Génome humain

3 000 000 000 bases:
10000 caractères / feuille
300 000 feuilles
600 tomes, épaisseur: 30 mètres

Le savant doit ordonner; on fait la science avec des faits comme une maison avec des pierres; mais une accumulation de faits n'est pas plus une science qu'un tas de pierres n'est une maison.

H. Poincaré

Analyse de séquence

- Codes IUPAC

Amino acid codes

A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic acid
C	Cys	Cysteine
Q	Gln	Glutamine
E	Glu	Glutamic acid
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
L	Leu	Leucine
K	Lys	Lysine
M	Met	Methionine
F	Phe	Phenylalanine
P	Pro	Proline
S	Ser	Serine
T	Thr	Threonine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
V	Val	Valine
B	Asx	Aspartic acid or Asparagine
Z	Glx	Glutamine or Glutamic acid
X	Xaa	Any amino acid

Nucleic acid codes

A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
R	Purine (A or G)
Y	Pyrimidine (C, T, or U)
M	C or A
K	T, U, or G
W	T, U, or A
S	C or G
B	C, T, U, or G (not A)
D	A, T, U, or G (not C)
H	A, T, U, or C (not G)
V	A, C, or G (not T, not U)
N	Any base

Machine manipule des ***chaines de caractères!***

Analyse de séquence

- Carte de restriction

```
                          30-CviJI
                          4-HaeIII                2-BtsI
                         3-Sau96I                 4-TspRI   3-MaeIII
                        1-Cac8I                 12-MnlI      2-HpyCH4III
     1-NspI           16-BslI                 2-MaeIII   4-Sau96I
     12-NlaIII      13-NlaIII                 2-Tsp45I   1-AvaII
     V              V V VVV                   V V  V     V   V
481 ATGCTGGGAAGTGCCATGAGCAGGCCCATCATACATTTCGGCAGTGACTATGAGGACCGT   540
          490       500       510       520       530
```

- Traduction conceptuelle
  - Double brin ADN = 6 phases de lectures
  - Phase ouverte de lecture (ORF)

```
Phase N0    3
>Seq_perso    530 a   712   Phase 2 (sens direct)     61 aa
     M  R  T  V  T  I  V  K  T  C  T  V  T  P  T  K  C  T  T  G  P  W
    ATGAGGACCGTTACTATCGTGAAAACATGCACCGTTACCCCAACCAAGTGTACTACAGGCCCATGG
  530       540       550       560       570       580       590
     A  T  R  T  T  L  C  T  T  A  S  I  S  Q  S  S  S  T  R  S  P  Q
    GCAACCAGAACAACTTTGTGCACGACTGCGTCAATATCACAATCAAGCAGCACACGGTCACCACAA
  605       615       625       635       645       655       665
     G  R  T  S  P  R  P  T  L  R  *
    GGGAGAACTTCACCGAGACCGACGTTAAGATGA
  680       690       700       710
```

- Caractéristiques protéiques
  - Poids moléculaire, point isoélectrique etc.

```
Nombre de résidus     =      285
Masse moléculaire (monoisotopique) =    30748.5234
Masse moléculaire (moyenne)        =    30768.4062
Index de polarité (%) =       38.25
Point isoélectrique   =        9.09
DO 260 (1mg/ml) =  1.881  DO 280 (1mg/ml) =  2.852
Composition atomique :
C =  1367  N =   400  O =   388  H =  1965  S =   16
```

Traduction conceptuelle

Sequence personnelle : insert cDNA CJD.265a du 28/10/2000 (Query)
Code génétique : Standard

```
ATGGCGAACCTTGGCTGCTGGATGCTGGTTCTCTTTGTGGCCACATGGAGTGACCTGGGC
        10         20         30         40         50         60
TACCGCTTGGAACCGACGACCTACGACCAAGAGAAACACCGGTGTACCTCACTGGACCCG

 M  A  N  L  G  C  W  M  L  V  L  F  V  A  T  W  S  D  L  G
  W  R  T  L  A  A  G  C  W  F  S  L  W  P  H  G  V  T  W  A
   G  E  P  W  L  L  D  A  G  S  L  C  G  H  M  E  .  P  G  P

 H  R  V  K  A  A  P  H  Q  N  E  K  H  G  C  P  T  V  Q  A
  P  S  G  Q  S  S  S  A  P  E  R  Q  P  W  M  S  H  G  P  G
   A  F  R  P  Q  Q  I  S  T  R  K  T  A  V  H  L  S  R  P  R


CTCTGCAAGAAGCGCCCGAAGCCTGGAGGATGGAACACTGGGGGCAGCCGATACCCGGGG
        70         80         90        100        110        120
GAGACGTTCTTCGCGGGCTTCGGACCTCCTACCTTGTGACCCCCGTCGGCTATGGGCCCC

 L  C  K  K  R  P  K  P  G  G  W  N  T  G  G  S  R  Y  P  G
  S  A  R  S  A  R  S  L  E  D  G  T  L  G  A  A  D  T  R  G
   L  Q  E  A  P  E  A  W  R  M  E  H  W  G  Q  P  I  P  G  A

 E  A  L  L  A  R  L  R  S  S  P  V  S  P  A  A  S  V  R  P
  R  C  S  A  G  S  A  Q  L  I  S  C  Q  P  C  G  I  G  P  A
   Q  L  F  R  G  F  G  P  P  H  F  V  P  P  L  R  Y  G  P  C


CAGGGCAGCCCTGGAGGCAACCGCTACCCACCTCAGGGCGGTGGTGGCTGGGGGCAGCCT
       130        140        150        160        170        180
GTCCCGTCGGGACCTCCGTTGGCGATGGGTGGAGTCCCGCCACCACCGACCCCCGTCGGA

 Q  G  S  P  G  G  N  R  Y  P  P  Q  G  G  G  G  W  G  Q  P
  R  A  A  L  E  A  T  A  T  H  L  R  A  V  V  A  G  G  S  L
   G  Q  P  W  R  Q  P  L  P  T  S  G  R  W  W  L  G  A  A  S

 L  A  A  R  S  A  V  A  V  W  R  L  A  T  T  A  P  P  L  R
  P  C  G  Q  L  C  G  S  G  V  E  P  R  H  H  S  P  A  A  E
   P  L  G  P  P  L  R  .  G  G  .  P  P  P  P  Q  P  C  G  .
```

Comparaison de séquence

- Objectif

Démontrer que 2 séquences sont homologues

Transfert mutuel de connaissances!

- Définition de l'homologie

2 séquences descendant par ***évolution divergente*** d'un ***ancètre commun*** sont **homologues**

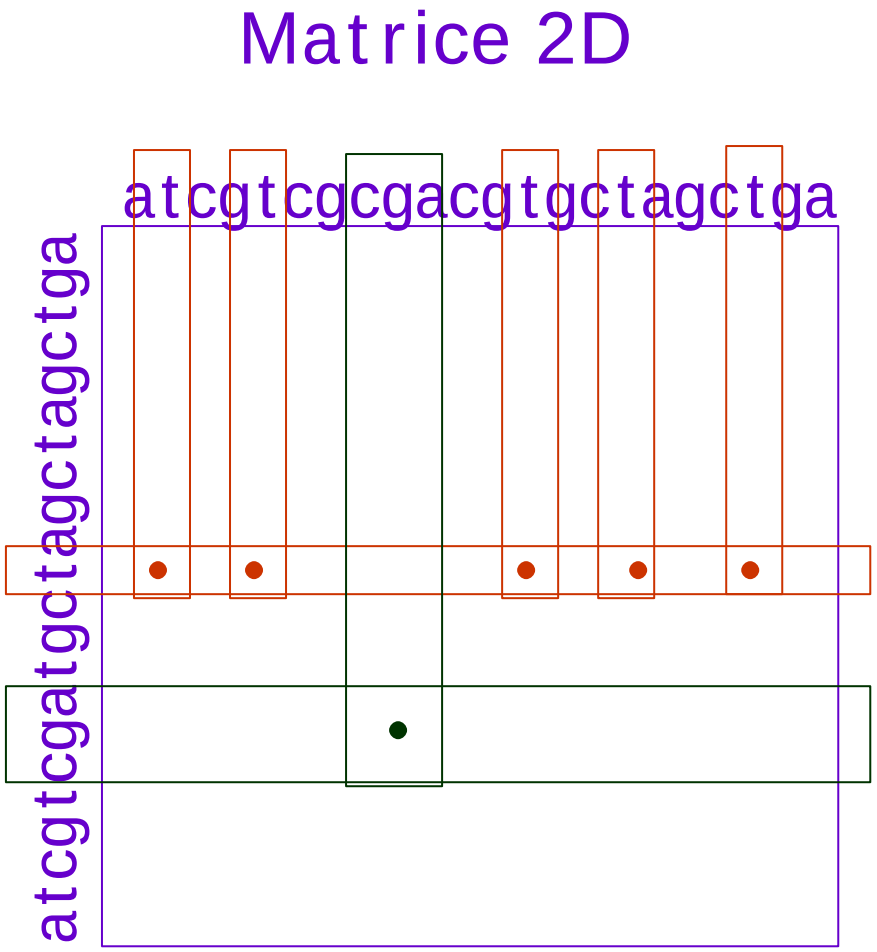
~~**Degré**~~ **Booléen!**

Orthologues: Divergence par ***Spéciation*** Organismes différents

Paralogues: Divergence par ***duplication*** de gène Même organisme

Comparaison de séquences

- Etude visuelle: nuage de point



- **LETASDEPIERRES LESTASDEFAITS**
- **ELUPARCETTECRAPULE**

Alignements de séquences

- Etats d'appariements

Apparié: **match** (**|** ou ***** ou **C**)
Non apparié: **mismatch** (néant)
Similaire**: Substitution** conservative (**+** ou **:** ou **.**)
Brèche: **gap** INDEL (**-** ou **.**)

- Représentations

```
4   ttttatacggatctccca----cctcgactgataattatgccattaaaggt 50
    ||||||||||||||| |||      | |   |||| ||||| || |||||| |
48  ttttatacggatccccctgttctcgcatttgatgattattccgttaaagat 76
```

identité INDE similarité

```
29   LTEWDVILKGPPDTLYEGGLFKAKIV--FPPKYPYEPPRLTFTSEMWHP 76
     +T+W+  + GPP + +E  ++   I    P YP  PP++TF S++  P
38   MTKWNGTILGPPHSNHENRIYSLSIDHTRCGPNYPDSPPKVTFISKINLP 84

 79  YSDGKLCISILHGDNAEEQGMTWSPAQKIDTVLLSVISLLNEPNPDSPAN 128
     | :| :|:|||      |:   | ||  |: ::::: :|||||| ::||:
 87  YPSGTVCLSIL.....EEDKDWRPAITINQLFIGIQELLNEPNIQEPAQ 130

YEAST   DEEYPNKPPHVKFLSEMFHPNVYANGEICLDILQNR--------WTP 111
ASFM2   PPKYPYEPPRLTFTSEMWHPNIYSDGKLCISILHGDNAEEQGMTWSP 116
ARATH   PPDYPFKPPKVAFRTKVFHPNINSNGSICLDILKEQ--------WSP 108
HUMAN   PAEYPFKPPKITFKTKIYHPNIDEKGQVCLPVISAEN-------WKP 110
          .** :**:: * ::::***:  .*.:*: ::           *.*
```

- Hypothèse: principe de parcimonie

→ *Alignement = historique évolution*

*Chemin (*mutations, insertions, délétions) le plus probable entre 2 séquences

Acides Aminés: critères de similarité

- Modélisation
  - Matrices de substitution
  - PAM, GONNET, BLOSUM…

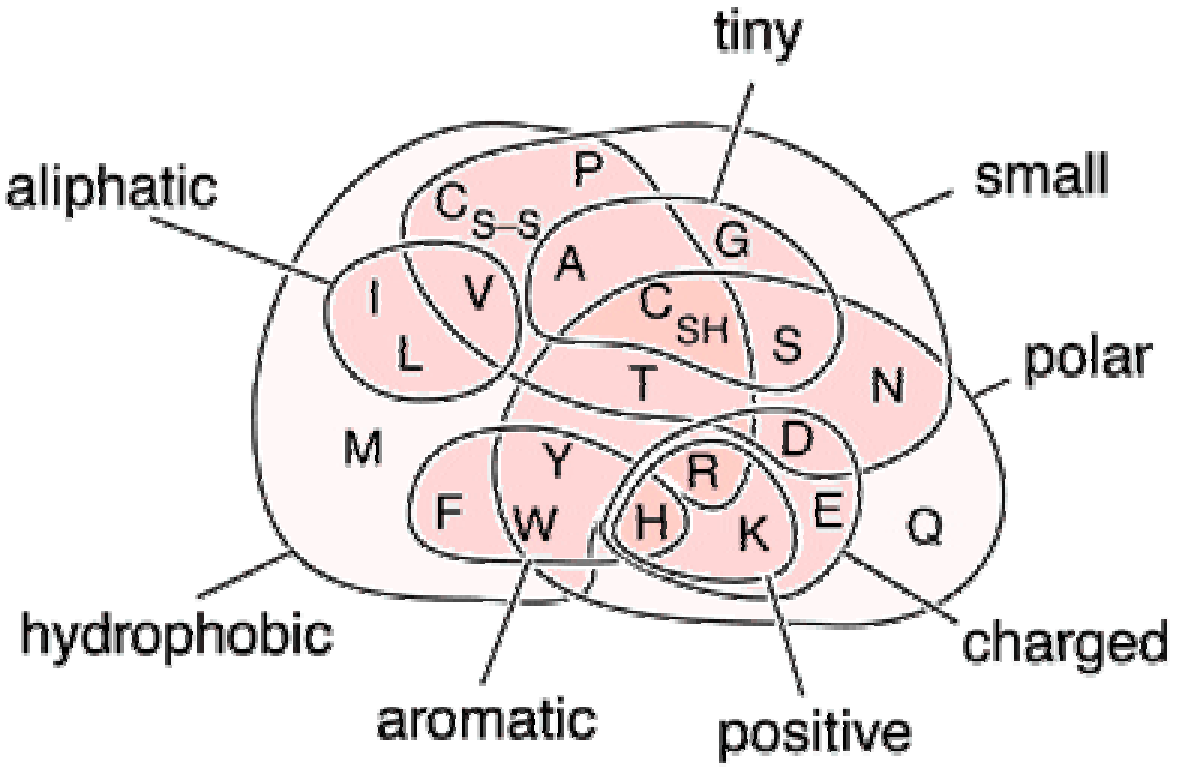


Diagramme de Venn

Alignements: critères de qualité

Bruits de fond:
Nucléique = ¼ = 25 %
Protéique = 1/20 = 5%

- % **identités**
- % **similarités**
- Somme des **scores** à chaque position
- Evaluation subjective:

→ *Alignement **non** dû au **hasard**: séquences **homologues***!

```
Identities = 36/61 (59%), Positives = 44/61 (72%)

 UBCv: 29  LTEWDVILKGPPDTLYEGGLFKAKIVFPPKYPYEPPRLTFTSEMWHPNIYSDGKLCISILH 90
           L  W+V + GPP+T YEGG FKA++ FP  YPY PP   F ++MWHPNIY  G +CISILH
UBC3h: 37  LYNWEVAIFGPPNTYYEGGYFKARLKFPIDYPYSPPAFRFLTKMWHPNIYETGDVCISILH 98


Identities = 16/47 (34%), Positives = 28/47 (59%)

  UBCv: 30 LTEWDVILKGPPDTLYEGGLFKAKIVFPPKYPYEPPRLTFTSEMWHP 76
           +T+W+  + GPP + +E  ++   I   P YP  PP++TF S++  P
UBCvar: 38 MTKWNGTILGPPHSNHENRIYSLSIDCGPNYPDSPPKVTFISKINLP 84


Identities = 19/56 (34%), Positives = 30/56 (54%)

UBCv: 3    SSFLLAEYKNLIVNPSEHFKISVNEDNLTEWDVILKGPPDTLFKAKIVFPPKYPYEP 64
           ++F+L   +    NPS+ FKI +   +    D ++K P  TLFK  +  P + P EP
 VTG: 1407 AAFMLGFSQKEQRNPSKQFKIILAVTSPNTIDTLIKAPKITLFKQAVQIPVQIPMEP 1463
```

Alignements automatisés

- Alignement manuel

  Intuition du biologiste

- Alignement bioinformatique

  Etudie tous alignements possibles

  Calcule score pour chaque cas

  Score le + élevé = alignement optimal

- Calcul du score

  2 résidus: matrice de substitution

  INDELS: doivent être pénalisés

  Gap de longueur L, pénalité G:

  G = - L x D pénalité linéaire

  G = - D + (L-1) E pénalité affine

moins de gaps mais plus **longs** = réalité biologique (1 **évenement** pour tout le INDEL)

- Gaps peu pénalisés

```
MANIAVQRIKR--EFKEVLKSEETSKNQIKVDLV-D---EN---FTELRGEIAGPPDTPY 51
MS-----SSKRRIET-DVMK-LLMSDHQ--VDLINDSMQEFHVKF--L-----GPKDTPY 44
*:       **  *  :*:*    *.:*  ***: *   *    *  *     ** ****

EGG--RYQL--EIKIPETYPFNPPKVRFITKIWHPNI---SSVTGAICLD-I------LK 97
ENGVWR--LHVEL--PDNYPYKSPSIGFVNKIFHPNIDIAS---GSICLDVINSTWSPLY 97
*.*  *  *  *:  *:.**::.*.: *:.**:****   *   *:**** *       *

DQWAAAMTLR-TVLLSLQALLAAAEPD--DPQDAVVANQYKQNPE-MFKQTAR---LWAH 150
D-------LINIVEWMIPGLLK--EPNGSDP-----LN----N-EAATLQ-LRDKKL--- 134
*       *   *   : .**   **:  **      *    * *    *  *   *

VYAGAPVSSPEY-----TKKIENLCAM-GFDRNAVIVALS--------SKSWDVE--TAT 194
-YEEKIK---EYIDKYATK--EKYQQMFGGD-NDSDDSDSGGDLQEEDSDS-D-EDMDGT 185
 *       **     **  *:   * * * *    : *        *.* * *   .*

----------EL--LLS-------------N- 200
GVSSGDDSVDELSEDLSDIDVSDDDDYDEVANQ 218
          **   **              *
```

- Gap fortement pénalisés

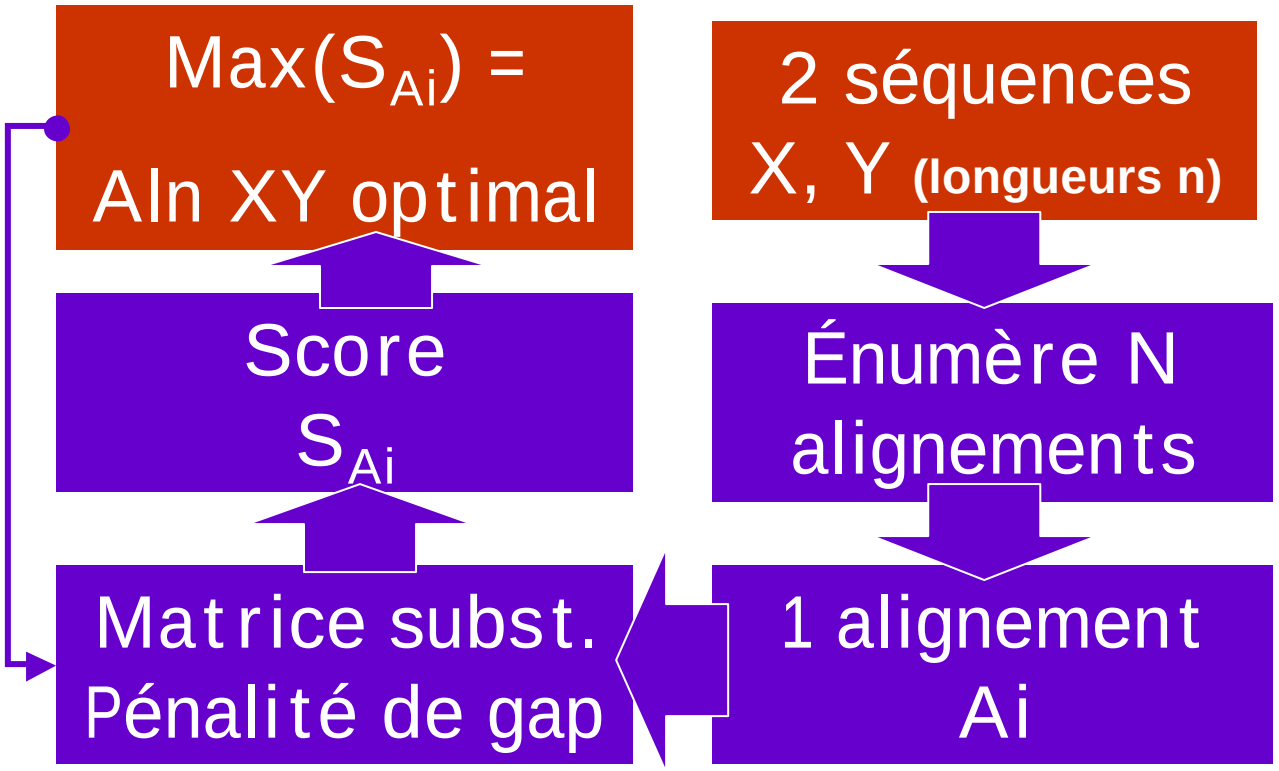
```
MANIAVQRIKREFKEVLKSEETSKNQIKVDLVDENFTELRGEIAGPPDTPYEGGRYQLEI 60
MS-SSKRRIETDVMKLLMS------DHQVDLINDSMQEFHVKFLGPKDTPYENGVWRLHV 53
*:  : :**: :. ::* *      : :***::.: *:: :: ** *****.* ::*.:

KIPETYPFNPPKVRFITKIWHPNISSVTGAICLDILKDQWAAAMTLR-TVLLSLQALLAA 119
ELPDNYPYKSPSIGFVNKIFHPNIDIASGSICLDVINSTWSPLYDLINIVEWMIPGLLKE 113
::*:.**::.*.: *:.**:****. .:*:****:::. *:.   *   *   : .**

AEPDDPQDAVVAN----QYKQNPEMFKQTARLWAHVY------AGAPVSSPEYTKKIENL 169
PNGSDPLNNEAATLQLRDKKLYEEKIKEYIDKYATKEKYQQMFGGDNDSDDSDSGGDLQE 173
.: .** :  .*.    : *   * :*:    :*          .*   *. . :    :

CAMGFDRNAVIVALSSKSWDVETATELLLSN------------- 200
EDSDSDEDMDGTGVSSGDDSVDELSEDLSDIDVSDDDDYDEVANQ 218
   . *.:   ..:** . .*:  :* * .
```

Recherche alignement optimal



$$N = \frac{(2n)!}{(n!)^2} = O(2^n)$$

= *calcul « NP complet »*

longueur	temps
10	1 sec
20	17 heures
50	35000 ans

$N=O(2^n)$

1 aln / msec

Programmation dynamique

- Approche « *divide & conquer* »

  Algo: *Needleman & Wunsh*

grand problème insoluble

↓

petits problèmes individuellement solubles

↓

Sol^{tion} G^{ale} optimale = $\mathcal{F}$ (sol^{tion} locales

Prérequis = principe de l'optimalité

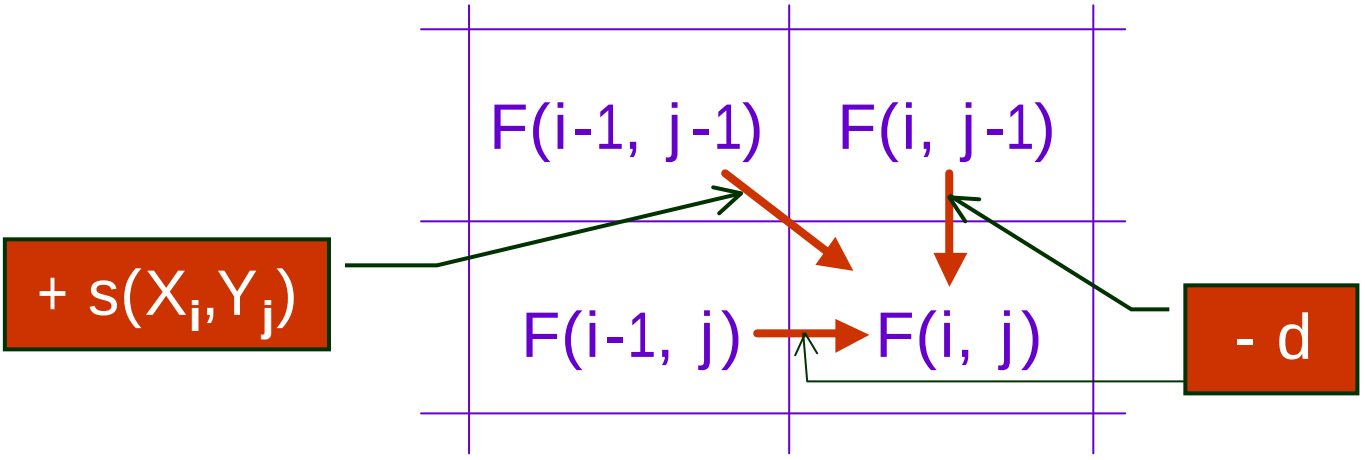
- Prog. Dyn. = $O(n^2)$

longueur	temps
10	0.1 sec
20	2.5 sec
1000	16 min

Programmation dynamique

		H	E	A	G	A	W	G	H	E	E
P											
A											
W											
H						F(i, j)					
E											
A											
E											

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



Exemple de Prog. Dyn.

- **Calcul des scores:**
  - **Match = + 2**
  - **Mismatch = - 1**
  - **Gap = - 2**

Gap < mismatch!

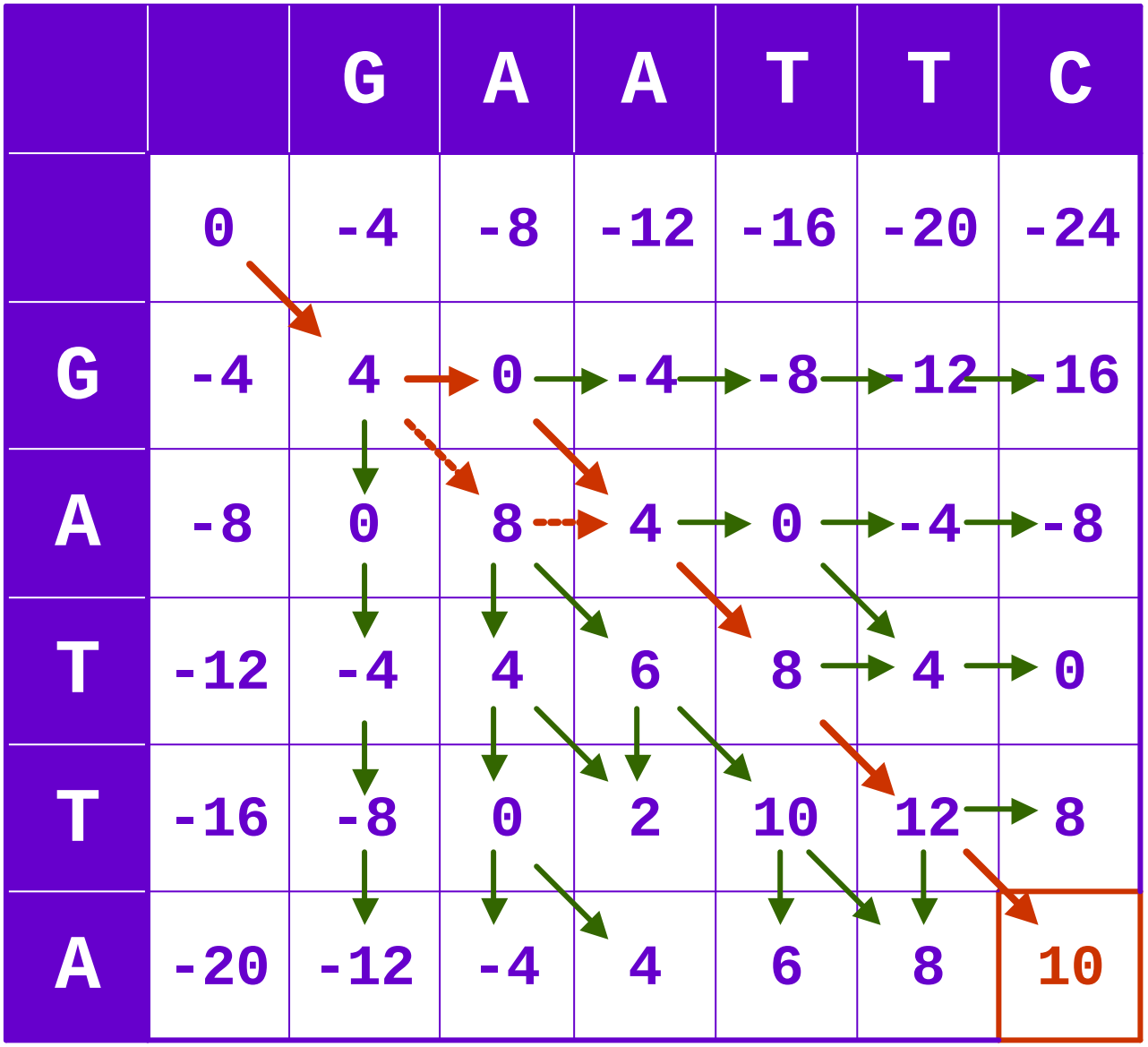
		G	A	A	T	T	C
	0	-2	-4	-6	-8	-10	-12
G	-2	2	0	-2	-4	-6	-8
A	-4	0	4	2	0	-2	-4
T	-6	-2	2	3	4	2	0
T	-8	-4	0	1	5	6	4
A	-10	-6	-2	2	3	4	5

GAATTC GAATTC
G-ATTA GA-TTA

Scores relatifs

- **Calcul des scores:**
  - **Match = +4**
  - **Mismatch = - 2**
  - **Gap = - 4**

valeurs relatives importent!

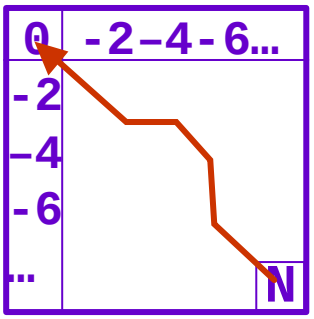


Variantes Prog. Dyn.

- **Alignement global**

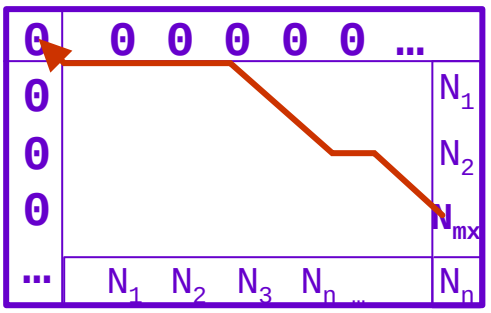
  Needleman & Wunsh

  Alns de 2 séquences sur toute leurs longueurs



  Variante: gaps extrêmités non-pénalisants

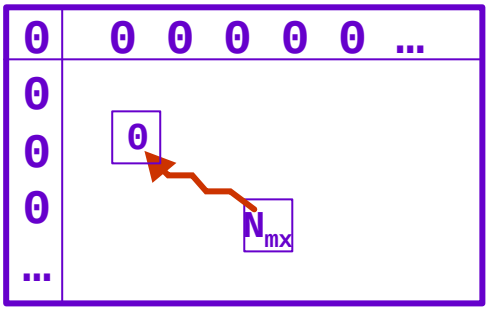
  Alns chevauchants, ou séquences de longueurs très diff.



- **Alignements locaux**

  Smith & Waterman

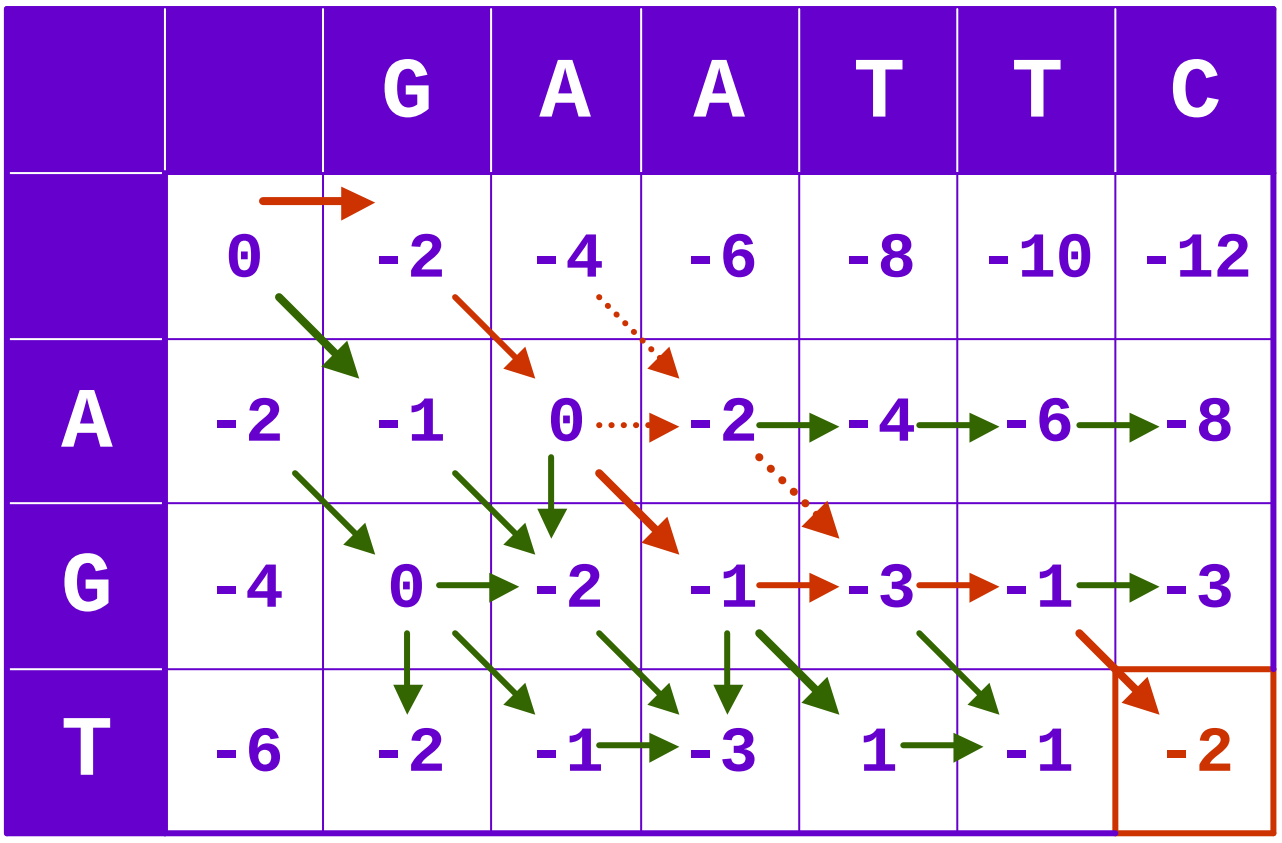
  Alns de sous régions (ex: domaines protéiques)



Gaps aux extrêmités

- **Calcul des scores:**
  - **Match = +2**
  - **Mismatch = - 1**
  - **Gap = - 2**

Longueurs très différentes



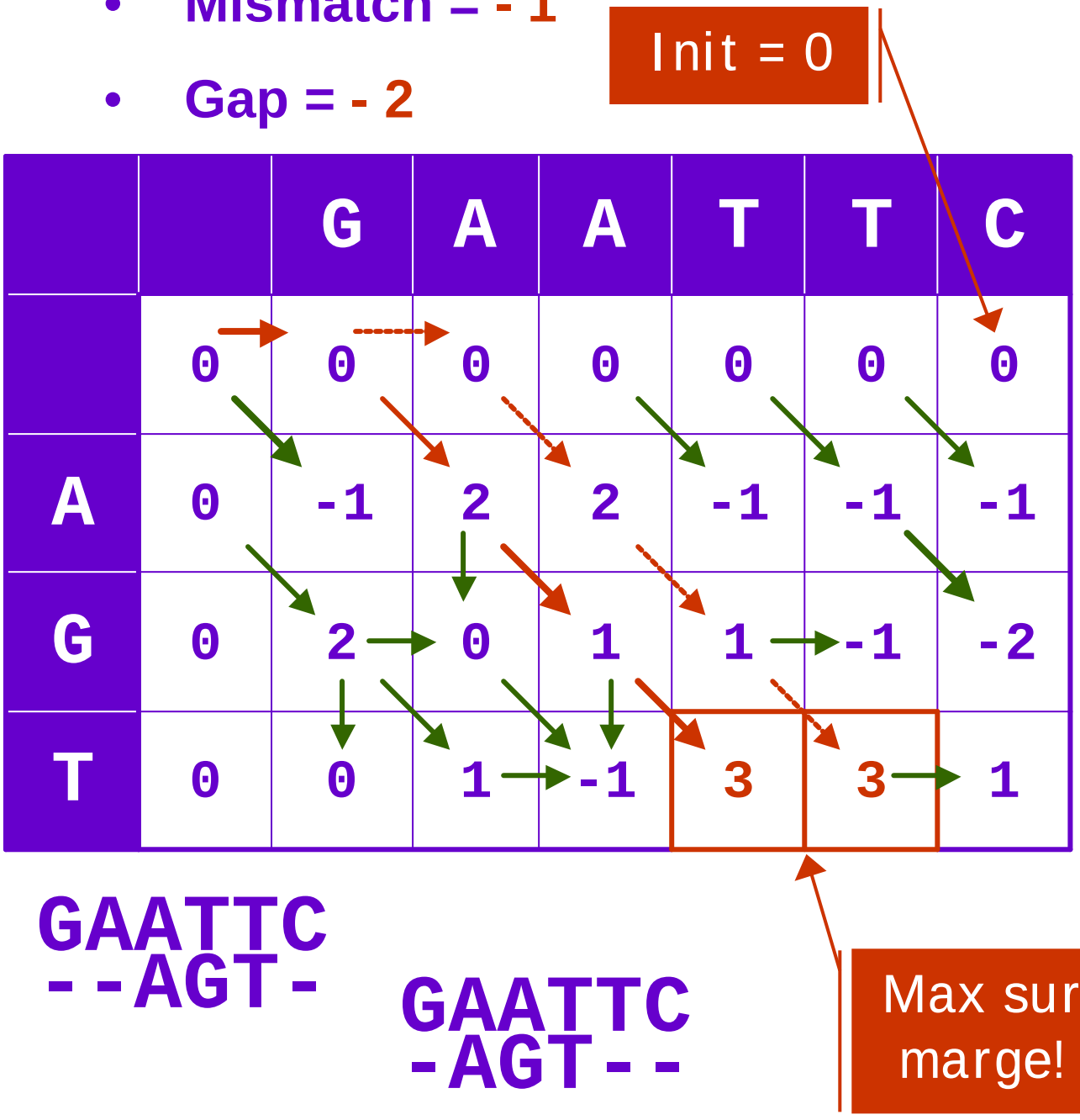
GAATTC
-AG--T

GAATTC
-A-G-T

GAATTC
--AG-T

Gaps aux extrêmités

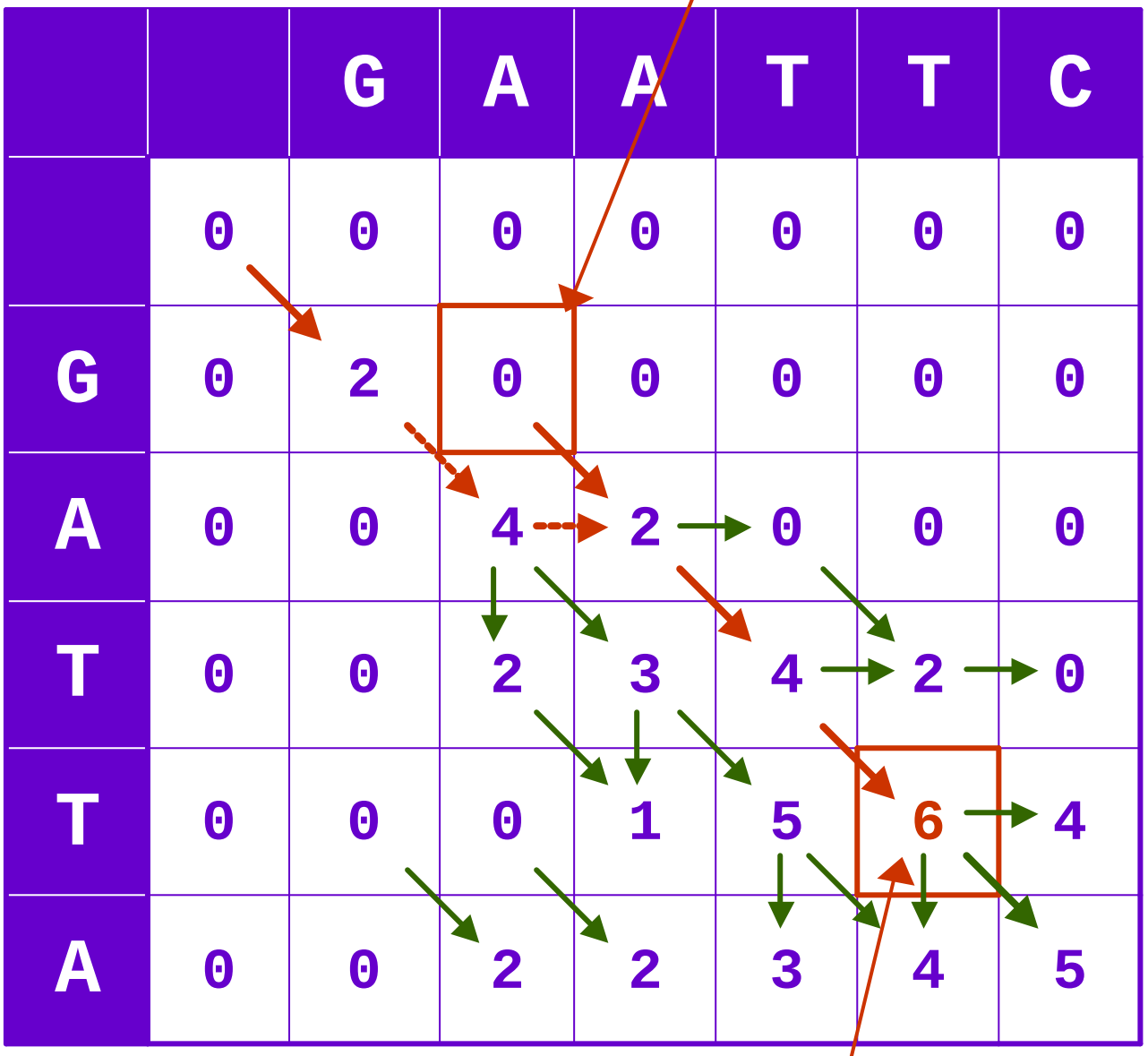
- **Calcul des scores:**
  - **Match = +2**
  - **Mismatch = - 1**
  - **Gap = - 2**



Alns locaux

- **Calcul des scores:**
  - **Match = + 2**
  - **Mismatch = - 1**
  - **Gap = - 2**

Négatifs = zéro = début



Max!

ATT GAATT
ATT GA-TT