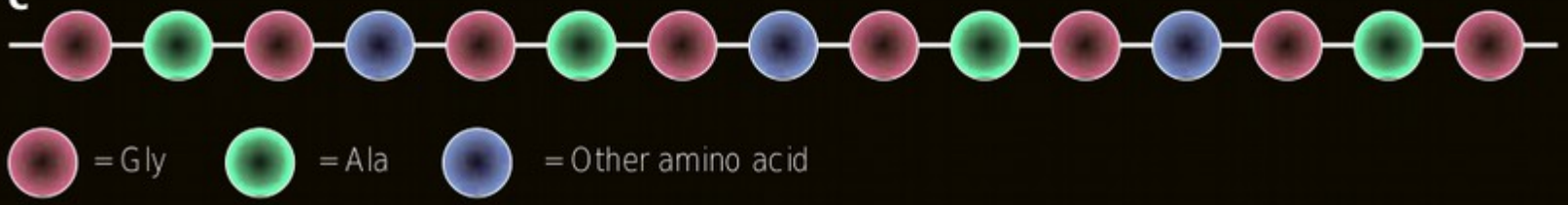
a



b



c



AVANT 1955:

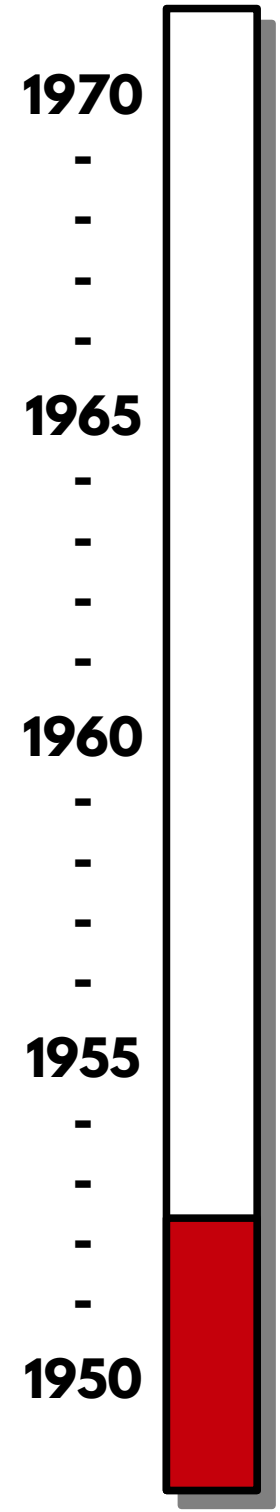
TROIS MODÈLES EN COMPÉTITION

a) Protéines = colloides amorphes sans structure moléculaire définie.

b) Structure polycyclique.

c) Protéines = simples chaînes répétitives d'acides aminés.

Hagen 2000 Nat Rev Genet



1955

MAIS NOUS EN SAVIONS BEAUCOUP PLUS SUR... LES PROTÉINES.

Découverte de la structure primaire de l'insuline
(Sanger et al. 1955)

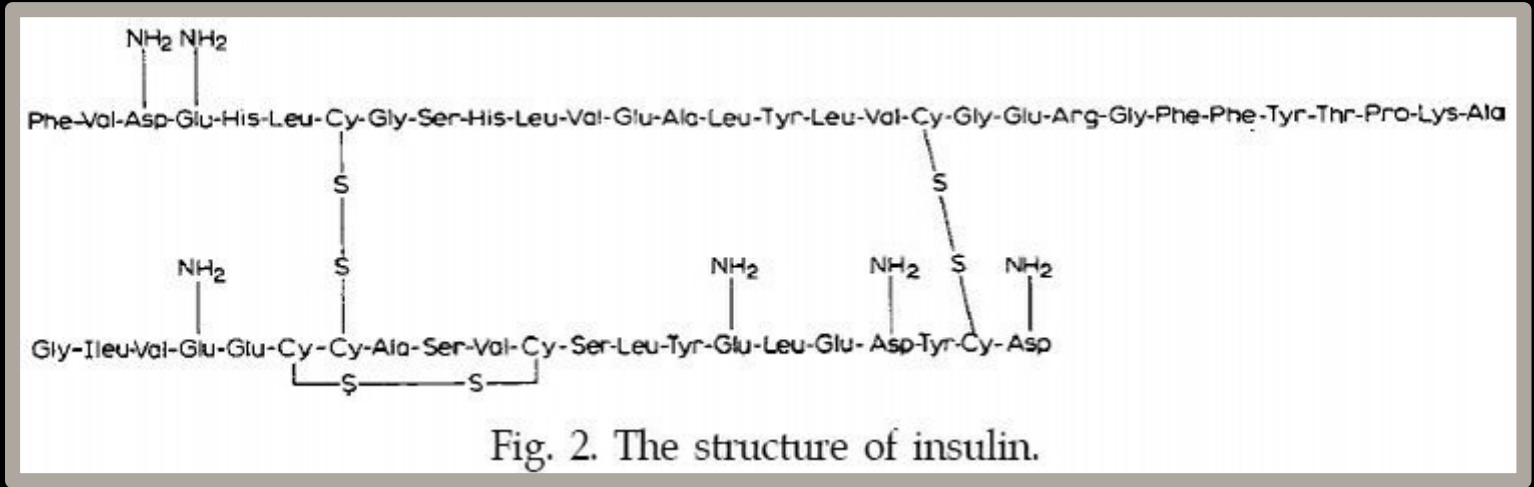


Fig. 2. The structure of insulin.

DÉBAT SUR LA NATURE POLYPEPTIDIQUE DES PROTÉINES: CLOS.

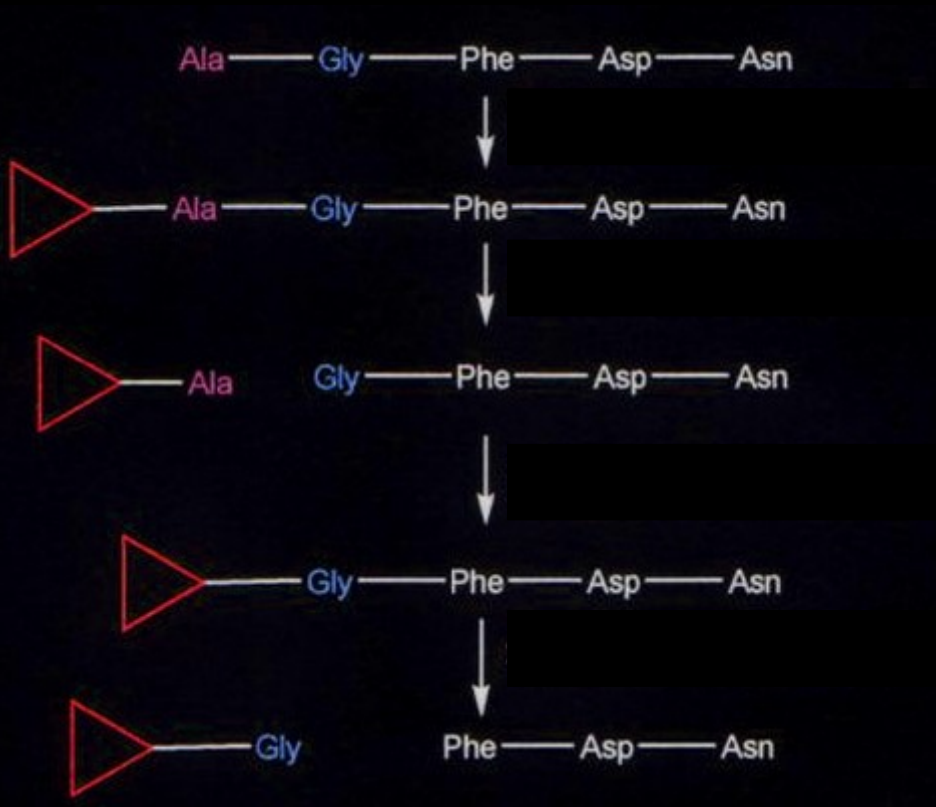
1970
1965
1960
1955
1950

1956-1960

OPTIMISATION DU SÉQUENÇAGE DES PROTÉINES.

RÉACTION DE DÉGRADATION D'EDMAN

= Séquençage à 1 a.a. à la fois, à partir du côté N-terminal.



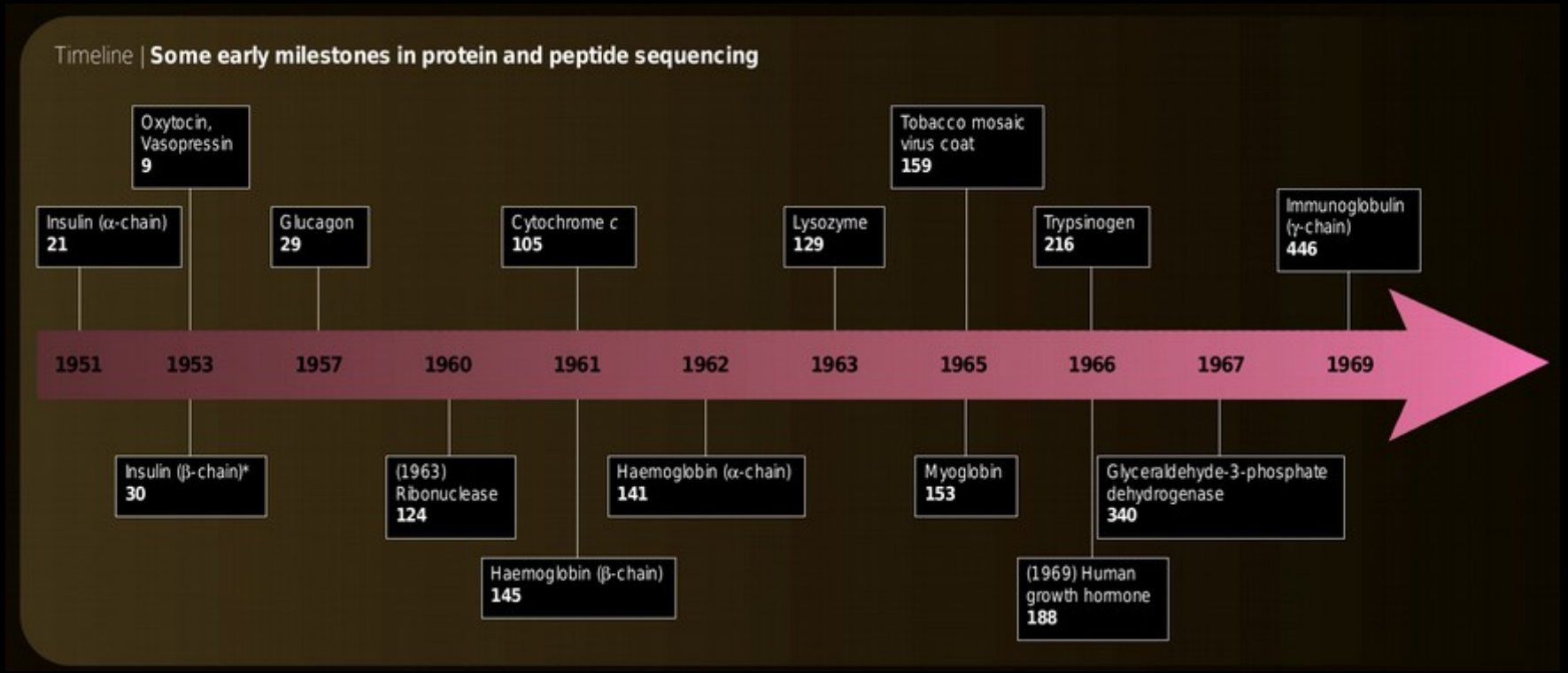
... MAINTENANT SEMI-AUTOMATISÉE !

1970
1965
1960
1955
1950

1960-1965

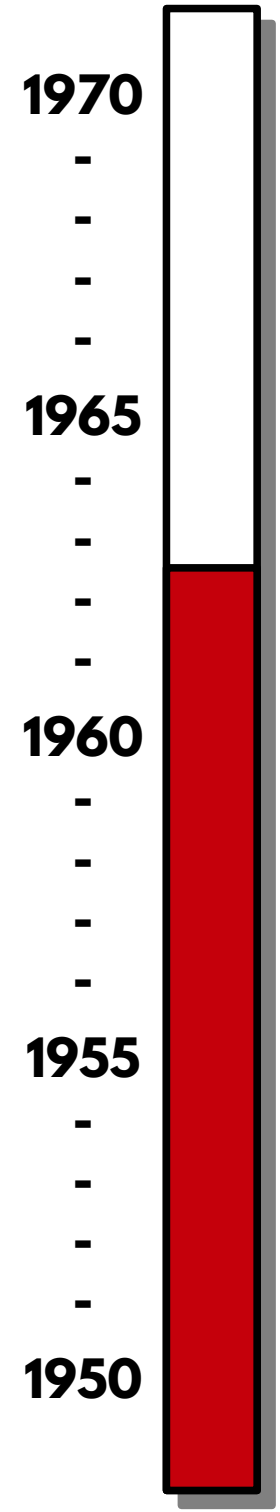
OPTIMISATION DU SÉQUENÇAGE DES PROTÉINES.

Dégradation d'Edman + Automatisation =



PUBLICATION MASSIVE DE SÉQUENCES DE PROTÉINES.

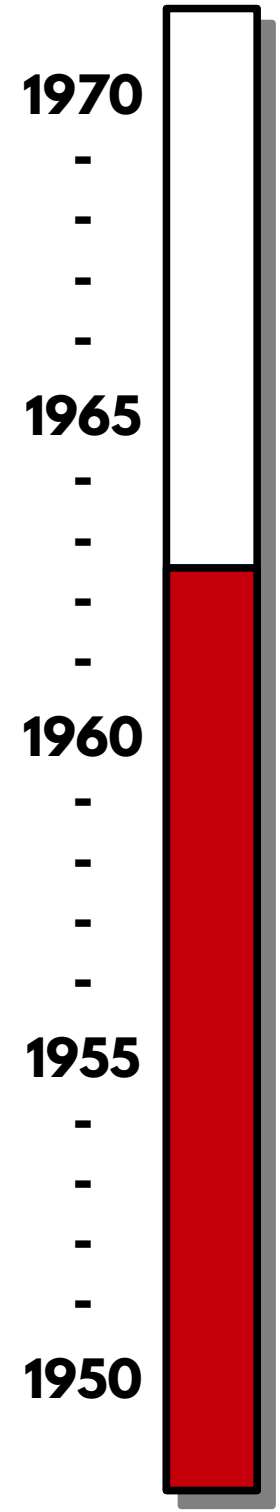
Hagen 2000 Nat Rev Genet



1960-1965

L'APOGÉE DU “PARADIGME DES PROTÉINES”

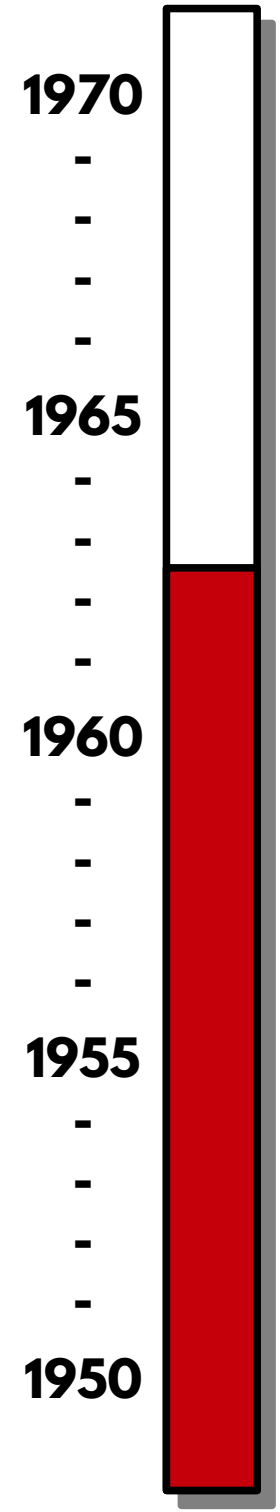
STRUCTURE → FONCTION



1960-1965

L'APOGÉE DU “PARADIGME DES PROTÉINES”

SÉQUENCE → **STRUCTURE** → **FONCTION**

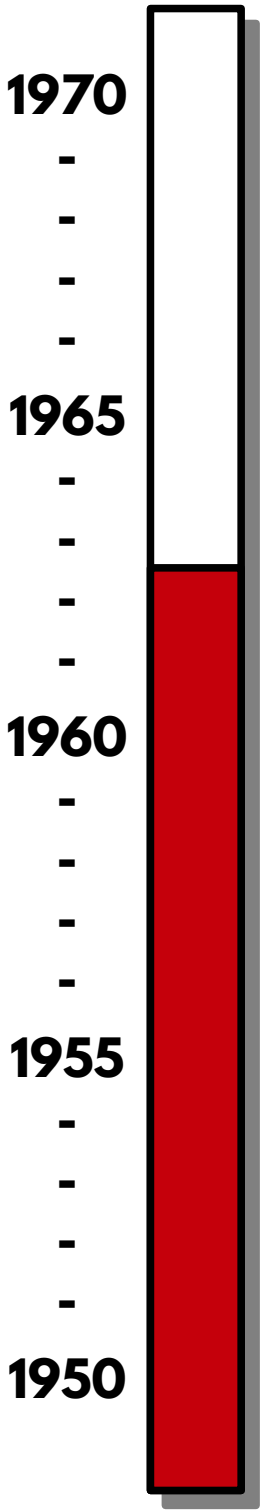


1960-1965

L'APOGÉE DU “PARADIGME DES PROTÉINES”

SÉQUENCE → **STRUCTURE** → **FONCTION**

AUTRE CONSTAT INTÉRESSANT…



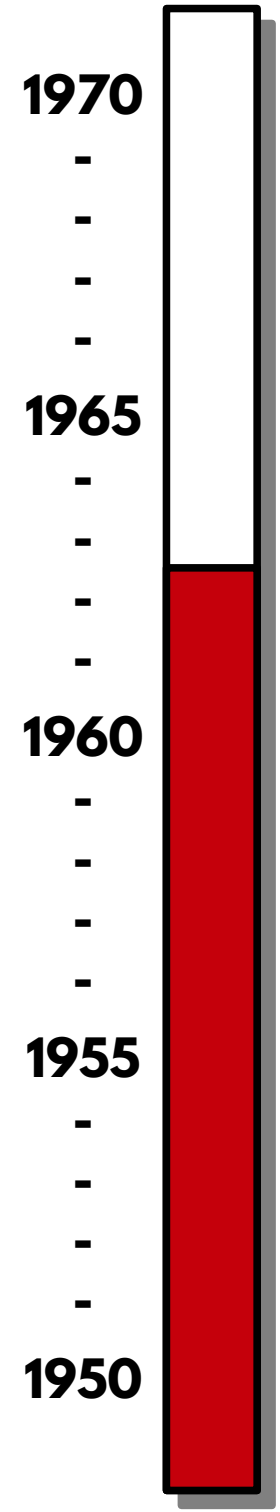
1960-1965

L'APOGÉE DU “PARADIGME DES PROTÉINES”

SÉQUENCE → **STRUCTURE** → **FONCTION**

AUTRE CONSTAT INTÉRESSANT...

	>MYOGLOBINE
Humain	NH_2-MIKTHECATATETHERATINTHEHAT...



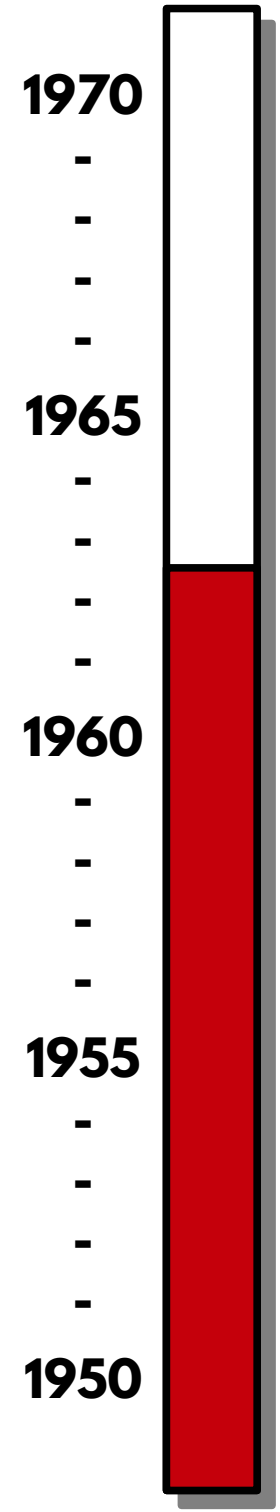
1960-1965

L'APOGÉE DU “PARADIGME DES PROTÉINES”

SÉQUENCE → **STRUCTURE** → **FONCTION**

AUTRE CONSTAT INTÉRESSANT...

	>MYOGLOBINE
Humain	NH_2-MIKTHECATATETHERATINTHEHAT...
Gorille	NH_2-MIKTHDCATATETHERGTINTHEHAT...



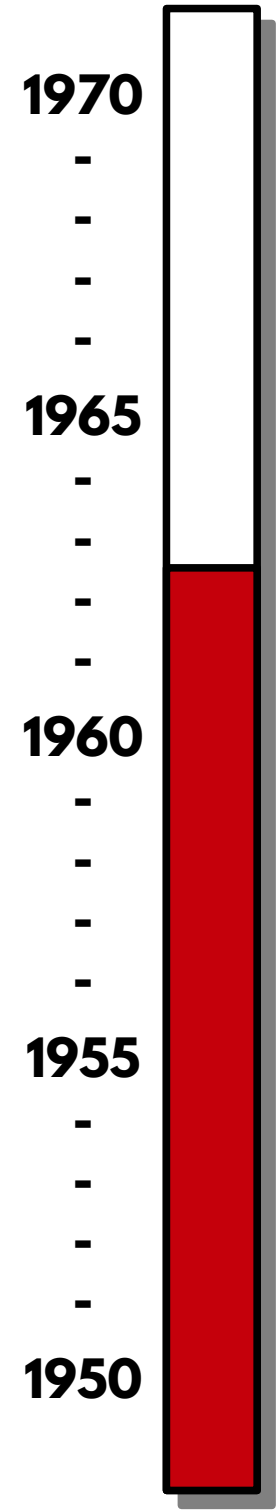
1960-1965

L'APOGÉE DU “PARADIGME DES PROTÉINES”

SÉQUENCE → **STRUCTURE** → **FONCTION**

AUTRE CONSTAT INTÉRESSANT...

	>MYOGLOBINE
Humain	NH_2-MIKTHECATATETHERATINTHEHAT...
Gorille	NH_2-MIKTHDCATATETHERGTINTHEHAT...
Chat	NH_2-MIKTHDCATATDTHERGTINTHEHAT...



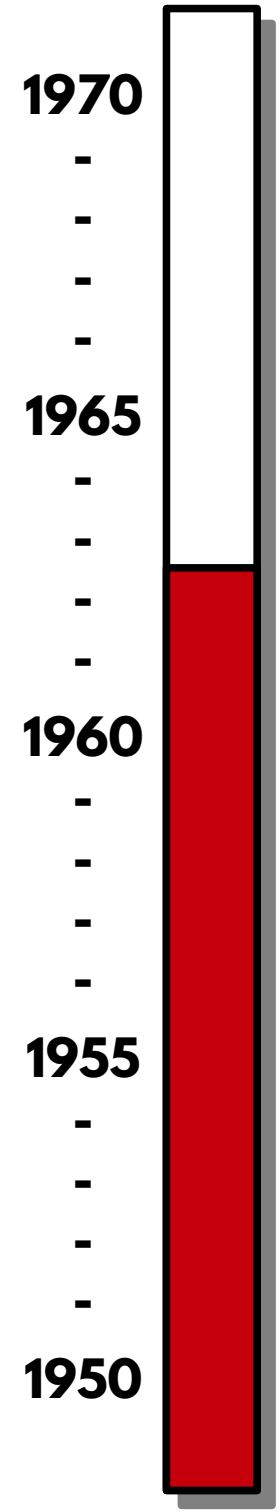
1960-1965

L'APOGÉE DU “PARADIGME DES PROTÉINES”

SÉQUENCE → **STRUCTURE** → **FONCTION**

AUTRE CONSTAT INTÉRESSANT...

	>MYOGLOBINE
Humain	NH_2-MIKTHECATATETHERATINTHEHAT...
Gorille	NH_2-MIKTHDCATATETHERGTINTHEHAT...
Chat	NH_2-MIKTHDCATATDTHERGTINTHEHAT...
Fourmi	NH_2-MLKTHDCATGTDTH-RGTIQTH--AT...



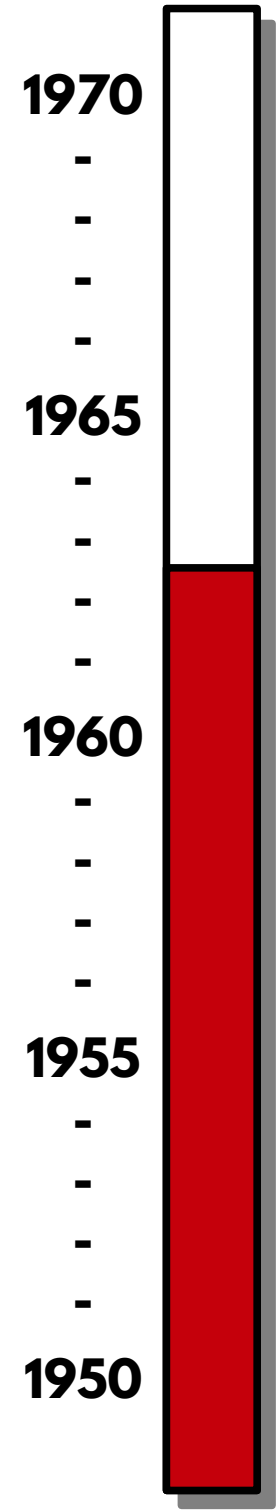
1960-1965

L'APOGÉE DU “PARADIGME DES PROTÉINES”

SÉQUENCE → **STRUCTURE** → **FONCTION**

AUTRE CONSTAT INTÉRESSANT...

	>MYOGLOBINE
Humain	NH_2-MIKTHECATATETHERATINTHEHAT...
Gorille	NH_2-MIKTHDCATATETHERGTINTHEHAT...
Chat	NH_2-MIKTHDCATATDTHERGTINTHEHAT...
Fourmi	NH_2-MLKTHDCATGTDTH-RGTIQTH--AT...
Pieuvre	NH_2-MLKTHDTGTGTDTH-RGTIQTH--AS...

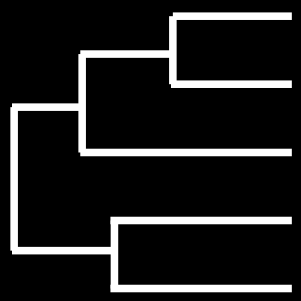


1960-1965

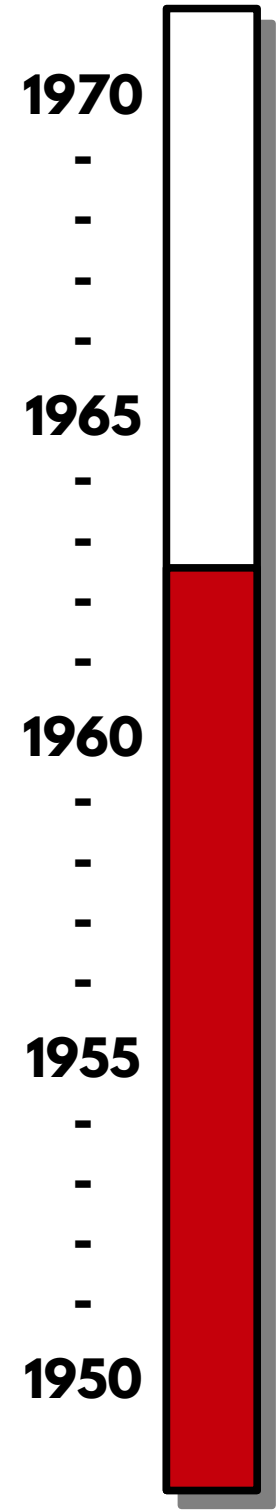
L'APOGÉE DU “PARADIGME DES PROTÉINES”

SÉQUENCE → **STRUCTURE** → **FONCTION**

AUTRE CONSTAT INTÉRESSANT...



```
                >MYOGLOBINE
Humain          NH2-MIKTHECATATETHERATINTHEHAT...
Gorille         NH2-MIKTHDCATATETHERGTINTHEHAT...
Chat            NH2-MIKTHDCATATDTHERGTINTHEHAT...
Fourmi          NH2-MLKTHDCATGTDTH-RGTIQTH--AT...
Pieuvre         NH2-MLKTHDTGTGTDTH-RGTIQTH--AS...
```

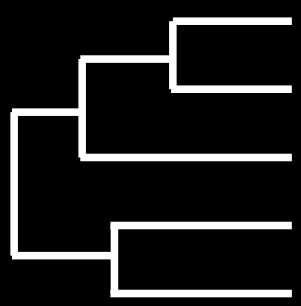


1960-1965

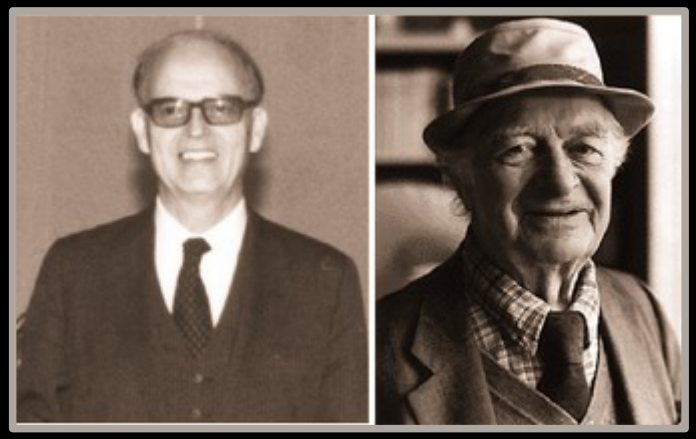
L'APOGÉE DU “PARADIGME DES PROTÉINES”

SÉQUENCE → **STRUCTURE** → **FONCTION**

AUTRE CONSTAT INTÉRESSANT...



```
                >MYOGLOBINE
Humain          NH2-MIKTHECATATETHERATINTHEHAT...
Gorille         NH2-MIKTHDCATATETHERGTINTHEHAT...
Chat            NH2-MIKTHDCATATDTHERGTINTHEHAT...
Fourmi          NH2-MLKTHDCATGTDTH-RGTIQTH--AT...
Pieuvre         NH2-MLKTHDTGTGTDTH-RGTIQTH--AS...
```



Emile ZUCKERKANDL

Linus PAULING

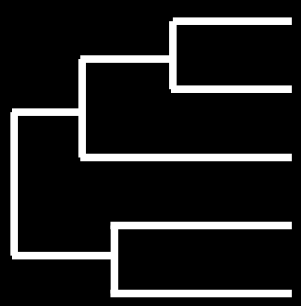
1970 - - - - 1965 - - - - - 1960 - - - - - 1955 - - - - 1950

1960-1965

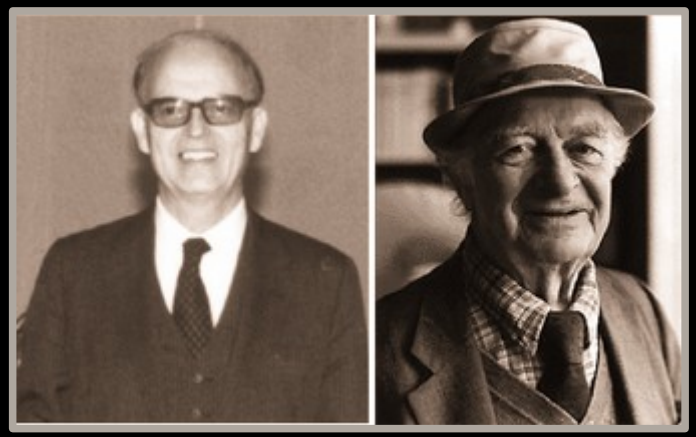
L'APOGÉE DU “PARADIGME DES PROTÉINES”

SÉQUENCE → **STRUCTURE** → **FONCTION**

AUTRE CONSTAT INTÉRESSANT...



```
                >MYOGLOBINE
Humain          NH2-MIKTHECATATETHERATINTHEHAT...
Gorille         NH2-MIKTHDCATATETHERGTINTHEHAT...
Chat            NH2-MIKTHDCATATDTHERGTINTHEHAT...
Fourmi          NH2-MLKTHDCATGTDTH-RGTIQTH--AT...
Pieuvre         NH2-MLKTHDTGTGTDTH-RGTIQTH--AS...
```



Emile ZUCKERKANDL

Linus PAULING

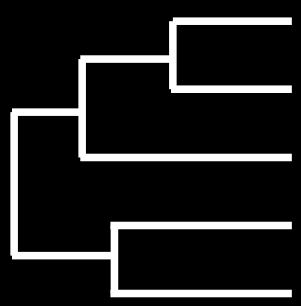
1970 - - - - 1965 - - - - - 1960 - - - - - 1955 - - - - - 1950

1960-1965

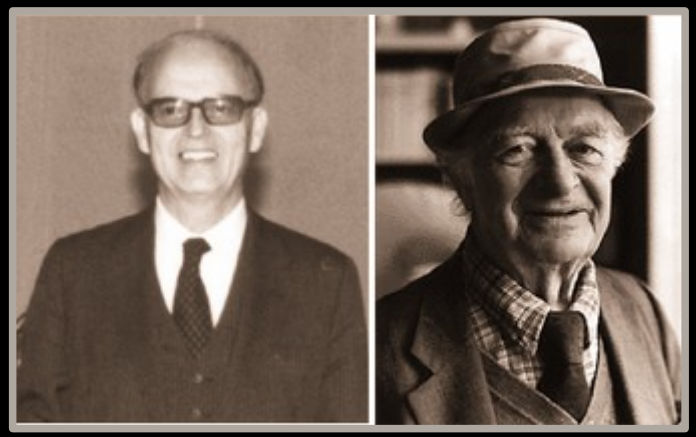
L'APOGÉE DU “PARADIGME DES PROTÉINES”

SÉQUENCE → **STRUCTURE** → **FONCTION**

AUTRE CONSTAT INTÉRESSANT...



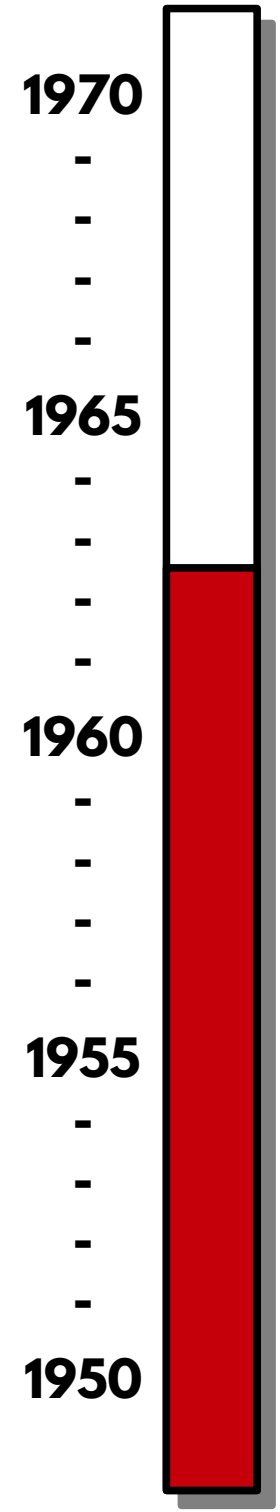
```
                >MYOGLOBINE
Humain          NH2-MIKTHECATATETHERATINTHEHAT...
Gorille         NH2-MIKTHDCATATETHERGTINTHEHAT...
Chat            NH2-MIKTHDCATATDTHERGTINTHEHAT...
Fourmi          NH2-MLKTHDCATGTDTH-RGTIQTH--AT...
Pieuvre         NH2-MLKTHDTGTGTDTH-RGTIQTH--AS...
```



Emile ZUCKERKANDL

Linus PAULING

Divergence entre séquences orthologues ~= Histoire évolutive des protéines (et des espèces) ?

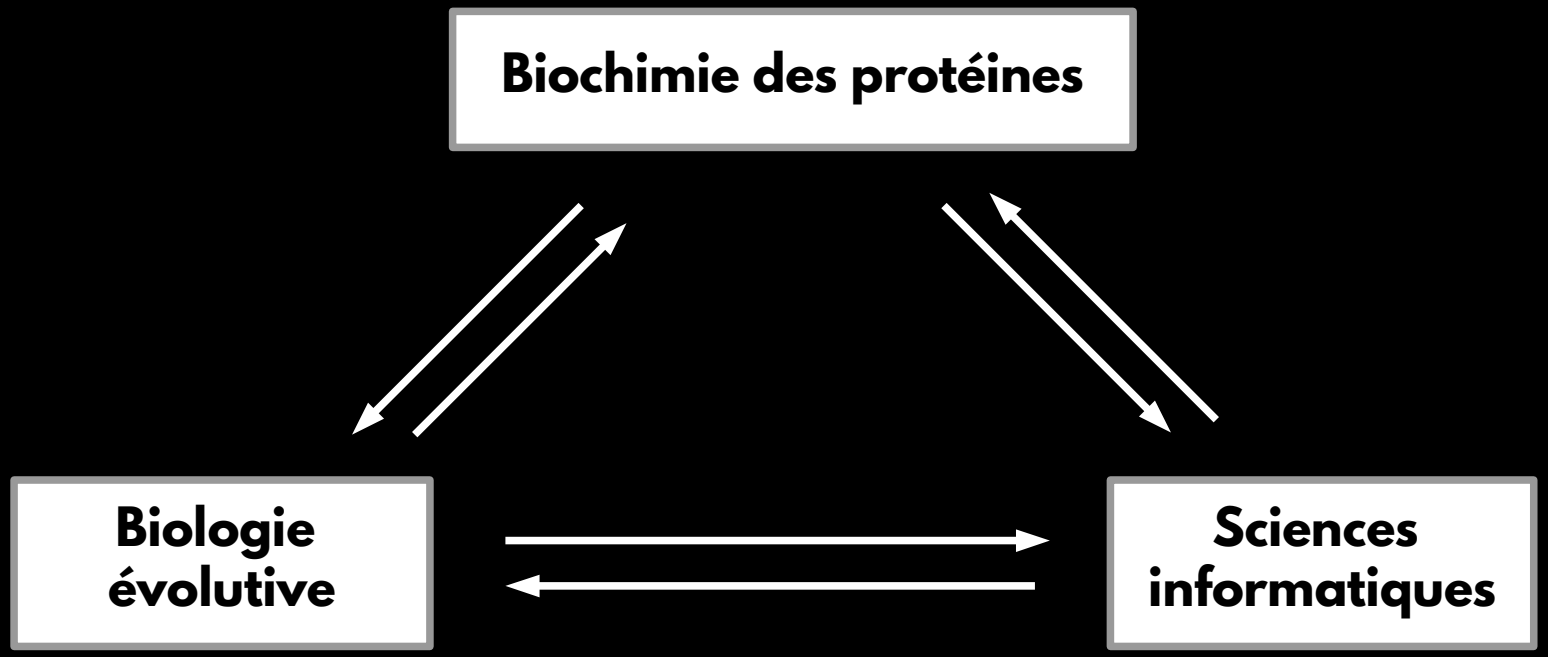


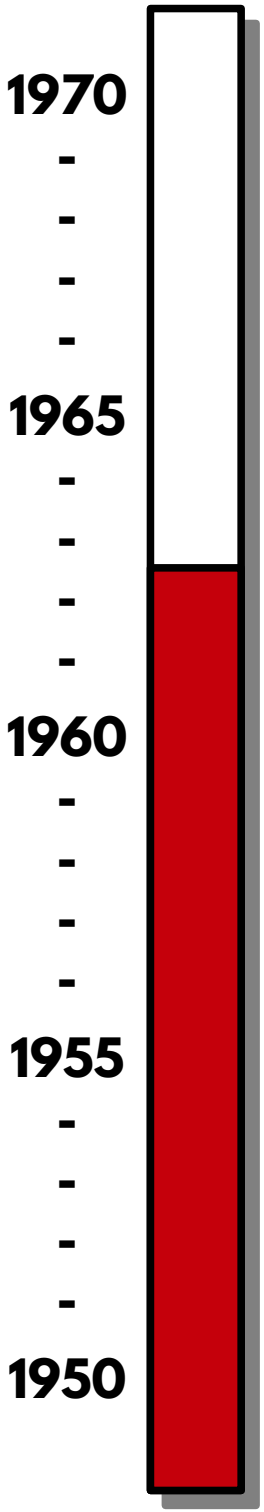
1960-1965

L'APOGÉE DU “PARADIGME DES PROTÉINES”

DÉBUT DE LA “ PALÉOGÉNÉTIQUE ”

Synthèse entre...



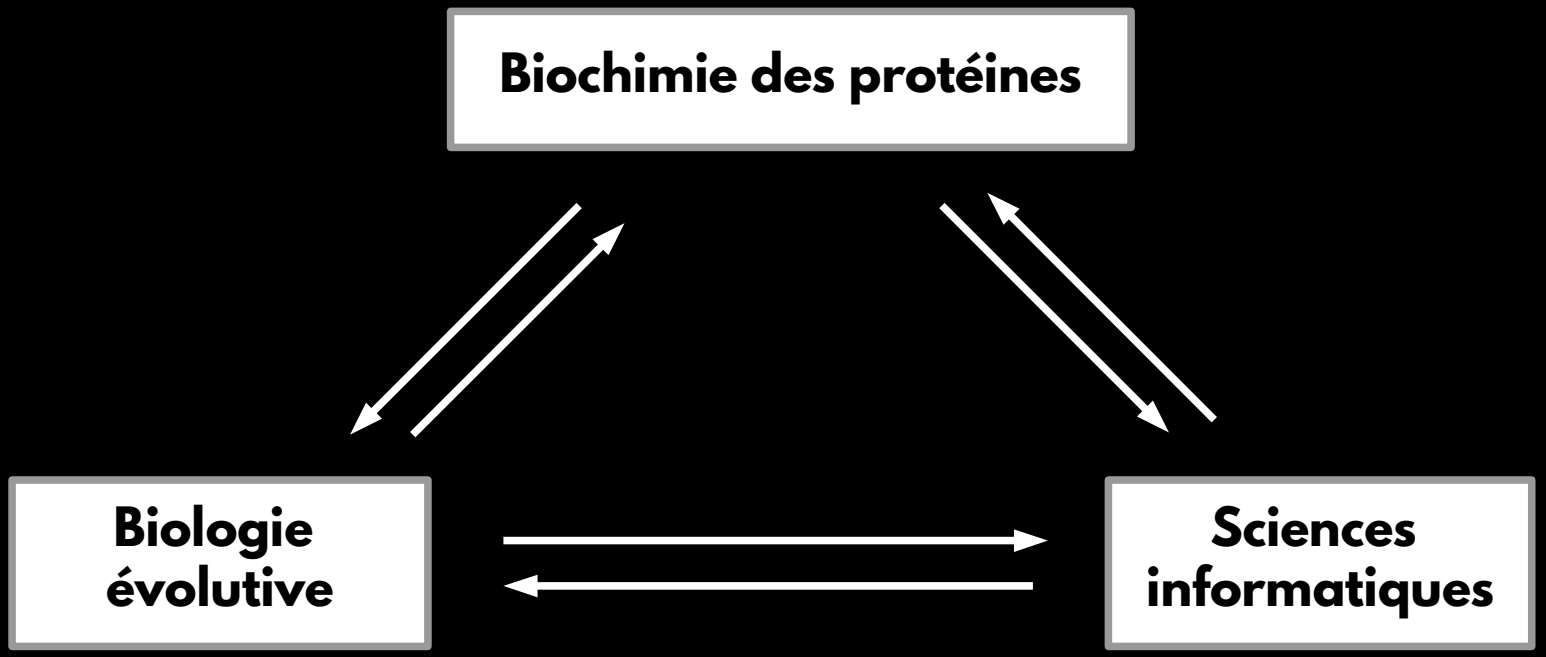


1960-1965

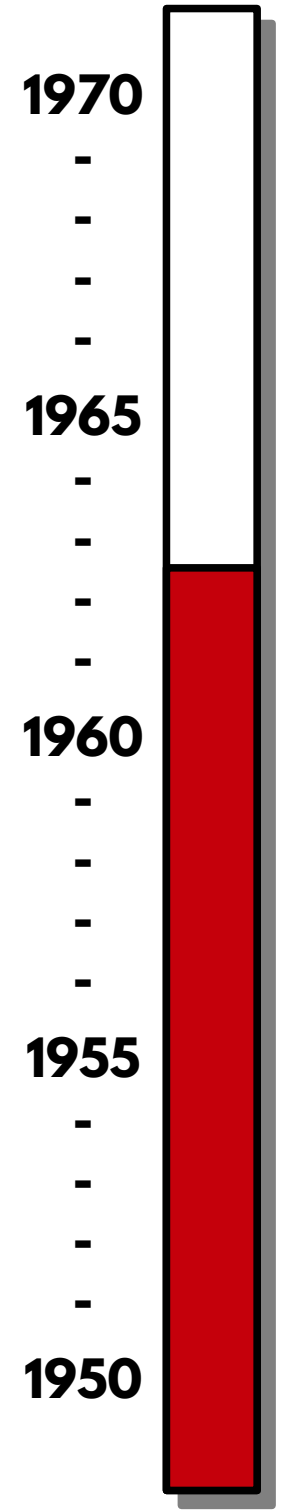
L'APOGÉE DU “PARADIGME DES PROTÉINES”

DÉBUT DE LA “ PALÉOGÉNÉTIQUE ”

Synthèse entre...



AU COEUR DE CETTE CHIMIE...



... SE TROUVE UNE FEMME.

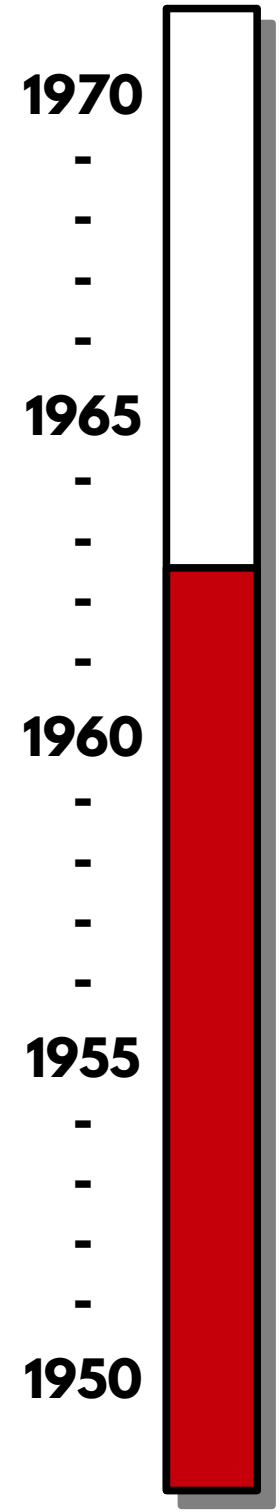
1970
1965
1960
1955
1950

1960-1965
GÉNÈSE DE LA “ BIOLOGIE COMPUTATIONNELLE ”

MARGARET OAKLEY DAYHOFF
(1925-1983)

“ ... the mother and father of bioinformatics”
- David Lipman (NCBI)

- **Directrice adjointe du National Biomedical Research Foundation (NBRF)**
- **A utilisé de manière extensive l'informatique lors de son Ph.D. en électrochimie.**
- **Voyait le potentiel de l'informatique en PALÉOGÉNÉTIQUE.**



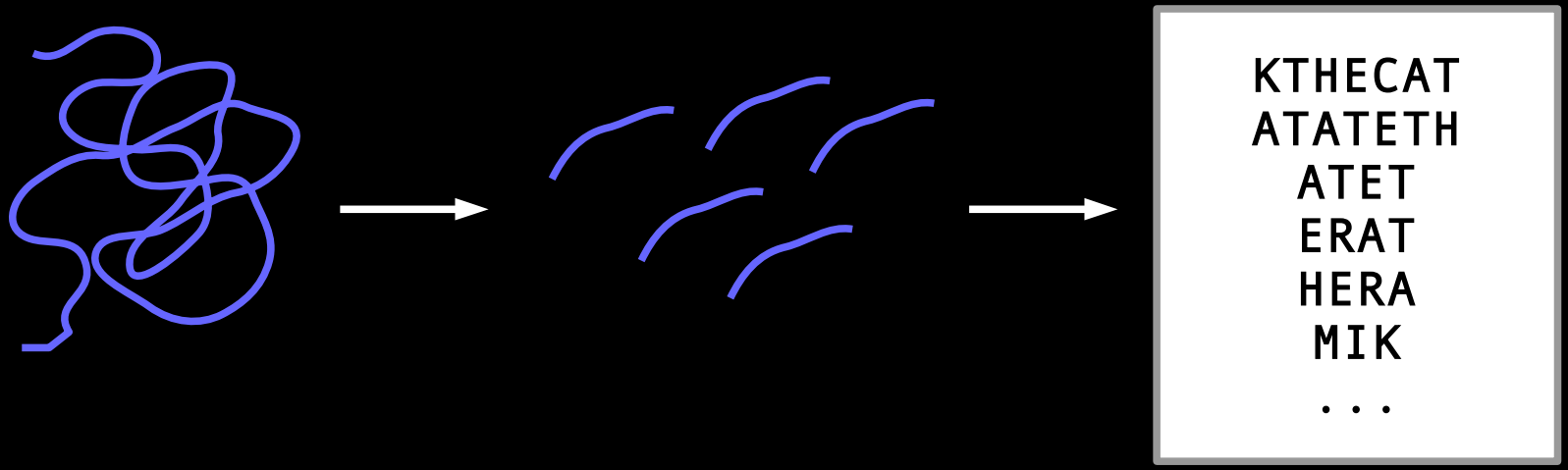
1960-1965

GÉNÈSE DE LA “ BIOLOGIE COMPUTATIONNELLE ”

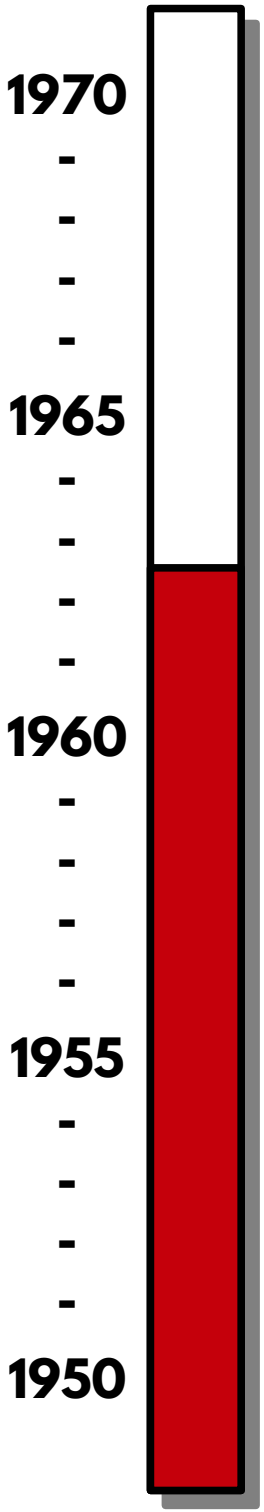
PRINCIPALES CONTRIBUTIONS DE M.O. DAYHOFF

DÉGRADATION D'EDMAN: *PROBLÈME*

- **Rendement: 98% = 50 a.a. MAX**
- **Une protéine > 50 a.a. doit être dégradée en plus petits peptides, qui eux seront séquencés.**



Séquence de la protéine complète ?



1960-1965

GÉNÈSE DE LA “ BIOLOGIE COMPUTATIONNELLE ”

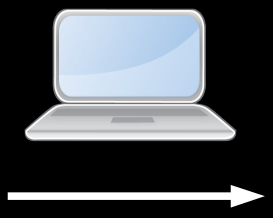
PRINCIPALES CONTRIBUTIONS DE M.O. DAYHOFF

1- Le premier “ outil bioinformatique ” (1962)

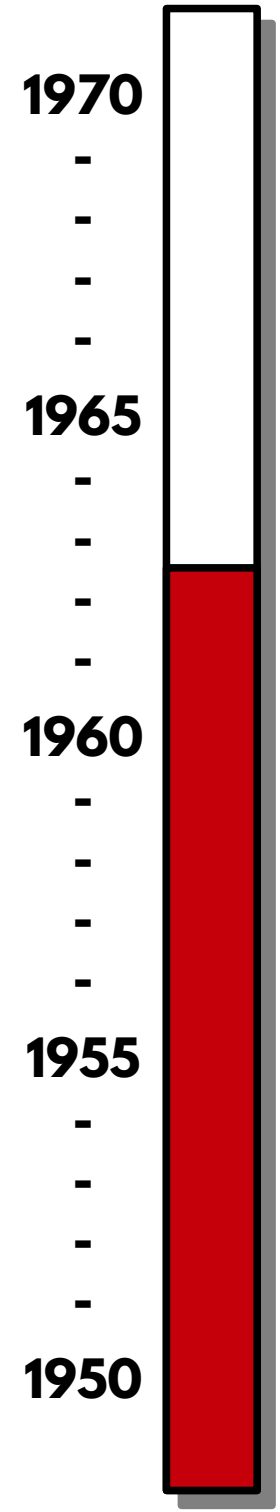
COMPROTEIN: A COMPUTER PROGRAM TO AID PRIMARY PROTEIN STRUCTURE DETERMINATION*

Margaret Oakley Dayhoff and Robert S. Ledley
National Biomedical Research Foundation
Silver Spring, Maryland

```
KTHECAT
ATATETH
  ATET
  ERAT
  HERA
  MIK
```



```
  KTHECAT                  Peptide
       ATATETH             Peptide
         ATET              Peptide
              ERAT         Peptide
             HERA          Peptide
MIK                        Peptide
--------------------
MIKTHECATATETHERAT         Protein
```



1960-1965

GÉNÈSE DE LA “ BIOLOGIE COMPUTATIONNELLE ”

FACILE ?

1970
1965
1960
1955
1950

1960-1965

GÉNÈSE DE LA “ BIOLOGIE COMPUTATIONNELLE ”

FACILE ?

Ordinateur typique de l'époque (IBM 7090)

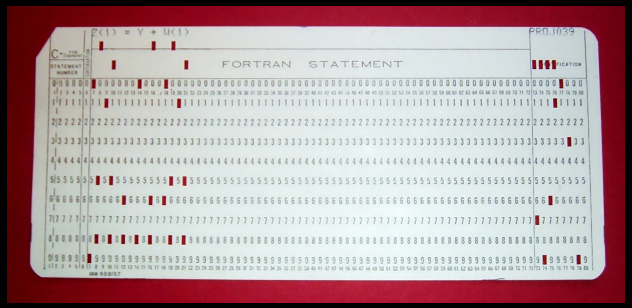
1970
1965
1960
1955
1950

1960-1965

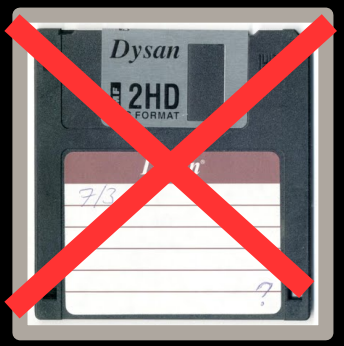
GÉNÈSE DE LA “ BIOLOGIE COMPUTATIONNELLE ”

FACILE ?

Ordinateur typique de l'époque (IBM 7090)



Une ligne de code en 1962 (FORTRAN)



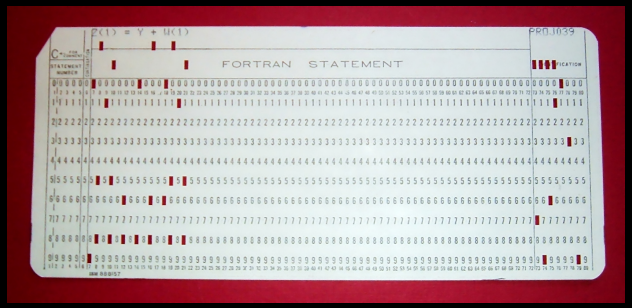
1970
-
-
-
-
1965
-
-
-
-
-
-
1960
-
-
-
-
-
-
1955
-
-
-
-
-
1950

1960-1965

GÉNÈSE DE LA “ BIOLOGIE COMPUTATIONNELLE ”

FACILE ?

Ordinateur typique de l'époque (IBM 7090)



Une ligne de code en 1962 (FORTRAN)

Code source d'un programme

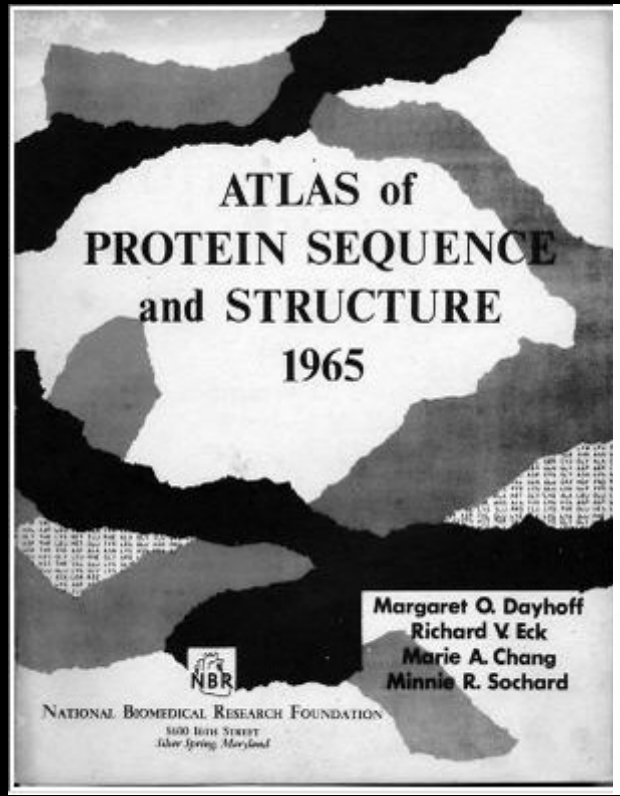
1970 - - - - 1965 - - - - - - - - 1960 - - - - - - - - 1955 - - - - - - - - 1950

1965

GÉNÈSE DE LA “ BIOLOGIE COMPUTATIONNELLE ”

PRINCIPALES CONTRIBUTIONS DE M.O. DAYHOFF

2- La première “base de données” de séquences de protéines…



BOVINE GROWTH HORMONE

```
          5        10        15        20        25        30
  1 A F P A M S L S G L F A N A V L R A Q H L H Q L A A D T F K
 31 E F E R T Y I P E G Q R Y S I Q N T Q V A F C F S E T I P A
 61 P T G K N E A Q Q K S D L E L L R I S L L L I Q S W L G P L
 91 Q F L S R V F T N S L V F G T S D R V Y E K L K D L E E G I
121 L A L M R E L E D G T P R A G Q I L K Q T Y D K F D T N M R
151 S D D A L L K N Y G L L S C F R K D L H K T E T Y L R V M K
181 C R R F G E A S C A R
```

COMPOSITION

15 ALA A	11 GLN Q	27 LEU L	13 SER S
13 ARG R	13 GLU E	11 LYS K	12 THR T
6 ASN N	10 GLY G	4 MET M	1 TRP W
10 ASP D	3 HIS H	13 PHE F	6 TYR Y
4 CYS C	7 ILE I	6 PRO P	6 VAL V

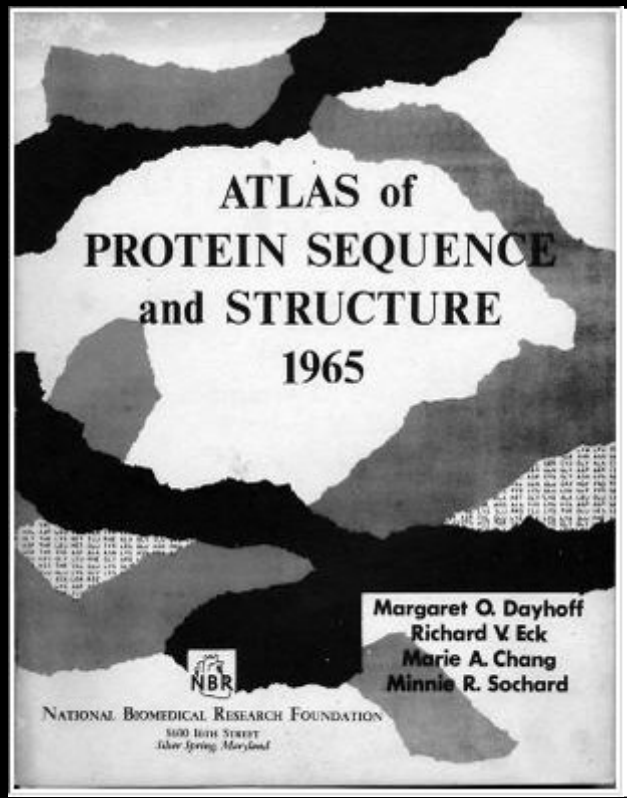
MOL. WT. = 21,816 NUMBER OF RESIDUES = 191

1970
1965
1960
1955
1950

1965
GÉNÈSE DE LA “ BIOLOGIE COMPUTATIONNELLE ”

PRINCIPALES CONTRIBUTIONS DE M.O. DAYHOFF

2- La première “base de données” de séquences de protéines...



Au début...

BEAUCOUP de variants interspécifiques d'une DIZAINE de protéines.

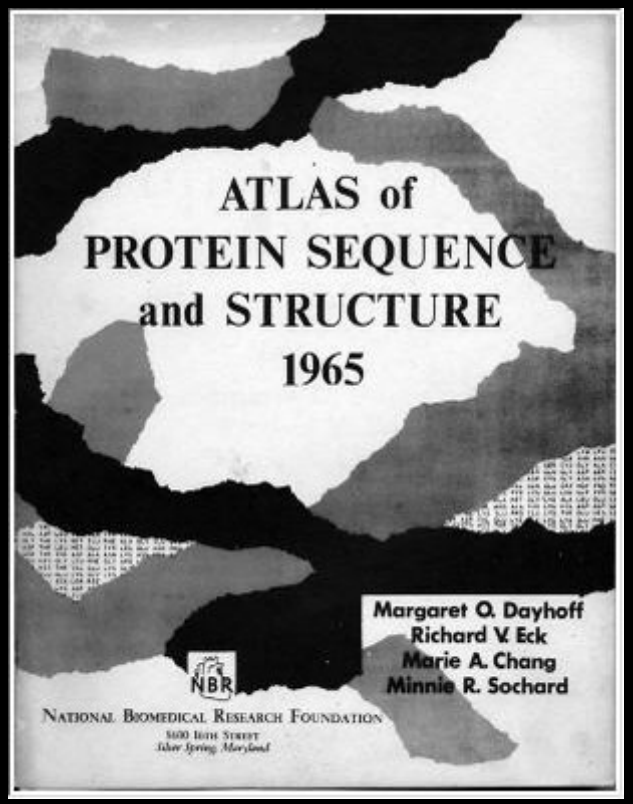
```
>PROT1
Humain     NH2-MIKTHECATATETHERATINTHE...-COOH
Gorille    NH2-MIKTHDCATATETHERGTINTHE...-COOH
Chat       NH2-MIKTHDCATATDTHERGTINTHE...-COOH
Fourmi     NH2-MLKTHDCATGTDTHRGTIQTHAT...-COOH
Pieuvre    NH2-MLKTHDTGTGTDTHRGTIQTHAS...-COOH
...
S. cerev   NH2-MLKNHDTGTGTDTHRGTIQTHAS...-COOH
E. coli    NH2-MLKTHNDTGTGARQTGTIQTHAS...-COOH
```

1970
1965
1960
1955
1950

1965
GÉNÈSE DE LA “ BIOLOGIE COMPUTATIONNELLE ”

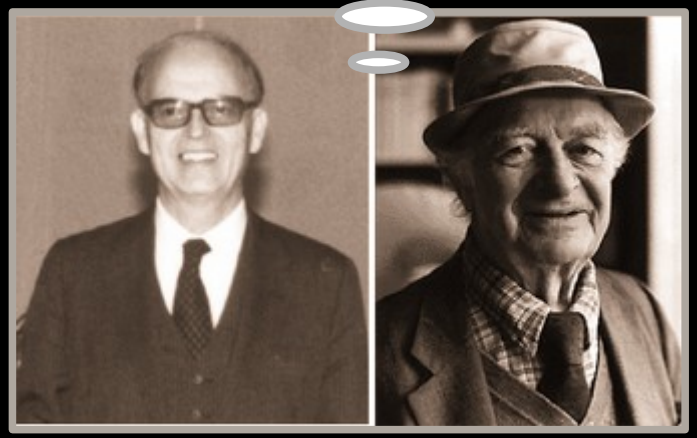
PRINCIPALES CONTRIBUTIONS DE M.O. DAYHOFF

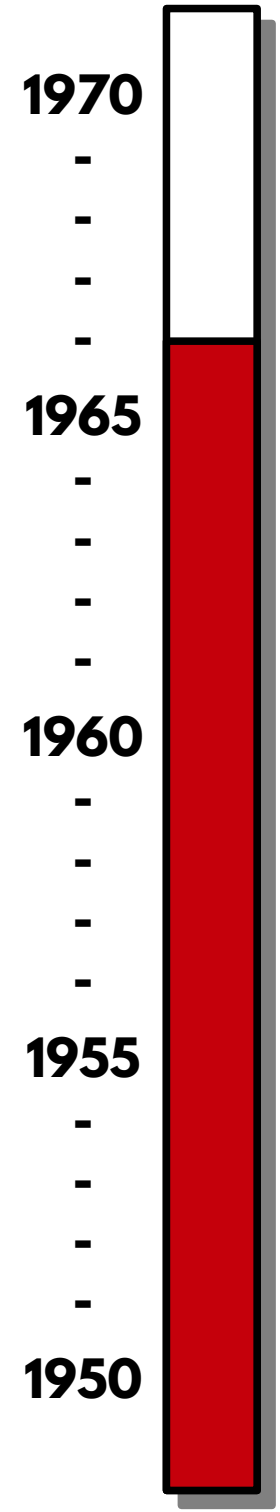
2- La première “base de données” de séquences de protéines...



Données idéales pour...

ancestry™

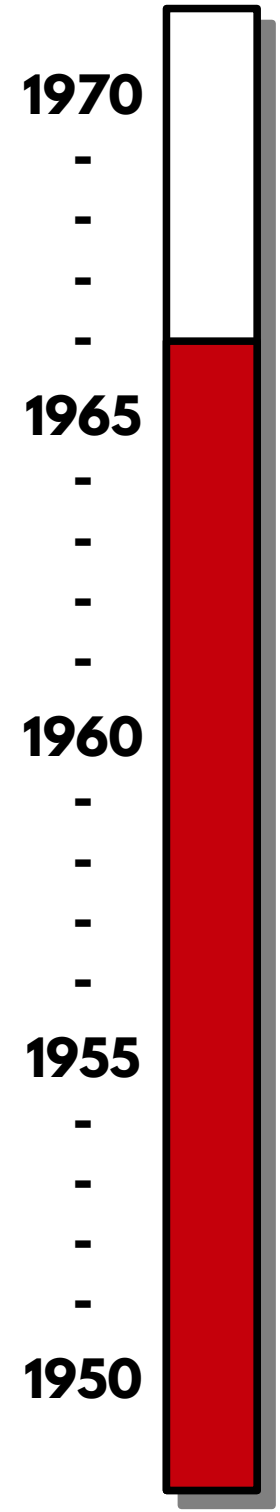




1966
DÉBUTS DE LA PHYLOGÉNIE MOLÉCULAIRE

Exemple...

Protéine 1 AVQH
Protéine 2 ALQH
Protéine 3 AIQH



1966
DÉBUTS DE LA PHYLOGÉNIE MOLÉCULAIRE

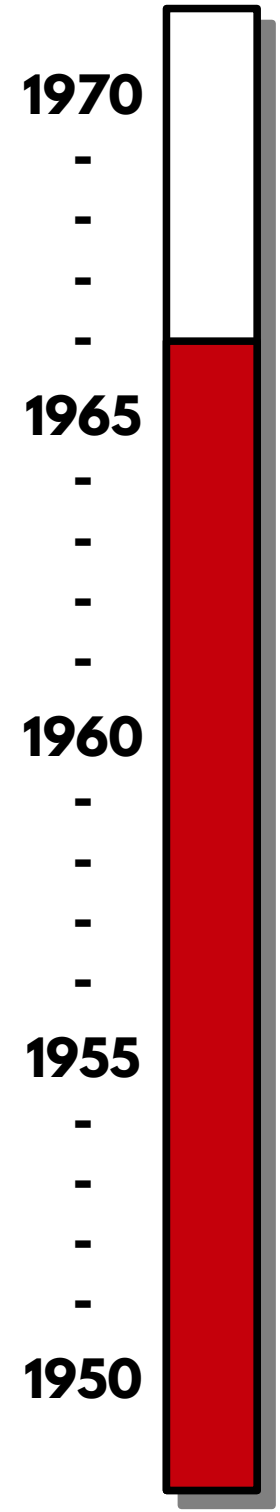
Exemple...

```
Protéine 1 AVQH
Protéine 2 ALQH
Protéine 3 AIQH
```

Phylogénie ? Trois possibilités :

Quelle hypothèse est la bonne?

Vérifions la distance entre les séquences
À L'AIDE D'ALIGNEMENTS.



1966
DÉBUTS DE LA PHYLOGÉNIE MOLÉCULAIRE

PROBLÈME :

```
Prot1   AVQH
        : ::
Prot2   ALQH

Prot3   AIQH
        : ::
Prot2   ALQH

Prot3   AIQH
        : ::
Prot1   AVQH
```

Tous ces alignements sont “ÉQUIVALENTS”

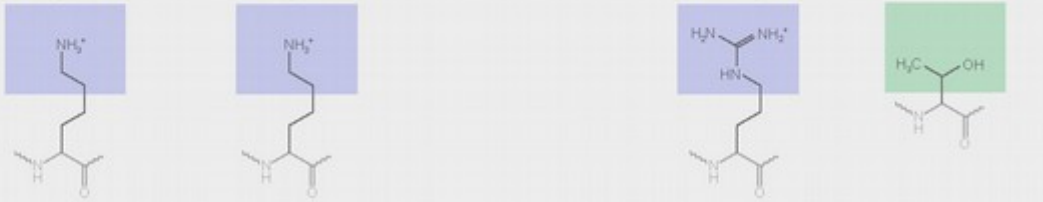
A.A. similaires, mais non-identiques.

Comment alors quantifier la distance entre `Prot1`, `Prot2` et `Prot3` ?

1970
1965
1960
1955
1950

1966
DÉBUTS DE LA PHYLOGÉNIE MOLÉCULAIRE

	No mutation	Point mutations			
		Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	Lys	Lys	STOP	Arg	Thr



+ + +
fréquent

- - -
fréquent

1970
1965
1960
1955
1950

1966
DÉBUTS DE LA PHYLOGÉNIE MOLÉCULAIRE

Vu que...

Séquence => Structure => Fonction

Certaines mutations ponctuelles
ont plus de chance de
se produire que d'autres.

1970
1965
1960
1955
1950

1966
DÉBUTS DE LA PHYLOGÉNIE MOLÉCULAIRE

CECI IMPLIQUE QUE...

Plus deux séquences comportent de substitutions rares, plus elles sont dissimilaires.

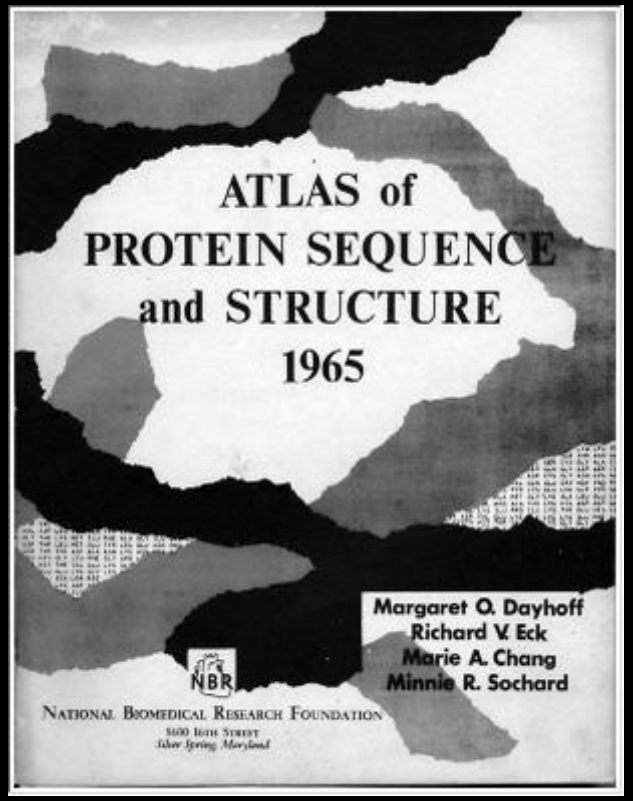
= POSSIBLE de discriminer des ALIGNEMENTS ÉQUIVALENTS.

1970
1965
1960
1955
1950

1966
DÉBUTS DE LA PHYLOGÉNIE MOLÉCULAIRE

PRINCIPALES CONTRIBUTIONS DE M.O. DAYHOFF

3- MATRICES DE SUBSTITUTION



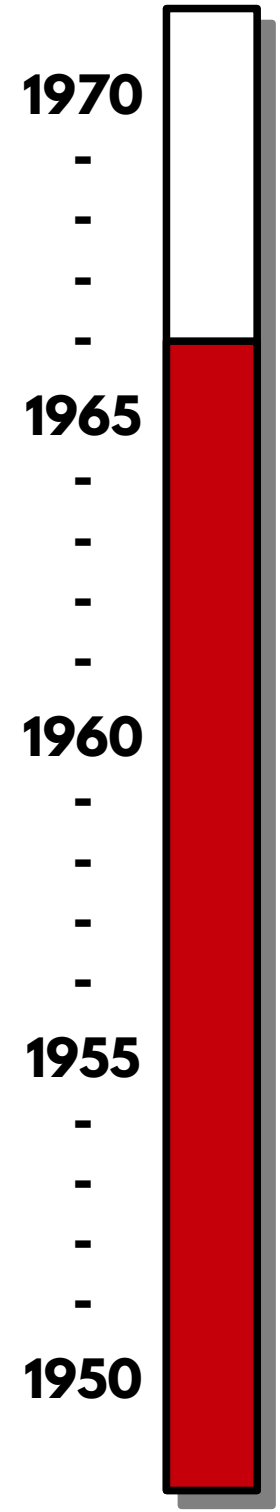
PAM (complétée en 1978!)
À partir de 1572 substitutions observées
dans 71 familles de protéines
partagent >85% d'identité

a.a. original

a.a. substitut

	A Ala	R Arg	N Asn	D Asp	C
A	9867	2	9	10	3
R	1	9913	1	0	1
N	4	1	9822	36	0
D	6	0	42	9859	0
C	1	1	0	0	9
Q	3	9	4	5	0

$10^4 \times P$ (Arg > Cys)



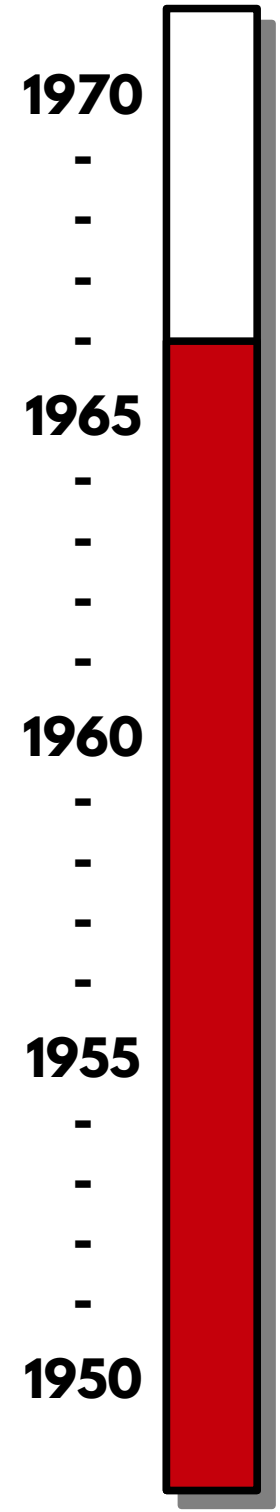
1966

DÉBUTS DE LA PHYLOGÉNIE MOLÉCULAIRE

RETOURNONS À NOS TROIS PROTÉINES...

Calcul des probabilités de substitution avec PAM1

```
Prot1  AVQH
       : ::
Prot2  ALQH

Prot3  AIQH
       : ::
Prot2  ALQH

Prot3  AIQH
       : ::
Prot1  AVQH
```



1966
DÉBUTS DE LA PHYLOGÉNIE MOLÉCULAIRE

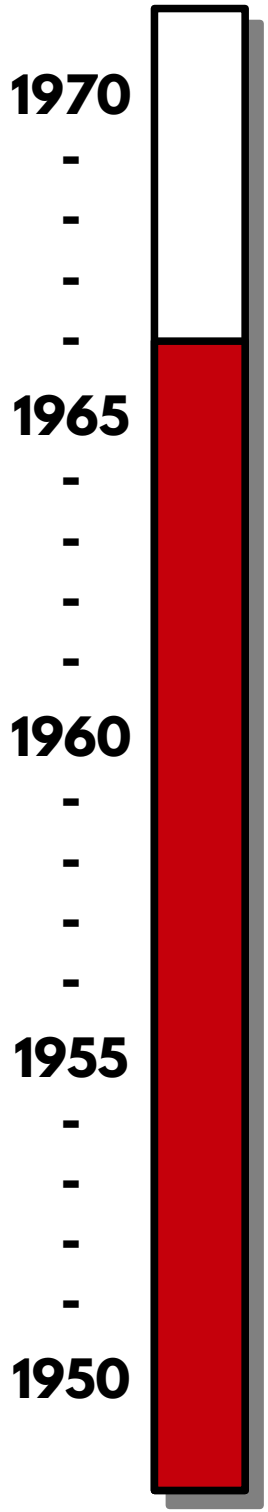
RETOURNONS À NOS TROIS PROTÉINES...
Calcul des probabilités de substitution avec PAM1

```
Prot1   AVQH
          : ::
Prot2   ALQH

Prot3   AIQH
          : ::
Prot2   ALQH

Prot3   AIQH
          : ::
Prot1   AVQH
```

$$P(V > L) = 0.0015$$
$$P(L > V) = 0.0011$$

$$P(I > L) = 0.0022$$
$$P(L > I) = 0.0009$$

$$P(I > V) = 0.0057$$
$$P(V > I) = 0.0033$$



1966

DÉBUTS DE LA PHYLOGÉNIE MOLÉCULAIRE

RETOURNONS À NOS TROIS PROTÉINES...

Calcul des probabilités de substitution avec PAM1

```
Prot1   AVQH
          : ::
Prot2   ALQH

Prot3   AIQH
          : ::
Prot2   ALQH

Prot3   AIQH
          : ::
Prot1   AVQH
```

$P(V > L) = 0.0015$

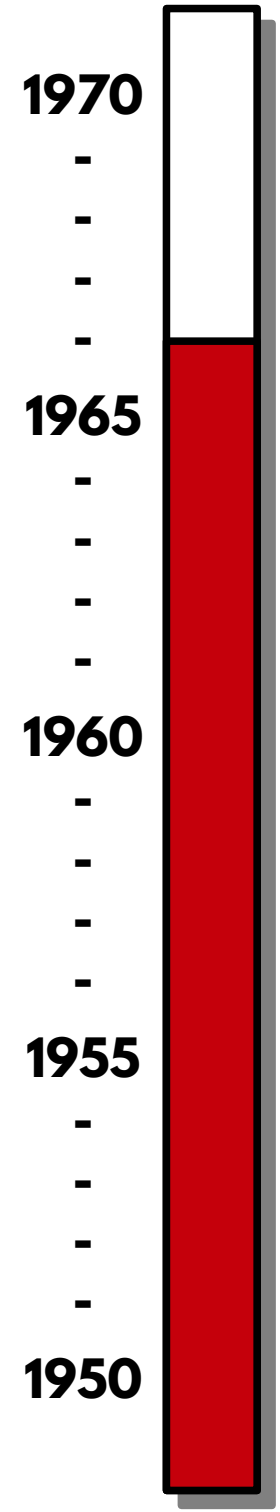
$P(L > V) = 0.0011$

$P(I > L) = 0.0022$

$P(L > I) = 0.0009$

$P(I > V) = 0.0057$

$P(V > I) = 0.0033$



1966

DÉBUTS DE LA PHYLOGÉNIE MOLÉCULAIRE

RETOURNONS À NOS TROIS PROTÉINES...

Calcul des probabilités de substitution avec PAM1

```
Prot1   AVQH
          : ::
Prot2   ALQH

Prot3   AIQH
          : ::
Prot2   ALQH

Prot3   AIQH
          : ::
Prot1   AVQH
```

$P(V > L) = 0.0015$

$P(L > V) = 0.0011$

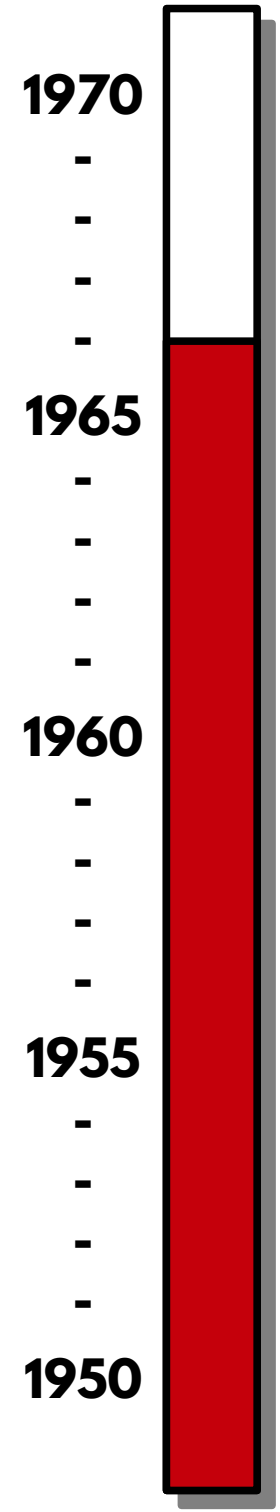
$P(I > L) = 0.0022$

$P(L > I) = 0.0009$

$P(I > V) = 0.0057$

$P(V > I) = 0.0033$

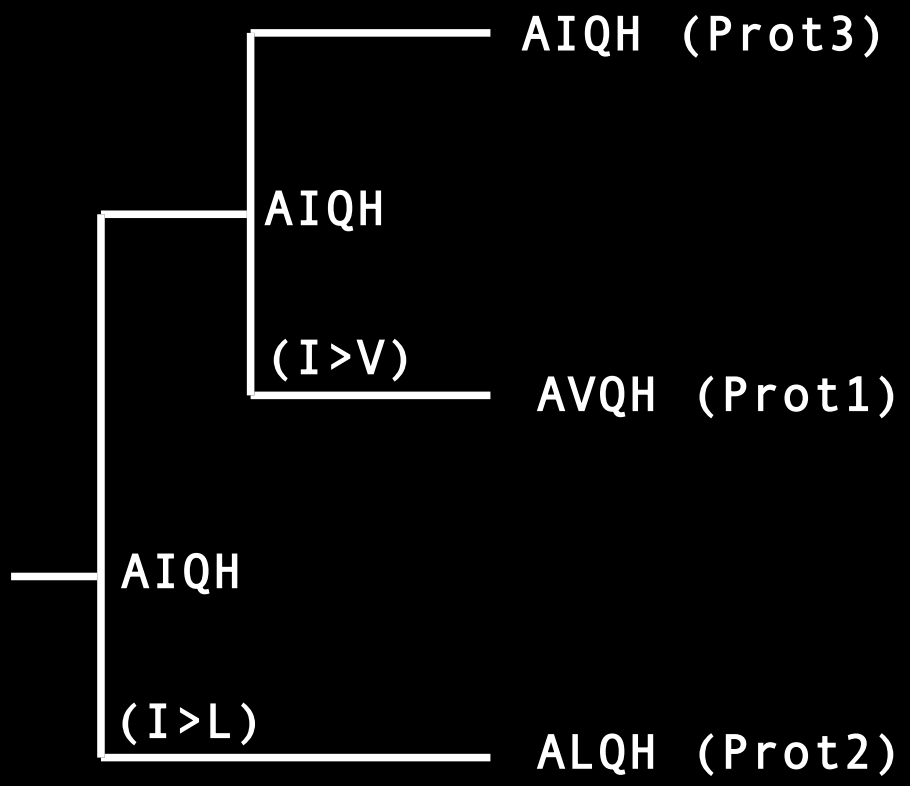
LES PLUS SIMILAIRES



1966
DÉBUTS DE LA PHYLOGÉNIE MOLÉCULAIRE

RETOURNONS À NOS TROIS PROTÉINES...
Calcul des probabilités de substitution avec PAM1

```
Prot1   AVQH
          : ::
Prot2   ALQH

Prot3   AIQH
          : ::
Prot2   ALQH

Prot3   AIQH
          : ::
Prot1   AVQH
```



Hypothèse la plus PROBABLE


DE RETOUR EN 2015 ...

La session prochaine...

PARTIE II (1971-1990)

- MINIATURISATION DE L'INFORMATIQUE;
- ARRIVÉE DU SÉQUENÇAGE DE L'ADN;
- PREMIERS RÉSEAUX D'ORDINATEURS;
- DÉBUTS DE L'ÈRE GÉNOMIQUE.