INSTITUT PASTEUR

Alignements multiples, recherche et découverte de motifs

I. Alignements multiples

II. Définition de motifs

III. Recherche de motifs

IV. Autre utilisation

Alignements multiples

- Caractériser des familles de protéines. Trouver des séquences similaires à une famille.

```
chite      ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat      --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
thybr      KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
unknown    -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
                 ***. ::: .: ..  .    :  . .      *  .  *: *

chite      AATAKQNYIRALQEYERNGG-
wheat      ANKLKGEYNKAIAAYNKGESA
trybr      AEKDKERYRKEM---------
Unknown    AKDDRIRYDNEMKSWEEQMAE
           *   : .* . :
```

Alignements multiples

- Identifier des régions et/ou des résidus conservées entre des séquences.
  - ↪ Définir des motifs et des domaines.

```
chite      ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat      --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
thybr      KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse      -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
                 ***. ::: .: ..  .    :  . .      *  .  *: *

chite      AATAKQNYIRALQEYERNGG-
wheat      ANKLKGEYNKAIAAYNKGESA
trybr      AEKDKERYRKEM---------
mouse      AKDDRIRYDNEMKSWEEQMAE
           *   : .* . :
```

Alignements multiples

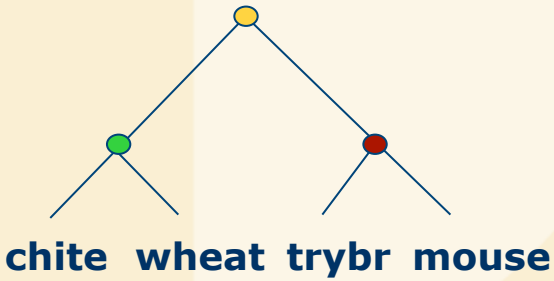
- Déterminer la **séquence consensus** de plusieurs séquences alignées.

```
chite      ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat      --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
thybr      KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse      -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP

cons         DSNKPKRAmSAFMfFsSEFREsiKrENpDlSsVvEVaKAAGEAWKnLSP

chite      AATAKQNYIRALQEYERNGG-
wheat      ANKLKGEYNKAIAAYNKGESA
trybr      AEKDKERYRKEM---------
mouse      AKDDRIRYDNEMKSWEEQMAE

cons       AeKDKeRYeKEM
```

Alignements multiples

- **Constituer le point de départ d'un traitement phylogénique.**

```
chite      ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat      --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
thybr      KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse      -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
                 ***. ::: .: ..  .    :  . .      *  .  *: *

chite      AATAKQNYIRALQEYERNGG-
wheat      ANKLKGEYNKAIAAYNKGESA
trybr      AEKDKERYRKEM---------
mouse      AKDDRIRYDNEMKSWEEQMAE
           *   : .* . :
```



INSTITUT PASTEUR

Alignements multiples

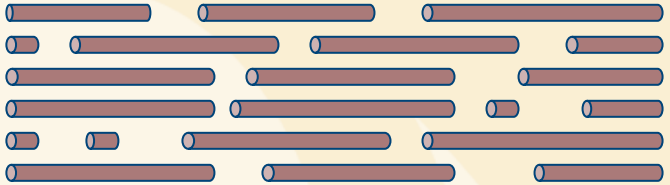
- **Les programmes d'alignement multiple ne retournent qu'un seul résultat et les résultats ne sont pas toujours parfaits.**

```
chite      ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat      --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
thybr      KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse      -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
                 ***. ::: .: ..  .    :  . .      *  .  *: *

chite      AATAKQNYIRALQEYERNGG-
wheat      ANKLKGEYNKAIAAYNKGESA
trybr      AEKDKERYRKEM---------
mouse      AKDDRIRYDNEMKSWEEQMAE
           *   : .* . :
```

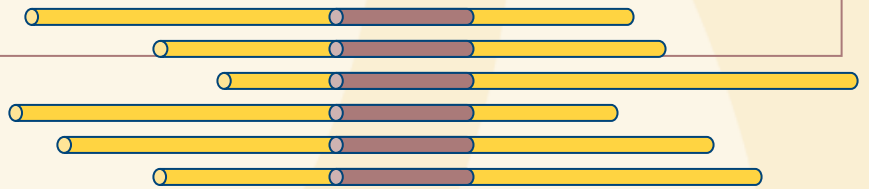
Alignements multiples

Alignement multiple global:
« tableau » contenant les k séquences, avec des indels

Séquences non alignées de longueur différentes



Alignement multiple local:
« tableau » contenant k portions de séquences, avec des indels



Motif, pattern, profil

INSTITUT PASTEUR

Alignements multiples, recherche et découverte de motifs

I. Alignements multiples
   1. Alignement multiple optimal
   2. Méthodes heuristiques
II. Définition de motifs
III. Recherche de motifs
IV. Autre utilisation

Construction d'un alignement multiple

- **Il faut quantifier l'alignement multiple:** ↪**Système de *SCORE***
  - ✓ adaptable à un nombre quelconque de séquences
  - ✓ indépendant de l'ordre des séquences
  - ✓ qui reflète la similitude

- **N séquences:**

$$\begin{aligned} S_1&: s1_1\ s1_2\ \ldots\ S1_{n1} \\ S_2&: s2_1\ s2_2\ \ldots\ S2_{n2} \\ &\ldots \\ S_N&: sN_1\ sN_2\ \ldots\ SN_{nN} \end{aligned}$$

- **Alignement des N séquences:**

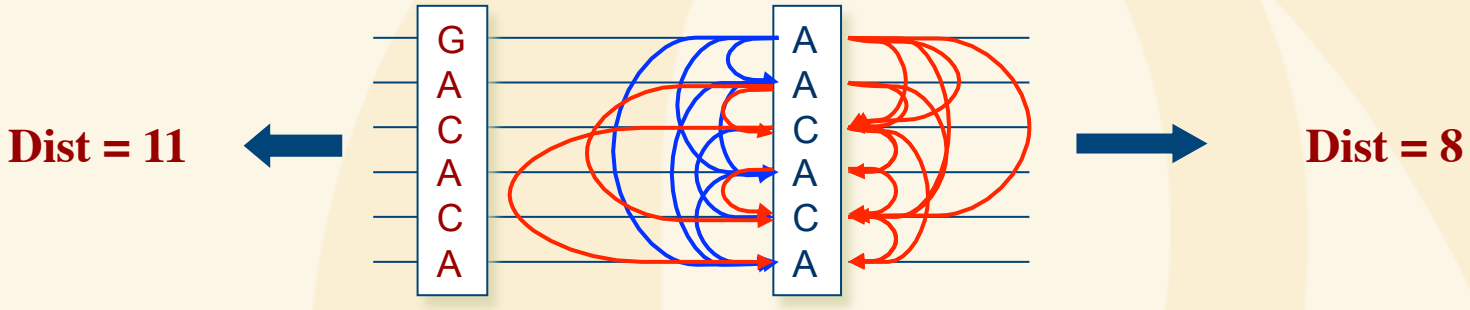
$$\begin{bmatrix} S1'_1 \\ S2'_1 \\ \ldots \\ sN'_1 \end{bmatrix} \begin{bmatrix} s1'_2 \\ s2'_2 \\ \ldots \\ sN'_2 \end{bmatrix} \ldots \begin{bmatrix} S1'_{n1} \\ S2'_{n2} \\ \ldots \\ SN'_{nN} \end{bmatrix}$$

- **Score de l'alignement:**
  (positions indépendantes)

$$Sc\begin{bmatrix} S1'_1 \\ S2'_1 \\ \ldots \\ sN'_1 \end{bmatrix} + Sc\begin{bmatrix} s1'_2 \\ s2'_2 \\ \ldots \\ sN'_2 \end{bmatrix} + \ldots + Sc\begin{bmatrix} S1'_{n1} \\ S2'_{n2} \\ \ldots \\ SN'_{nN} \end{bmatrix}$$

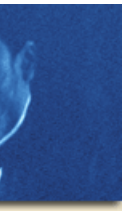
Score : Somme of Pair(SP)

- Le score est la somme des scores de tous les couples non ordonnés.
- Tous les couples sont indépendants.
- Les colonnes sont indépendantes.

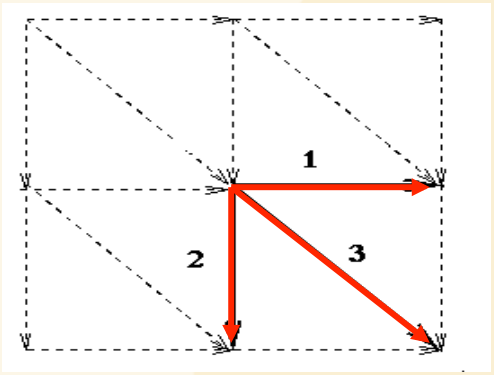


Identité = 0 et substitution = 1

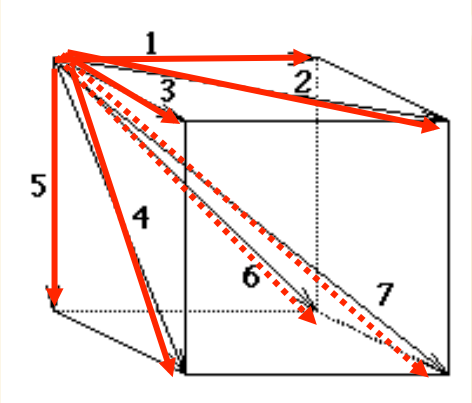
Alignements multiples simultanés



- **Alignements multiples simultanés de toutes les séquences par programmation dynamique.**



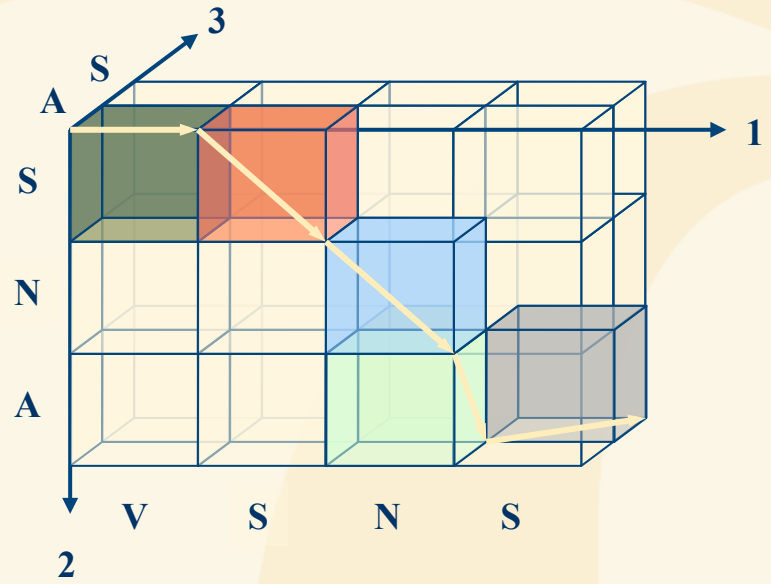
3 chemins



7 chemins

- **2^k - 1 possibilités de résoudre l'alignement multiple**

Alignements multiples simultanés: score SP

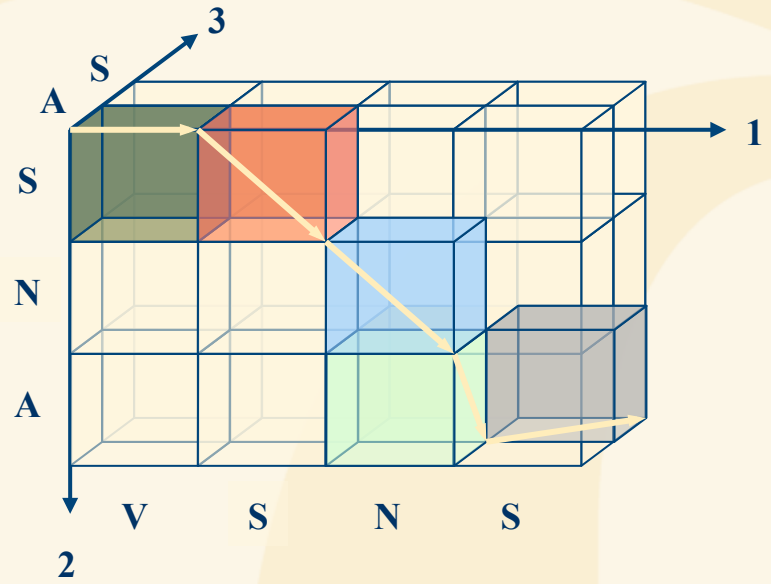


1. V S N – S
2. – S N A –
3. – – – A S

- **Complexité en temps**: $O(n^k)$
- **Exemple**: globin (150 aa)
  - 2 globins: 1s
  - 3 globins: 2 mn 30
  - 4 globins: 6 h 15

5 globins: 39 jours
6 globins: 16 ans
7 globins: 2400 ans ….

Alignements multiples simultanés: score SP



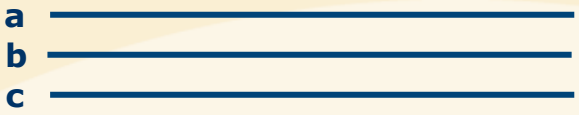
1. V S N – S
2. – S N A –
3. – – – A S

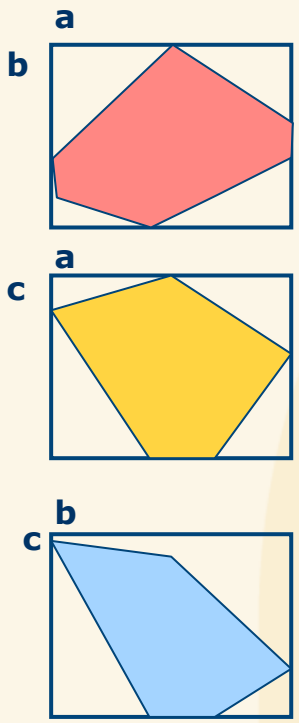
- **Complexité en temps**: $O(n^k)$ => Nécessité de programmes heuristiques

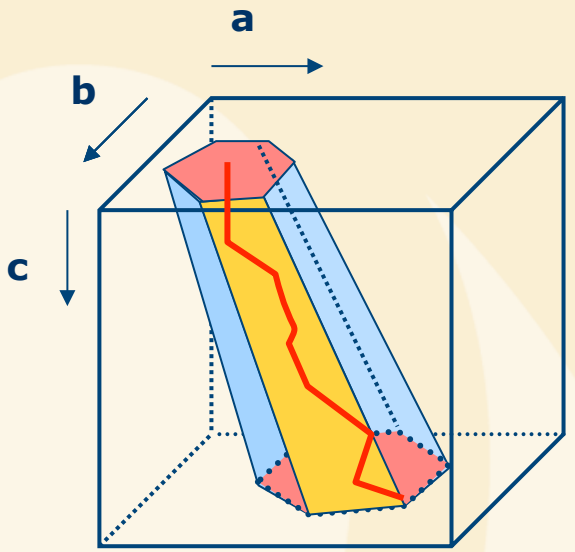
- **Amélioration de l'algorithme**:

  - Carillo-Lipman bound: MSA (1989)
  - Divide and Conquer: DCA (1997)

Alignements multiples simultanés: MSA

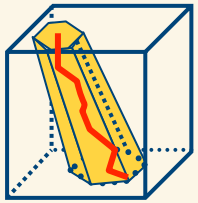


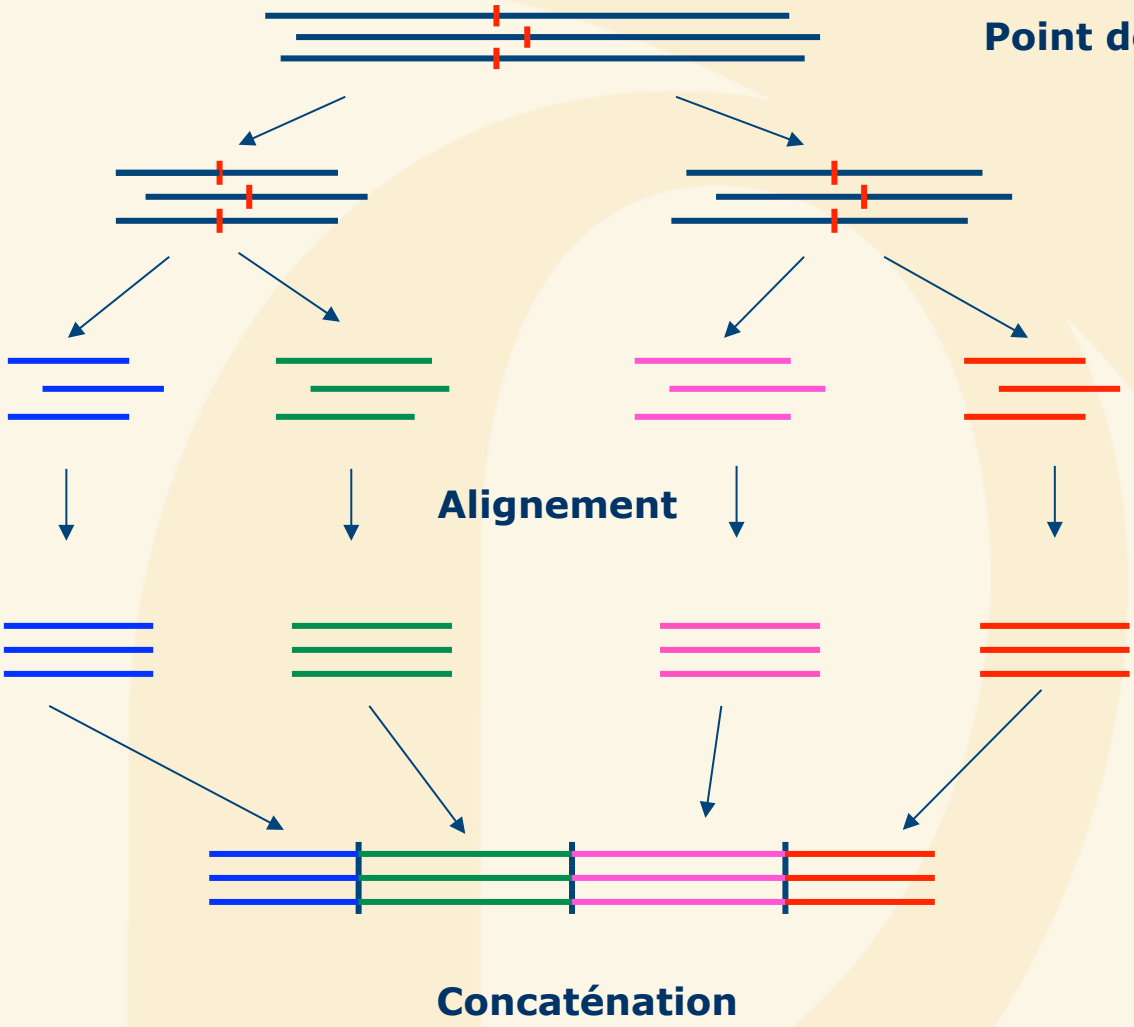




A_{pq} : Alignement induit de abc

Alignements multiples simultanés: DCA





INSTITUT PASTEUR

INSTITUT PASTEUR

Alignements multiples, recherche et découverte de motifs

I. **Alignements multiples**
   1. Alignement multiple optimal
   2. Méthodes heuristiques
II. Définition de motifs
III. Recherche de motifs
IV. Autre utilisation

Alignement multiple progressif: principe

- L'approche la plus courante consiste à aligner progressivement des paires de séquences.

- Les différentes approches se distinguent par:

  - la façon de choisir la paire initiale de séquences

  - la progression dans l'alignement:

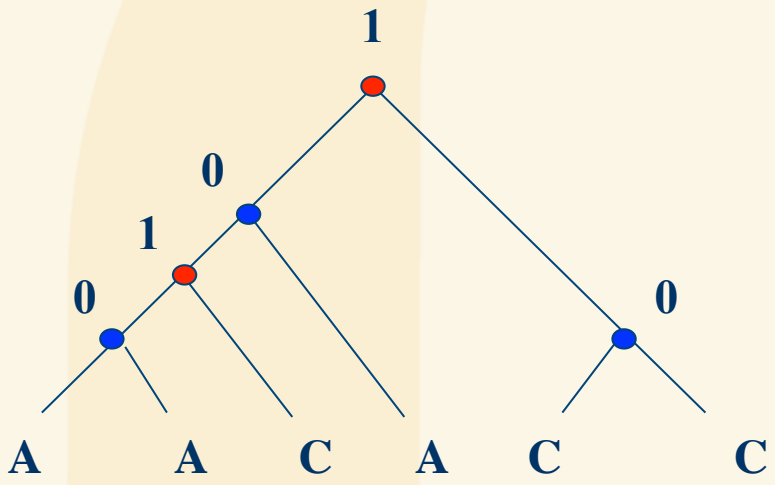
    - Soit aligner chaque séquence les unes après les autres à un alignement unique enrichi à chaque étape.
    - Soit créer des sous-familles de séquences d'abord alignées au sein de ces familles puis entre les familles.

  - la méthode de pondération des alignements individuels des paires de séquences et des alignements cumulés

Score: Tree Alignment (TA)

- Les séquences sont **SUPOSEES** liées par une **histoire évolutive**.
- Elles sont représentées suivant un arbre.

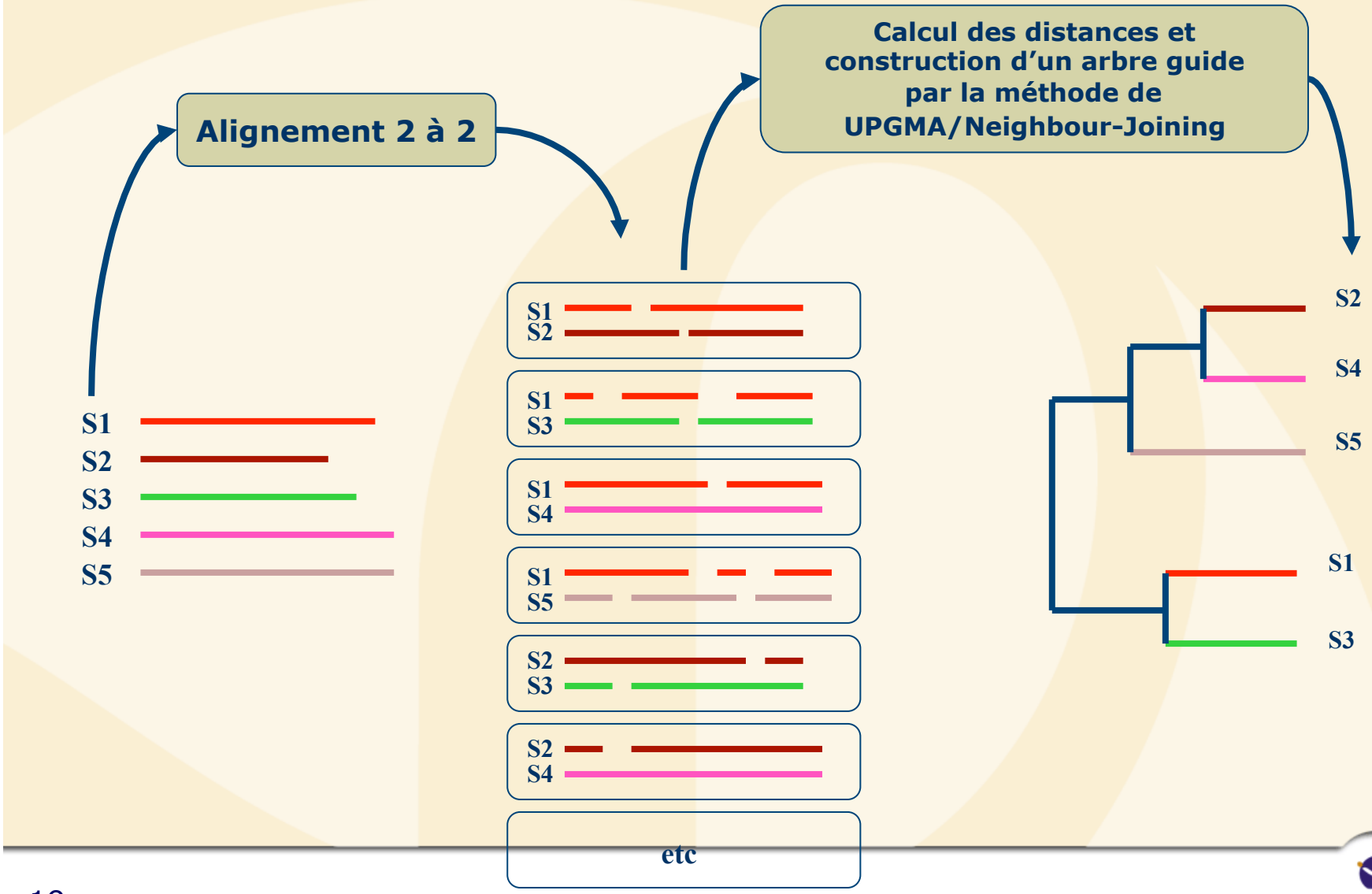
- Le score est la somme des scores des couples liés par une branche de l'arbre.



Dist = 2

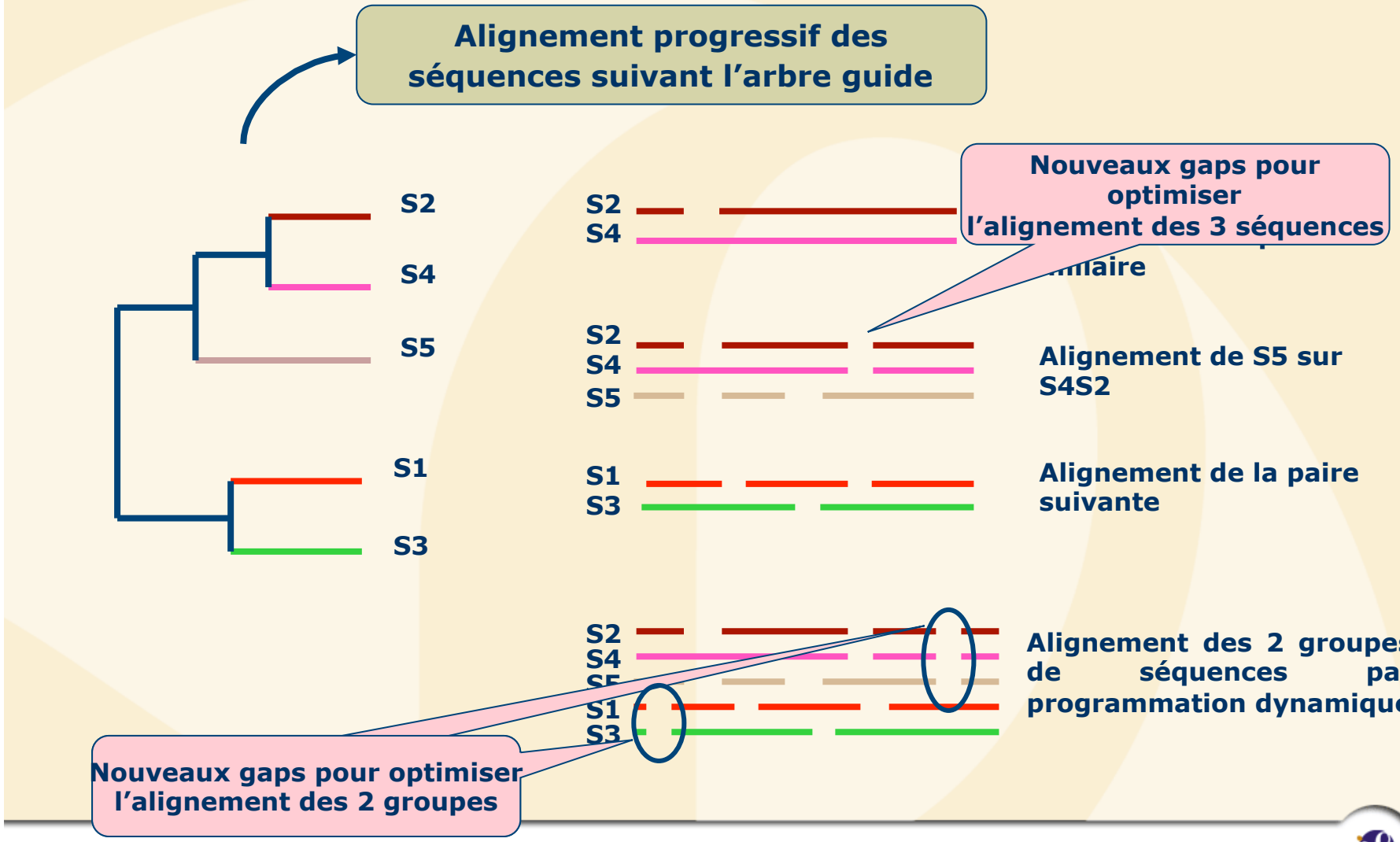
INSTITUT PASTEUR

Alignements progressifs: Clustalw



INSTITUT PASTEUR

Alignements progressifs: Clustalw



INSTITUT PASTEUR

Alignements progressifs: Clustalw

```
S1:  PEEKSAVTALWGKVNVDEVGG
S2:  AADKTNVKAAWSKVGGHAGEYGA
S3:  GEEKAAVLALWDKVNEEEVGG
S4:  PADKTNVKAAWGKVGAHAGEYGA
S5:  EHEWQLVLHVWAKVEADVAGHGQ
```

- **Système de score:** identité = 1 , substitution = -1 , insertion/délétion = -2

- **Etape 1**: calcul des scores de similitude de tous les alignements globaux 2 à 2.

```
S1: PEEKSAVTALWGKVNVD--EVGG
    |  | | | ||     | |
S2: AADKTNVKAAWSKVGGHAGEYGA
```

```
S2: AADKTNVKAAWSKVGGHAGEYGA
    |  | | | ||     | |
S3: GEEKAAVLALWDKVNEE--EVGG
```

```
S3: GEEKAAVLALWDKVNEE--EVGG
      |   ||  | ||       |
S5: EHEWQLVLHVWAKVEADVAGHGQ
```

```
S1: PEEKSAVTALWGKVNVDEVGG
    ||| || | | |||  ||||
S3: GEEKAAVLALWDKVNEEEVGG
```

```
S2: AADKTNVKAAWSKVGGHAGEYGA
    |||||||||| ||| |||||||
S4: PADKTNVKAAWGKVGAHAGEYGA
```

```
S4: PADKTNVKAAWGKVGAH-AGEYGA
          |   | || |  ||  |
S5: EHEWQLVLHVWAKVEADVAGH-GQ
```

```
S1: PEEKSAVTALWGKVNVD--EVGG
    |  |  | | ||||     | |
S4: PADKTNVKAAWGKVGAHAGEYGA
```

```
S2: AADKTNVKAAWSKVGGHAGEYGA
      |   | ||       |
S5: EHEWQLVLHVWAKVEADVAGHGQ
```

```
S1: PEEKSAVTALWGKVNVD--EVGG
      |   |  | ||  |    |
S5: EHEWQLVLHVWAKVEADVAGHGQ
```

```
S3: GEEKAAVLALWDKVNEE--EVGG
    |  | | | ||     | |
S4: PADKTNVKAAWGKVGAHAGEYGA
```

Alignements progressifs: Clustalw

- **Etape 1** (suite): Tableau des scores des alignements

	s1	s2	s3	s4	s5
s1		-9	7	-5	-11
s2	-9		-9	17	-13
s3	7	-9		-9	-11
s4	-5	17	-9		-10
s5	-11	-13	-11	-10	

· ***Regroupe les 2 séquences qui ont les meilleurs scores entre elles: s2s4***

- **Etape 2** : Calcul des nouveaux scores avec les autres séquences

	s1	s3	s2s4	s5
s1		7	-7	-11
s2s4	-7	-9		-11.5
s3	7		-9	-11
s5	-11	-11	-11.5	

. Moyenne des scores des séquences regroupées:
Ex: s2s4 avec s1= (s1s2+s1s4)/2 (-9-5)/2=-7

. Regroupe les 2 séquences qui ont les meilleurs scores entre elles: s1s3

INSTITUT PASTEUR

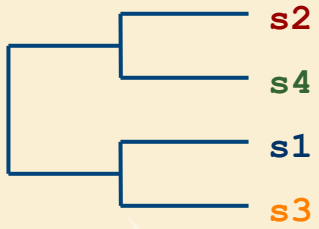
Alignements progressifs: Clustalw

- **Etape 2-2**: Calcul des nouveaux scores avec les autres séquences

	s1s3	s2s4	s5
s2s4	-8		-11.5
s1s3		-8	-11
s5	-11	-11.5	

. Moyenne des scores des séquences regroupées:
Ex: s1s3 avec les autres

. Regroupe les 2 séquences qui ont les meilleurs scores entre elles: s1s3/s2s4



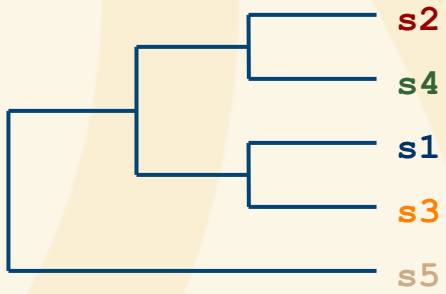
- **Etape 2-3** : Calcul des nouveaux scores avec les autres séquences

	s1s3/s2s4	s5
s1s3/ s2s4		-11.25
s5	-11.25	

. Moyenne des scores des séquences regroupées:

Ex: s1s3s2s4 avec les autres

. Regroupe les 2 dernières séquences



INSTITUT PASTEUR

Alignements progressifs: Clustalw

- **Etape 3 : Alignement progressif des séquences incorporées suivant l'ordre de l'arbre guide**

1er groupe

```
S2: AADKTNVKAAWSKVGGHAGEYGA
    |||||||||| ||| |||||||
S4: PADKTNVKAAWGKVGAHAGEYGA
```

2eme groupe

```
S1: PEEKSAVTALWGKVNVDEVGG
    ||| || | | |||  ||||
S3: GEEKAAVLALWDKVNEEEVGG
```

3eme groupe

```
S2: AADKTNVKAAWSKVGGHAGEYGA
S4: PADKTNVKAAWGKVGAHAGEYGA
S1: PEEKSAVTALWGKVNVD--EVGG
S3: GEEKAAVLALWDKVNEE--EVGG
```

4eme groupe

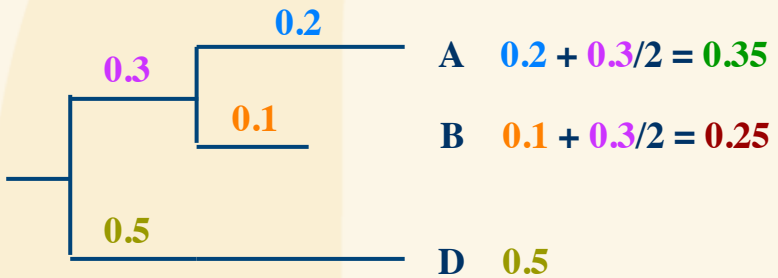
```
S2: AADKTNVKAAWSKVGGH-AGEYGA
S4: PADKTNVKAAWGKVGAH-AGEYGA
S1: PEEKSAVTALWGKVNVD---EVGG
S3: GEEKAAVLALWDKVNEE---EVGG
S5: EHEWQLVLHVWAKVEADVAGH-GQ
```

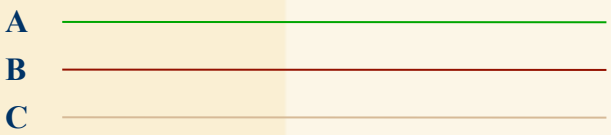
Alignements progressifs: Clustalw

• **Pondération** des séquences en fonction de leur distance par rapport à la racine de l'arbre (midd-point): Attribuer un poids à chaque branche de l'arbre.

• poids d'une branche = longueur branche / nombre taxa sur la branche.

• poids d'une séquence = somme des branches pondérées de la racine jusqu'au taxon considéré.





INSTITUT PASTEUR

Alignements progressifs: Clustalw

- **Erreurs dans les alignements des sous-groupes initiaux se propagent dans tous l'alignement et ne sont jamais corrigés.**

- **Adaptation des matrices de similitude au fil de l'algorithme en fonction de la divergence des séquences à aligner ➔ On ne choisit pas sa matrice !!!**

  Blosum 80 pour aligner les séquences proches
  Blosum 50 pour aligner les séquences distantes

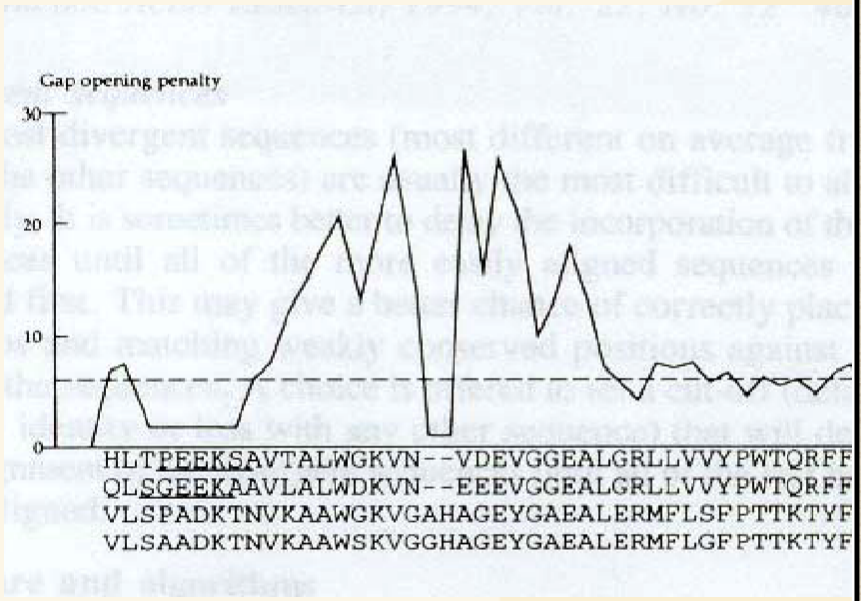
- **Pénalité de gaps réduite dans les régions hydrophobes.**

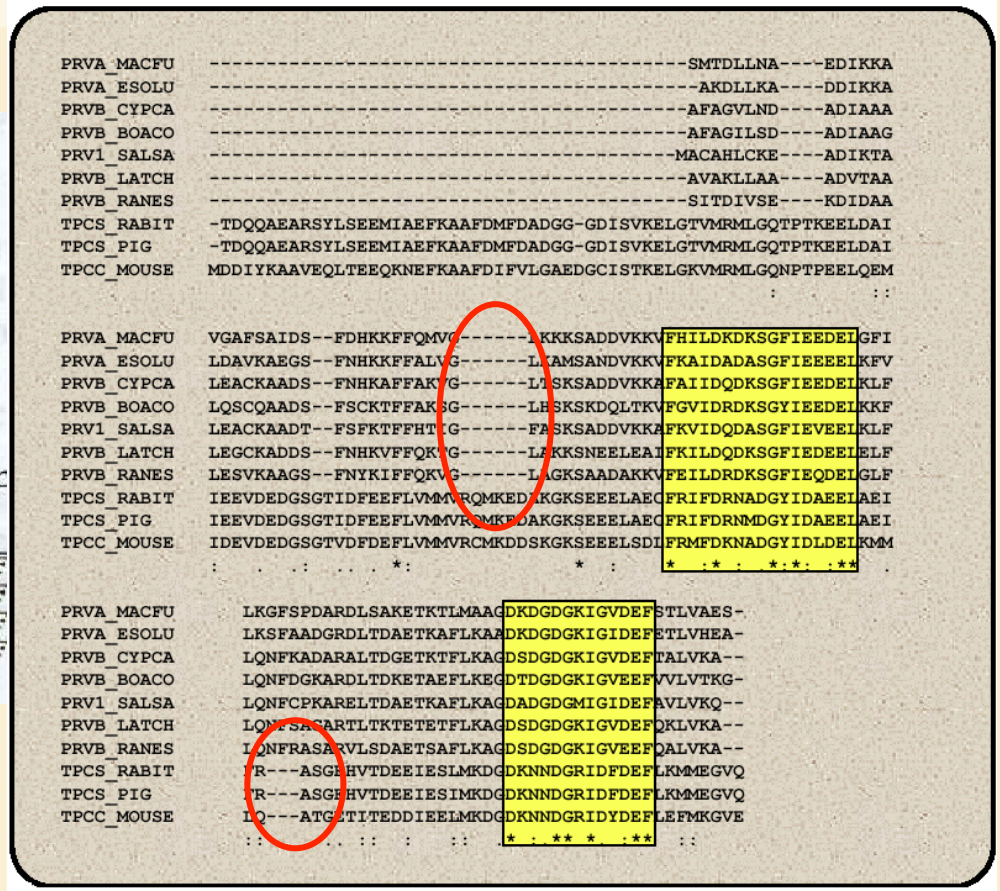
- **Encourage la formation de gaps dans les boucles plutôt que dans les régions structurées.**

- **Pénalité de gaps diminuée dans le voisinage d'autre gaps.**

- **Evite la formation de petits gaps voisins, au profit de longs gaps.**

Alignements progressifs: Clustalw



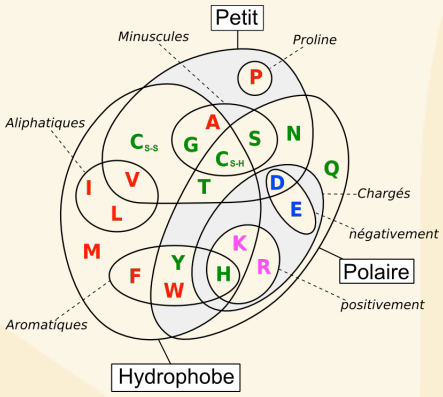


INSTITUT PASTEUR

Alignements progressifs: Clustalw

```
chite      ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD 47
mouse      -----KPKRPRSAYNIYVSESFQEAKDDSAQGK-LKLVN----EAWKNLSP 41
wheat      --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE 49
thybr      KKDSNAPKRAMTSFMFFS----SDFRSKHSDLS-IVEMSKAAGAAWKELGP 46
                 ***. ::: .:     .. : . .. . :  :        *: *
```

* -- caractère identique dans toute la colonne
: -- substitutions conservatives (suivant la table des couleurs)
. -- substitutions semi conservatives



AVFPMILW	Petits + hydrophobes (incl. aromatiques sans Y)
DE	Acides
RK	Basiques - H
STYHCNGQ	Hydroxyl + sulfhydryl + amine + G
Autres	Unusual amino/imino acids . . .

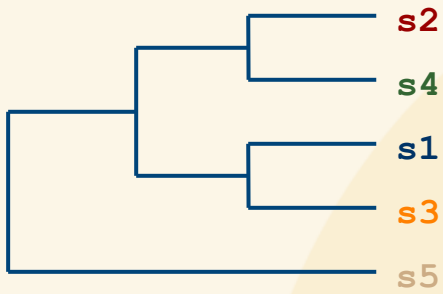
INSTITUT PASTEUR

Alignements progressifs: MultAlign

- 1-a Calcule une matrice de similarité des paires
- 1-b Construit un arbre de clustering hiérarchique suivant la méthode UPGMA
- 1-c Construit l'alignement multiple en suivant l'arbre
- **2- Reconstruit un arbre de clustering hiérarchique avec les nouveaux alignements paire à paire issus de l'alignement trouvée**
- **3- Réitère le processus jusqu'à stabilisation de l'arbre de clustering**

Alignements progressifs: MultAlign

- **Etape 1-c : Alignement progressif des séquences incorporées suivant l'ordre de l'arbre**




```
S2: AADKTNVKAAWSKVGGH-AGEYGA
S4: PADKTNVKAAWGKVGAH-AGEYGA
S1: PEEKSAVTALWGKVNVD---EVGG
S3: GEEKAAVLALWDKVNEE---EVGG
S5: EHEWQLVLHVWAKVEADVAGH-GQ
```

- **Etape 2-a : Recalcule les scores à partir des paires de séquences issus de l'alignement multiple**

- **Etape 2-b : Reconstruit une arbre de clustering hiérarchique avec les scores par la méthode UPGMA**

- **Etape 2-b : Reconstruit un alignement progressif**

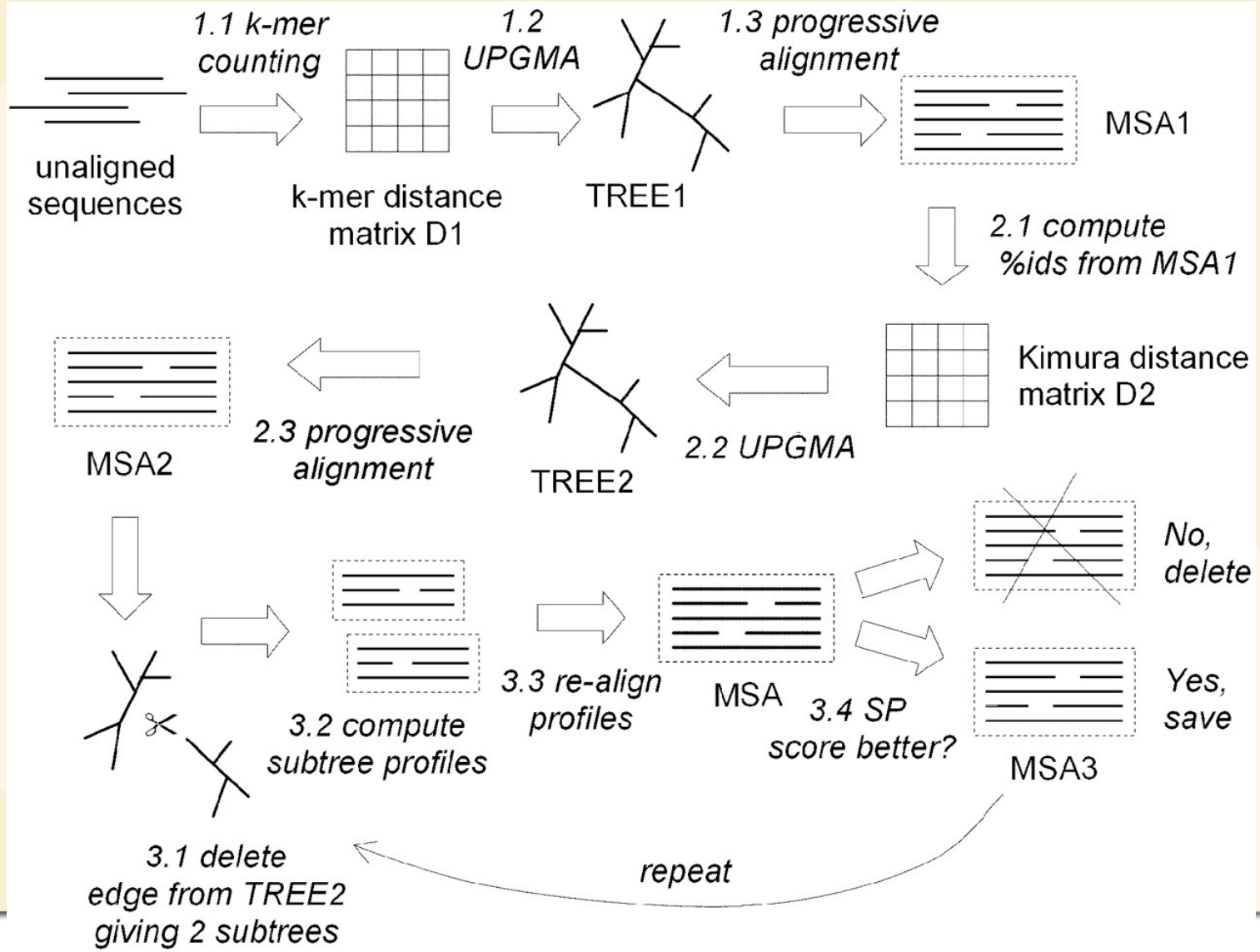
Alignements progressifs: MultAlign

- **Etape 3:** **Réitère le processus jusqu'à la stabilisation de l'arbre de clustering**

```
                1       10        20 23
                |-------+---------+--|
          s2    AADKTNVKAAWSKVGGHAGEYGA
          S4    PADKTNVKAAWGKVGAHAGEYGA
          S1    PEEKSAVTALWGKVNVDEVGG
          S3    GEEKAAVLALWDKVNEEEVGG
          S5    EHEWQLVLHVWAKVEADVAGHGQ
   Consensus    ..#k..V.a.W.KV.....g.g.
```

Rouge	**fort consensus, lettre majuscule**
Bleu	**faible consensus, lettre minuscule**
Noire	**neutre, pas de caractère reporté**

Alignements progressifs: Muscle



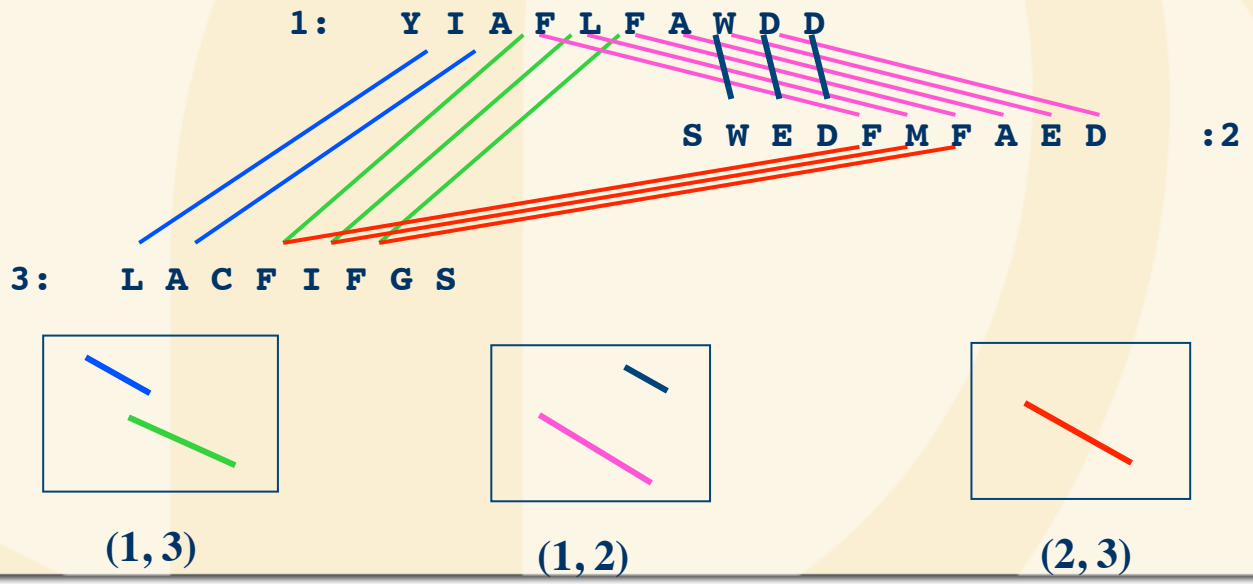
INSTITUT PASTEUR

Alignements progressifs: Dialign2

- **Principe**: ***Mixte entre alignement global et local***

  - ✓ **Alignement des régions très conservées.**
  - ✓ **Ancrage de l'alignement sur ce premier alignement**
  - ✓ **Alignement des régions moins conservées**

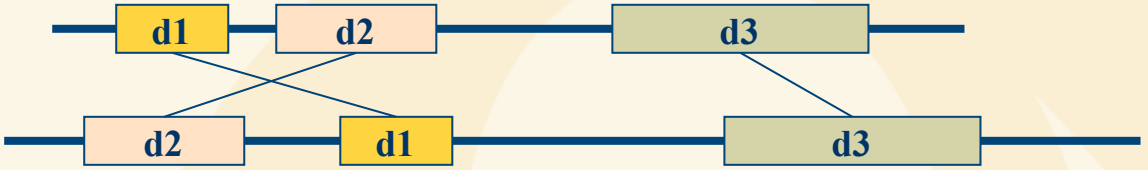
- **<u>Etape 1</u>: Détection des diagonales pour chaque paire de séquences et calcul du score.**



INSTITUT PASTEUR

Alignements progressifs: Dialign2

- **Etape 2**: Pour chaque couple de séquences, les diagonales compatibles (pas de croisement, pas de chevauchement) sont stockées en fonction du score.



- **Etape 3**: Mise en correspondance les zones conservées (score décroissant).

```
1:   y I A - F L F A W D d
3:   - L A c F I F g s - -
2:   s w e d F M F A E D -
```

Alignements: Quelles méthodes?

Clustalw

```
chite     ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD 47
mouse     -----KPKRPRSAYNIYVSESFQEAKDDSAQGK-LKLVN----EAWKNLSP 41
wheat     --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE 49
thybr     KKDSNAPKRAMTSFMFFS----SDFRSKHSDLS-IVEMSKAAGAAWKELGP 46
                ***. ::: .:     .. : . .. . :  :       *: *
```

Multalin

```
                    1       10        20        30        40        51
                    |-------+---------+---------+---------+---------+|
chite          ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat        DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
thybr     KKDSNAPKRAMTSFMFFSSDFRSKHSDLS-----IVEMSKAAGAAWKELGP
mouse           KPKRPRSAYNIYVSESFQEAKDDS-----AQGKLKLVNEAWKNLSP
Consensus  .....kPKRp.sa%....se.r.e.kd.s..........K..geaWk.Lsp
```

Dialign

```
chite          1   ---ADKPKRP LSAYMLWLNS ARESIKRENP DFK-VTEVAK KGGELWRGLK  D
wheat          1   --DPNKPKRA PSAFFVFMGE FREEFKQKNP KNKsVAAVGK AAGERWKSLS  E
thybr          1   kkDSNAPKRA MTSFMFFSSD FRSKHSDLSI VEM-SKAA-- --GAAWKELG  P
mouse          1   -----KPKRP RSAYNIYVSE SFQEAKDDSA QGK-LKLV-- --NEAWKNLS  P

                   0012479999 9999999888 8888666666 6660444411 1134444443  2
```

Alignements: Quelles méthodes?

Muscle

```
mouse          -----KPKRPRSAYNIYVSESFQEA-----KDDSAQGKLKLVNEAWKNLSP
chite          ---ADKPKRPLSAYMLWLNSARESIKREN-PDFKVTEVAKKGGELWRGLKD
wheat          --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
thybr          KKDSNAPKRAMTSFMFFSSDFRSKH-----SDLSIVEMSKAAGAAWKELGP
                     ***. ::: .: ..  ..       : .     *  .  *. *
```

Mafft

```
CLUSTAL format alignment by MAFFT (v6.859b)


chite          ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat          --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
thybr          KKDSNAPKRAMTSFMFFSSDFRSKHSDLS-----IVEMSKAAGAAWKELGP
mouse          -----KPKRPRSAYNIYVSESFQEAKDDSAQGK-----LKLVNEAWKNLSP
                     ***. ::: .: ..  .. .  .          *  .  *: *
```

T-coffee

```
chite          ---ADKPKRPLSAYMLWLNSARESIKRENP-DFKVTEVAKKGGELWRGLK D
mouse          -----KPKRPRSAYNIYVSESFQEAK-----DDSAQGKLKLVNEAWKNLS P
thybr          KKDSNAPKRAMTSFMFFSSDFRSKHS-----DLSIVEMSKAAGAAWKELG P
wheat          --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLS E
                     ***. ::: .: ..  .. .     : .     *  .  *: *
```

Prosite

```
TFAM_MOUSE/154-218      PKRPRSAYNIYVSESFQEAKDD--.--SAQGKLKLVNEAWKN
HMG1B_CHITE/5-71        PKRPLSAYMLWLNSARESIKRENP.DFKVTEVAKKGGELWRG
HMGL_WHEAT/42-111       PKRAPSAFFVFMGEFREEFKQKNPkNKSVAAVGKAAGERWKS
HMGL_TRYBR/206-270      PKRAMTSFMFFSSDFRSK--HS--.DLSIVEMSKAAGAAWKE
                        ***  ::: .: ..  ..         .     *  .  *:
```

INSTITUT PASTEUR

Alignements multiples, recherche et découverte de motifs

I. Alignements multiples

II. Définition de motifs

III. Recherche de motifs

IV. Autre utilisation

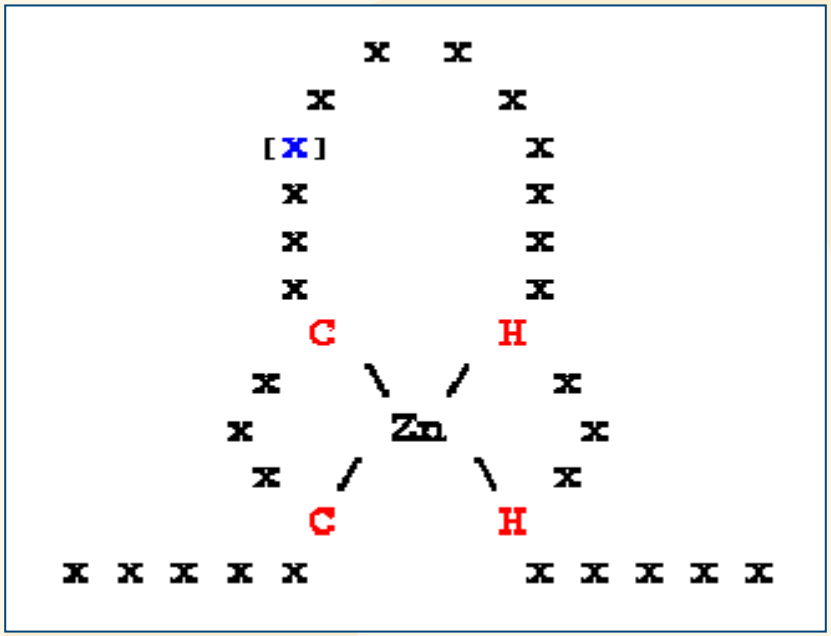
Domaine conservé

Exemple : Doigt de zinc

```
TYY1_HUMAN/383-407     YVCPF-DGCN---KKFAQSTNLKSHILT---H
YKQ8_CAEEL/78-102      YKCT---VCR---KDISSSESLRTHMFKQ-HH
BAS0_HUMAN/719-742     FQCD---ICK---KTFKNACSVKIHHKN--MH
ZG29_XENLA/62-84       FVCT---VCG---KTYKYKHGLNTHLHS---H
P43_XENBO/106-130      LKCSV-PGCK---RSFRKKRALRIHVSE---H
IKAR_MOUSE/488-512     FECN---MCG---YHSQDRYEFSSHITRG-EH
Q92610/1043-1069       YTCG---YCTEDSPSFPRPSLLESHISL--MH
TRA1_CAEEL/306-331     YKCEF-ADCE---KAFSNASDRAKHQNR--TH
ZN10_HUMAN/383-405     YKCN---QCG---IIFSQNSPFIVHQIA---H
GLI1_XENLA/283-310     FVCHW-QDCSRELRPFKAQYMLVVHMRR---H
XFIN_XENLA/276-298     FRCS---ECS---RSFTHNSDLTAHMRK---H
TF3A_BUFAM/72-97       CKCET-ENCN---LAFTTASNMRLHFKR--AH
ZG58_XENLA/174-196     FVCT---ECN---LSFAGLANLRSHQHL---H
P43_XENBO/163-187      YRCSY-EDCQ---TVSPTWTALQTHLKK---H
TSH_DROME/354-378      FRCV---WCK---QSFPTLEALTTHMKDS-KH
ZN76_HUMAN/165-189     FRCGY-KGCG---RLYTTAHHLKVHERA---H
TF3A_BUFAM/219-244     YRCPR-ENCD---RTYTTKFNLKSHILT--FH
SUHW_DROAN/349-373     YACK---ICG---KDFTRSYHLKRHQKYS-SC
ZN76_HUMAN/285-309     YTCPE-PHCG---RGFTSATNYKNHVRI---H
SRYC_DROME/469-492     FKCN---YCP---RDFTNFPNWLKHTRR--RH
EVI1_HUMAN/761-784     YRCK---YCD---RSFSISSNLQRHVRN--IH
...
```

Extrait de Pfam, entrée zf-C2H2

Domaine conservé



Motif :

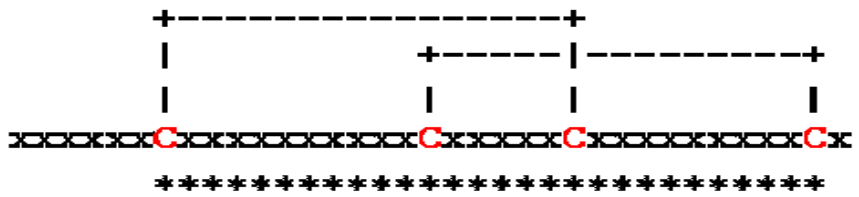
C-x (2,4)-C-x (3)-[LIVMFYWC]-x (8)-H-x (3,5)-H

Site de fixation de la cellulose

```
GUN1_TRIRE/427-455   HWGQCGGI---GYSGC--K-TCTSGTTCQYSNDYYSQCL
GUX1_TRIRE/481-509   HYGQCGGI---GYSGP--T-VCASGTTCQVLNPYYSQCL
GUX1_PHACH/484-512   QWGQCGGI---GYTGS--T-TCASPYTCHVLNPYYSQCY
GUX2_TRIRE/30-58     VWGQCGGQ---NWSGP--T-CCASGSTCVYSNDYYSQCL
GUN5_TRIRE/209-237   LYGQCGGA---GWTGP--T-TCQAPGTCKVQNQWYSQCL
GUNF_FUSOX/21-49     IWGQCGGN---GWTGA--T-TCASGLKCEKINDWYYQCV
GUX3_AGABI/24-52     VWGQCGGN---GWTGP--T-TCASGSTCVKQNDFYSQCL
Q01763/473-500       --SQCGGL---GYAGP--TgVCPSPYTCQALNIYYSQCI
GUX1_PENJA/505-533   DWAQCGGN---GWTGP--T-TCVSPYTCTKQNDWYSQCL
GUXC_FUSOX/482-510   QWGQCGGQ---NYSGP--T-TCKSPFTCKKINDFYSQCQ
GUX1_HUMGR/493-521   RWQQCGGI---GFTGP--T-QCEEPYICTKLNDWYSQCL
GUX1_NEUCR/484-512   HWAQCGGI---GFSGP--T-TCPEPYTCAKDHDIYSQCV
Q9Y894/23-53         PWGQCGGP---GWTGPttT-CCVTGCTCPVTND-YSQCL
PSBP_PORPU/26-54     LYEQCGGI---GFDGV--T-CCSEGLMCMKMGPYYSQCR
GUNB_FUSOX/29-57     VWAQCGGQ---NWSGT--P-CCTSGNKCVKLNDFYSQCQ
PSBP_PORPU/69-97     PYGQCGGM---NYSGK--T-MCSPGFKCVELNEFFSQCD
GUNK_FUSOX/339-370   AYYQCGGSKSAYPNGN--L-ACATGSKCVKQNEYYSQCV
PSBP_PORPU/128-156   EYAACGGE---MFMGA -K-CCKFGLVCYETSGKWSQCR
```

Extrait de Prosite, entrée PS00562

C-G-G-x(4,7)-G-x(3)-C-x(5)-C-x(3,5)-[NHG]-x-[FYWM]- x(2)-Q-C



Représentation de l'information contenue dans un alignement multiple

- ***Séquence consensus***: **pour chaque colonne de l'alignement, l'acide aminé le plus représenté est conservé: Une partie de l'information de l'alignement est perdu.**

```
G H E G V G K V V K L G A G A
G H E K K G Y F E D R G P S A
G H E G Y G G R S R G G G Y S
G H E F E G P K G C G A L Y I
G H E L R G T T F M P A L E C
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
G	H	E	GKFL	VKYER	G	KYGPT	VFRKT	VESGF	KDRCM	LRGP	GA	APGL	GSYE	ASIC

Consensus: G H E g e G g f s R G G L Y A

Représentation de l'information contenue dans un alignement multiple

• **Expression régulière/rationnelle: une certaine ambiguïté autorisé est sur chaque colonne de l'alignement: On perd une petite partie de l'information de l'alignement.**

G	H	E	G	V	G	K	V	V	K	L	G	A	G	A
G	H	E	K	K	G	Y	F	E	D	R	G	P	S	A
G	H	E	G	Y	G	G	R	S	R	G	G	G	Y	S
G	H	E	F	E	G	P	K	G	C	G	A	L	Y	I
G	H	E	L	R	G	T	T	F	M	P	A	L	E	C

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
G	H	E	GKFL	VKYER	G	KYGPT	VFRKT	VESGF	KDRCM	LRGP	GA	APGL	GSYE	ASIC

Pattern: G-H-E-x(2)-G-x(5)-[GA]-x(3)

Représentation de l'information contenue dans un alignement multiple

- **Expression régulière/rationnelle: Exemples**

- Post-translational signatures:
  - Protein splicing signature: [DNEG]-x-[LIVFA]-[LIVMY]-[LVAST]-H-N-[STC]
  - Tyrosine kinase phosphorylation site: [RK]-x(2)-[DE]-x(3)-Y or [RK]-x(3)-[DE]-x(2)-Y
  - ...
- DNA-RNA interaction signatures:
  - Histone H4 signature: G-A-K-R-H
  - p53 signature: M-C-N-S-S-C-[MV]-G-G-M-N-R-R
  - ...
- Enzymes:
  - L-lactate dehydrogenase active site: [LIVMA]-G-[EQ]-H-G-[DN]-[ST]
  - Ubiquitin-activating enzyme signature: P-[LIVM]-C-T-[LIVM]-[KRH]-x-[FT]-P
  - ...

Représentation de l'information contenue dans un alignement multiple

- **Expression régulière/rationnelle**

Residue Property	Residue Groups
Small	Ala, Gly
Small hydroxyl	Ser, Thr
Basic	Lys, Arg
Aromatic	Phe, Tyr, Trp
Basic	His, Lys, Arg
Small hydrophobic	Val, Leu, Ile
Medium hydrophobic	Val, Leu, Ile, Met
Acidic/amide	Asp, Glu, Asn, Gln
Small/polar	Ala, Gly, Ser, Thr, Pro

[ASGPT]-D-[IVLM]-G-X(5)-C-[DENQ]-R-[FYW]2-Q

↓ + Stricte

[AS]-D-[IVL]-G-X(4)-{PG}-C-[DE]-R-[FY]2-Q

Expressions rationnelles: Prosite

Alignement multiple (15 séquences)

A-[DE]-x (2)-W-G(2,4)

2400 mots

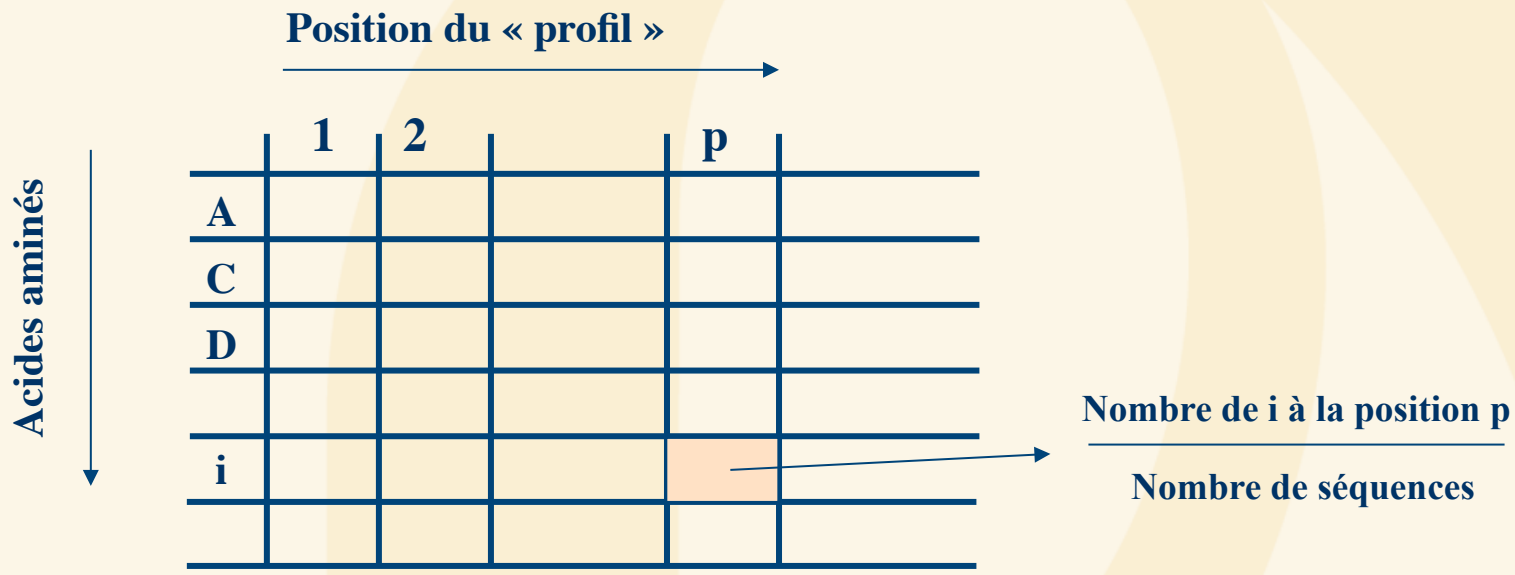
ADxxWGG
ADxxWGGG
ADxxWGGGG
AExxWGG
AExxWGGG
AExxWGGGG

ADAAWGG
ADACWGG
ADADWGG
:
:

- **L'information de l'alignement multiple est amplifiée.**

Représentation de l'information contenue dans un alignement multiple

- **Matrice de fréquence ou profile (Position-Spécifique-Score-Matrice: pssm): montre la distribution de probabilité d'un résidu (ligne) pour chaque position de l'alignement (colonne)**



Position Specific Scoring/Weight Matrices

G	H	E	G	V	G	K	V	V	K	L	G	A	G	A
G	H	E	K	K	G	Y	F	E	D	R	G	P	S	A
G	H	E	G	Y	G	G	R	S	R	G	G	G	Y	S
G	H	E	F	E	G	P	K	G	C	G	A	L	Y	I
G	H	E	L	R	G	T	T	F	M	P	A	L	E	C

Nombre de i à la position p

A	0	0	0	0	0	0	0	0	0	0	0	2	1	0	2
C	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
D	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
E	0	0	5	0	1	0	0	0	1	0	0	0	0	1	0
F	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0
G	5	0	0	2	0	5	1	0	1	0	2	3	1	1	0
H	0	5	0	0	0	0	1	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
K	0	0	0	1	1	0	0	1	0	1	0	0	0	0	0
L	0	0	0	1	0	0	0	0	0	0	1	0	2	0	0
M	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
P	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0
R	0	0	0	0	1	0	0	1	0	1	1	0	0	0	0
S	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1
T	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
V	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0
Y	0	0	0	0	1	0	1	0	0	0	0	0	0	2	0

Position Specific Scoring/Weight Matrices

Nombre de i à la position p

Nombre de séquences

G	H	E	G	V	G	K	V	V	K	L	G	A	G	A
G	H	E	K	K	G	Y	F	E	D	R	G	P	S	A
G	H	E	G	Y	G	G	R	S	R	G	G	G	Y	S
G	H	E	F	E	G	P	K	G	C	G	A	L	Y	I
G	H	E	L	R	G	T	T	F	M	P	A	L	E	C

A	0	0	0	0	0	0	0	0	0	0	0	0.4	0.2	0	0.4
C	0	0	0	0	0	0	0	0	0	0.2	0	0	0	0	0.2
D	0	0	0	0	0	0	0	0	0	0.2	0	0	0	0	0
E	0	0	1	0	0.2	0	0	0	0.2	0	0	0	0	0.2	0
F	0	0	0	0.2	0	0	0	0.2	0.2	0	0	0	0	0	0
G	1	0	0	0.4	0	1	0.2	0	0.2	0	0.4	0.6	0.2	0.2	0
H	0	1	0	0	0	0	0.2	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2
K	0	0	0	0.2	0.2	0	0	0.2	0	0.2	0	0	0	0	0
L	0	0	0	0.2	0	0	0	0	0	0	0.2	0	0.4	0	0
M	0	0	0	0	0	0	0	0	0	0.2	0	0	0	0	0

Prob(GHEFKGYKSMLAAYA)= 1*1 *1*0.2*0.2*1*0.2*0.2*0.2*0.2*0.2*0.4*0.2*0.4*0.4 =0.1024

Ln[Prob(GHEFKGYKSMLAAYA)]= 4 ln(1) + 8 ln(0.04) +3 ln(0.4) = 0 - 12.87 -2.75 = -15.62

Prob(AHEFKGYKSMLAAYA)= 0*1 *1*0.2*0.2*1*0.2*0.2*0.2*0.2*0.2*0.4*0.2*0.4*0.4 =0

Position Specific Scoring/Weight Matrices

- **Dans la pratique les profils tiennent compte d'autres paramètres**
  - ✓ **Pénalité pour les insertions/délétions**
  - ✓ **Problème si un acide aminé est sous ou pas représenté à une position donnée (la probabilité sera toujours nulle). L'acide aminé sera proscrit définitivement du model.**

- **Utilisation de pseudo-comptes**

$$x_{ip} = n_{i,p} + z_i$$

- **Utilisation des matrices de substitutions**

$$x_{ip} = \sum_j n_{j,p} \times M_{ij}$$

Position Specific Scoring/Weight Matrices

$$\frac{\text{Nb de i à la position p + 1}}{\text{Nb de seq (5) + Nb d'aa possible en p (20)}}$$

G	H	E	G	V	G	K	V	V	K	L	G	A	G	A
G	H	E	K	K	G	Y	F	E	D	R	G	P	S	A
G	H	E	G	Y	G	G	R	S	R	G	G	G	Y	S
G	H	E	F	E	G	P	K	G	C	G	A	L	Y	I
G	H	E	L	R	G	T	T	F	M	P	A	L	E	C

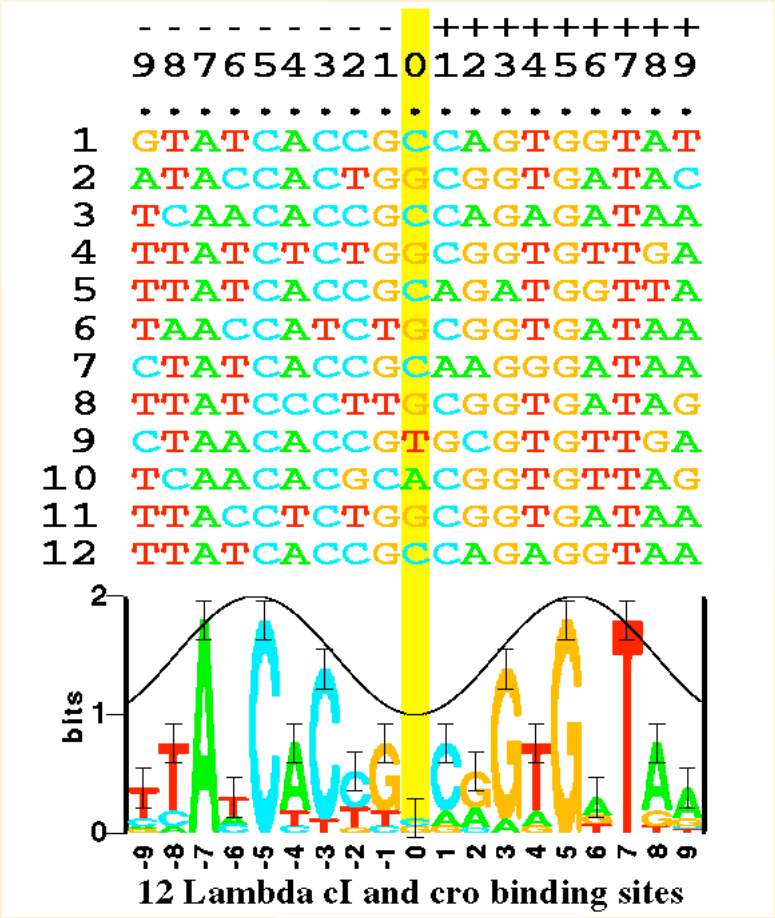
A	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.12	0.08	0.04	0.12
C	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.08	0.04	0.04	0.04	0.04	0.08
D	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.08	0.04	0.04	0.04	0.04	0.04
E	0.04	0.04	0.24	0.04	0.08	0.04	0.04	0.04	0.08	0.04	0.04	0.04	0.04	0.08	0.04
F	0.04	0.04	0.04	0.08	0.04	0.04	0.04	0.08	0.08	0.04	0.04	0.04	0.04	0.04	0.04
G	0.24	0.04	0.04	0.12	0.04	0.24	0.08	0.04	0.08	0.04	0.12	0.16	0.08	0.08	0.04
H	0.04	0.24	0.04	0.04	0.04	0.04	0.08	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
I	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.08
K	0.04	0.04	0.04	0.08	0.08	0.04	0.04	0.08	0.04	0.08	0.04	0.04	0.04	0.04	0.04
L	0.04	0.04	0.04	0.08	0.04	0.04	0.04	0.04	0.04	0.04	0.08	0.04	0.12	0.04	0.04
M	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.08	0.04	0.04	0.04	0.04	0.04

Prob(AHEFKGYKSMLAAYA)= 0*1 *1*0.2*0.2*1*0.2*0.2*0.2*0.2*0.2*0.4*0.2*0.4*0.4 =0

Prob(AHEFKGYKSMLAAYA)= 0.04*0.24 *0.24*0.08*0.08*0.24*0.08*0.08 *0.08*0.08*008*0.12*0.08*0.12*0.12 =$1.6e^{-15}$

Position Specific Scoring/Weight Matrices

- **Les informations d'un profile PSSMs peuvent être visualisées avec des séquences « LOGOs ».**



Schneider 1994

Hidden Markov Model (HMM)

***Application à l'analyse de séquence*:**

- **reconnaissance et classification: *alignement multiple et appartenance à une famille de protéines*.**
- **segmentation de séquence: *recherche de gènes et régions codantes***

***Deux utilisations des HMM avec les alignements multiples*:**

- ***séquences alignées: utiliser l'alignement pour estimer les paramètres du HMM*.**
- ***séquences NON alignées: utiliser l'architecture des HMM pour apprendre les propriétés intrinsèques des séquences.***

INSTITUT PASTEUR

Alignements multiples, recherche et découverte de motifs

I. Alignements multiples

II. Définition de motifs

III. Recherche de motifs

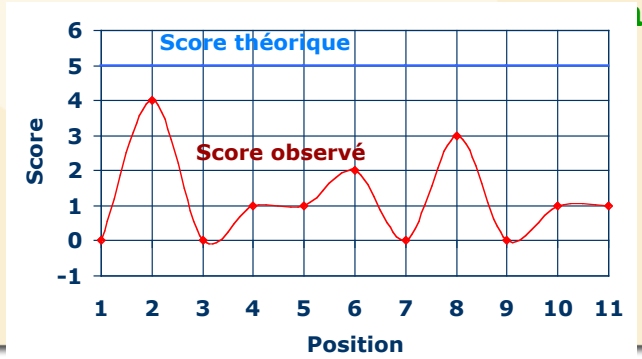
IV. Autre utilisation

Recherche d'un mot sur une séquence

- **Recherche des positions du mot *x* dans la séquence par balayages (gauche-droite) et décalages.**

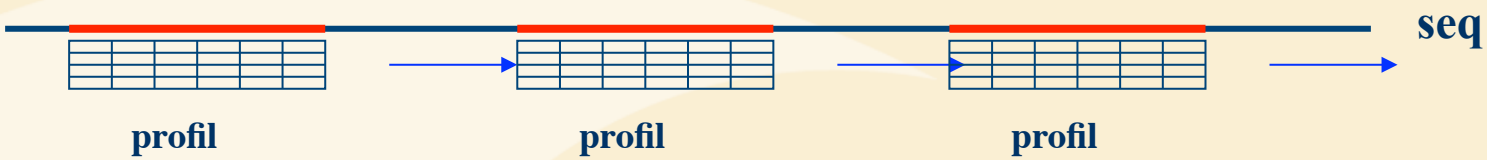
```
    a g c g a
... c a g c t a c a g t t a c a g...
```

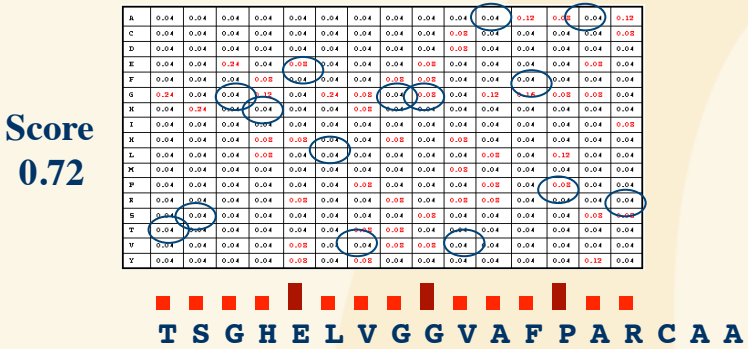
```
                                          Smax
c a g c t                  : 0 + 0 + 0 + 0 + 0 = 0
  a g c t a                : 1 + 1 + 1 + 0 + 1 = 4
    g c t a c              : 0 + 0 + 0 + 0 + 0 = 0
      c t a c a            : 0 + 0 + 0 + 0 + 1 = 1
        t a c a g          : 0 + 0 + 1 + 0 + 0 = 1
          a c a g t        : 1 + 0 + 0 + 1 + 0 = 2
            c a g t t      : 0 + 0 + 0 + 0 + 0 = 0
              a g t t a    : 1 + 1 + 0 + 0 + 1 = 3
                g t t a c  : 0 + 0 + 0 + 0 + 0 = 0
                  t t a c a: 0 + 0 + 0 + 0 + 1 = 1
                    t a c a g: 0 + 0 + 1 + 0 + 0 = 1
```



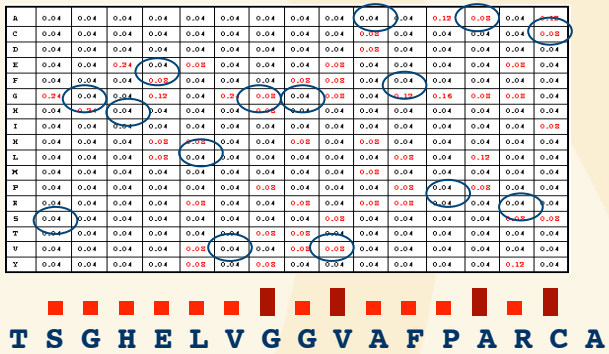
INSTITUT PASTEUR

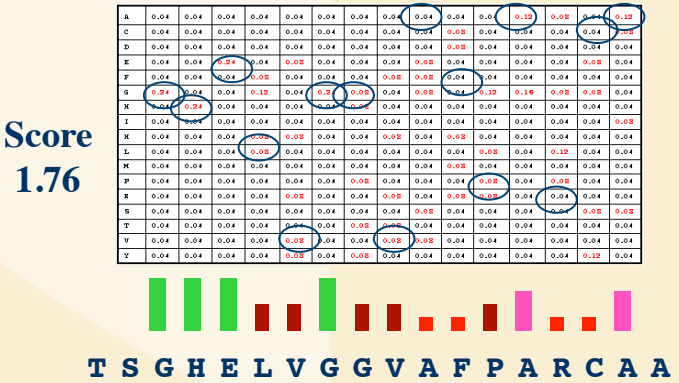
Recherche d'un profil sur une séquence





position + 1





position + 1

INSTITUT PASTEUR

Banques/Bases de motifs

- **Blocks** Multiply aligned ungapped segments corresponding to the most highly conserved regions of protein
- **Pfam** A Collection of Protein Domain Families (HMM)
- **Prints** Protein Motif Fingerprint Database
- **Prodom** A protein domain database which consists of homologous domains detected in the SWISS-PROT database
- **Prosite** Dictionary of Protein Sites and Patterns
- **Smart** Simple Modular Architecture Research Tool (HMM)
- **Tigrfams** Protein families based on Hidden Markov Models
- **Interpro** CATH, Uniprot, Prosite, Pfam, Prints, ProDom, Smart, Tigrfams, PirSF, Superfamily, gene3D, Panther, SCOP, SwissModel, ModBase

Prosite

Home ScanProsite ProRule Documents Downloads Links Funding

Database of protein domains, families and functional sites

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them [More details / References / Disclaimer / Commercial users].
PROSITE is complemented by **ProRule**, a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [More details].

Release 20.38, of 14-Oct-2008 (1539 documentation entries, 1315 patterns, 818 profiles and 819 ProRule)

PROSITE access

e.g: PDOC00022, PS50089, SH3, zinc finger

Search

add wildcard '*'

Browse:

- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hit

SRS - Sequence Retrieval System

PROSITE tools

Scan a sequence against PROSITE patterns and profiles - quick scan

(Output includes graphical view and feature detection)

Enter your sequence or a UniProtKB (Swiss-Prot or TrEMBL) ID or AC [help]:

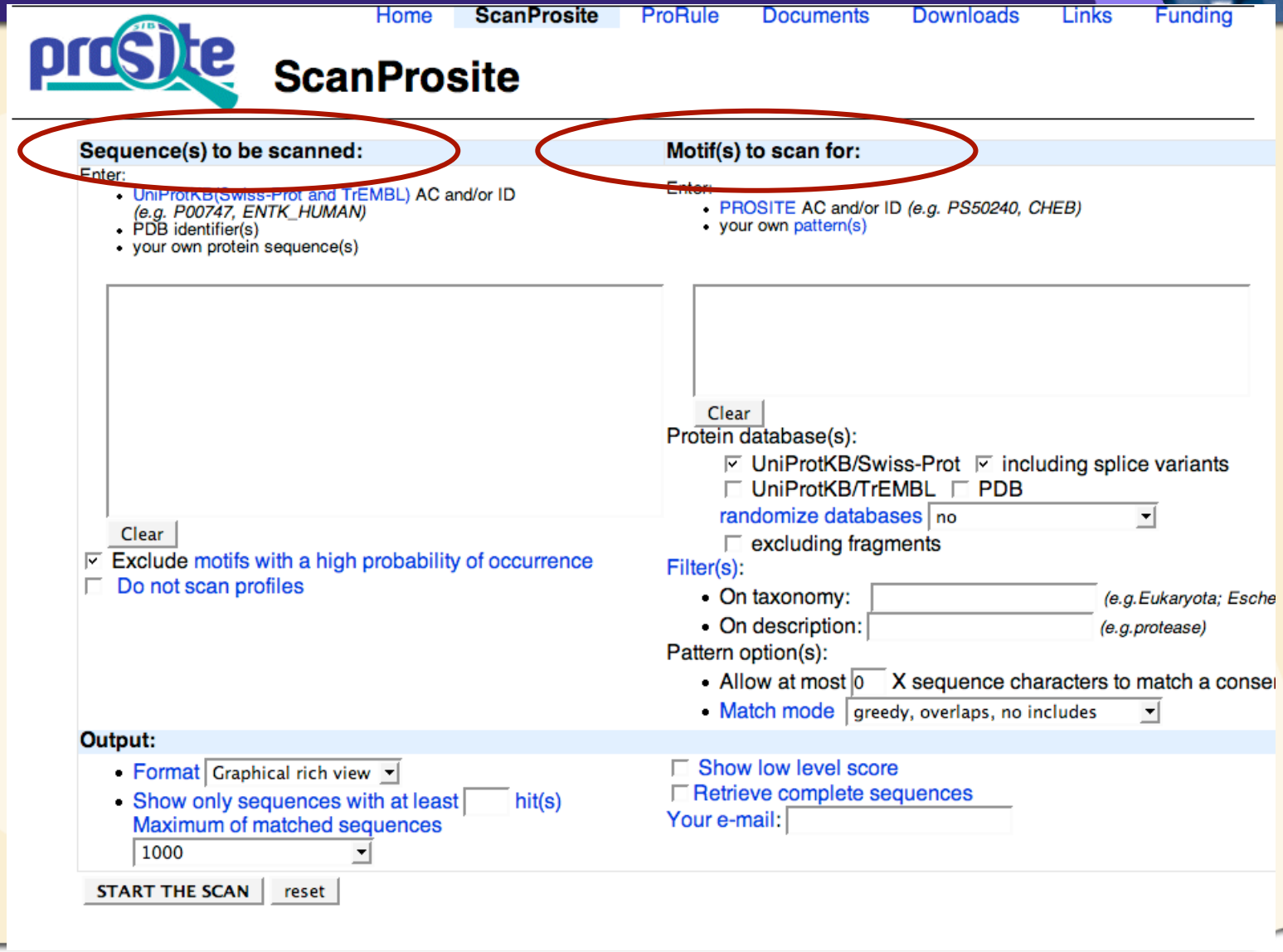
Scan Clear

exclude patterns with a high probability of occurrence

- **ScanProsite** - advanced scan
- **PRATT** - allows to interactively generate conserved patterns from a series of unaligned proteins.
- **MyDomains - Image Creator** new - allows to generate custom domain figures.

INSTITUT PASTEUR

Prosite



INSTITUT PASTEUR

Prosite

ScanProsite Results Viewer

This view shows ScanProsite results together with ProRule-based predicted intra-domain features (help).

Hits for USERPAT1{P-X(4)-C-D-X-R-[LIVM](2)-X-[KR]-X(14)-C} motif on all UniProtKB/Swiss-Prot (releas

found: 25 hits in 25 sequences

ruler: 1 100 200 300 400 500 600 700 800 900 1000

hits by patterns: [25 hits (by 1 pattern) on 25 sequences]

Hits by **USERPAT0** :
Pattern: **P - X(4) - C - D - X - R - [LIVM](2) - X - [KR] - X(14) - C**
Approximate number of expected random matches [Ref: PMID 11535175] in ~ 100'000 sequences (50'000'000 residues): 1.059938e-02

A3FFS8
(EPO_BOSMU) (192 aa) individual view

Erythropoietin. *Bos mutus grunniens (Wild yak) (Bos grunniens)*

27 - 54: ParliCDsRVLeRyileareaenatmgC

P48617
(EPO_BOVIN) (192 aa) individual view

Erythropoietin. *Bos taurus (Bovine)*

27 - 54: ParliCDsRVLeRyileareaenatmgC

Prosite: pattern

General information about the entry	
Entry name	**EPO_TPO**
Accession number	**PS00817**
Entry type	PATTERN
Date	OCT-1993 (CREATED); NOV-1995 (DATA UPDATE); JUL-1998 (INFO UPDATE).
PROSITE documentation	PDOC00644
Name and characterization of the entry	
Description	Erythropoietin / thrombopoietin signature.
Pattern	P-x(4)-C-D-x-R-[LIVM](2)-x-[KR]-x(14)-C.
Numerical results	

- SWISS-PROT release number: **40.7**, total number of sequence entries in that release: **103373**.
- Total number of hits in SWISS-PROT: **14 hits in 14 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **14 hits in 14 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **0 hits in 0 different sequences**
- Number of false hits (on unrelated proteins): **0 hits in 0 different sequences**
- Number of known missed hits: **0**
- Number of partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences: **1**
- Precision (true hits / (true hits + false positives)): **100.00 %**
- Recall (true hits / (true hits + false negatives)): **100.00 %**

Prosite: pattern

Comments

- Taxonomic range: **Eukaryotes**
- Maximum known number of repetitions of the pattern in a single protein: **1**
- `Interesting' site in the pattern: **3,disulfide**
- `Interesting' site in the pattern: **11,disulfide**

Cross-references

SWISS-PROT

True positive hits:

EPO_BOVIN (P48617), EPO_CANFA (P33707), EPO_FELCA (P33708),
EPO_HUMAN (P01588), EPO_MACFA (P07865), EPO_MACMU (Q28513),
EPO_MOUSE (P07321), EPO_PIG (P49157), EPO_RAT (P29676),
EPO_SHEEP (P33709), TPO_CANFA (P42705), TPO_HUMAN (P40225),
TPO_MOUSE (P40226), TPO_RAT (P49745)

`Potential' hits (partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences):

TPO_PIG (P42706)

Retrieve an alignment of SWISS-PROT true positive hits:

[Clustal format, color, condensed view] [Clustal format, color] [Clustal format, plain text] [Fasta format]

Prosite: profile

General information about the entry	
Entry name	**INTEIN_C_TER**
Accession number	**PS50818**
Entry type	MATRIX
Date	MAY-2002 (CREATED); MAY-2002 (DATA UPDATE); MAY-2002 (INFO UPDATE).
PROSITE documentation	PDOC00687
Name and characterization of the entry	
Description	Intein C-terminal splicing motif profile.
Matrix / Profile	/GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=22; /DISJOINT: DEFINITION=PROTECT; N1=3; N2=20; /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=0.8533; R2=0.02263959; TEXT='NScore'; /CUT_OFF: LEVEL=0; SCORE=290; N_SCORE=7.4; MODE=1; TEXT='!'; /CUT_OFF: LEVEL=-1; SCORE=249; N_SCORE=6.5; MODE=1; TEXT='?'; /DEFAULT: M0=-8; D=-20; I=-20; B0=-60; B1=-60; E0=-60; E1=-60; MI=-105; MD=-105; IM=-105; DM=-105; A B C D E F G H I K L M N P Q R S T V W Y Z /I: B0=0; B1=0; BI=-105; BD=-105; /M: SY='Y'; M=-15,-11,-29, -9, -6, 3,-23, -5,-14, -1,-13,-10,-10, -2, -9, 0,-12, -8,-15, -5, 12, -9; /M: SY='V'; M= 1,-25,-13,-27,-26, -1,-24,-27, 20,-18, 7, 6,-24,-26,-25,-17, -8, 1, 34,-27, -9,-26; /M: SY='Y'; M=-19,-17,-28,-18,-17, 29,-28, 9, -2,-11, -1, -1,-16,-27,-11, -9,-18,-10, -9, 17, 55,-17; /M: SY='D'; M=-13, 27, -1, 33, 5,-29,-12, -6,-30, -8,-25,-23, 15,-17, -6,-12, 4, -3,-22,-41,-21, -1; /M: SY='L'; M= -6,-25,-20,-27,-22, 3,-30,-24, 23,-25, 26, 13,-24,-25,-21,-21,-18, -4, 20,-23, -3,-23; /M: SY='T'; M= -1, 1,-18, -1, 4,-16,-10,-10,-14, -7,-16,-13, 3,-12, -3, -9, 11, 12, -9,-28,-11, 0; /M: SY='V'; M= 3,-18, -3,-21,-22, -8,-21,-25, 10,-18, 1, 0,-17,-24,-22,-19, -4, 4, 23,-30,-13,-22; /I: I=-5; MD=-27; /M: SY='E'; M=-10, 0,-25, 7, 15,-17,-19, -7,-21, 2,-19,-15, -5, 11, 0, -6, -6, -9,-21,-21, -9, 6; D=-5; /I: I=-5; MD=-27; /M: SY='N'; M= -2, 10,-17, 7, 0,-18, 5, -1,-18, -2,-18,-12, 13,-11, -3, -3, 2, -3,-16,-22,-13, -2; D=-5; /I: I=-5; MI=-27; MD=-27; IM=-27; DM=-27; /M: SY='H'; M=-12, 0,-26, 1, 5,-22, -6, 42,-25, -8,-18, -6, 5,-16, 7, -4, -4,-13,-24,-27, 0, 4; D=-5; /I: I=-5; DM=-27; /M: SY='N'; M= -5, 14,-19, 2, -5,-10, -8, -2,-13, -4,-15,-11, 27,-20, -5, -2, 3, 0,-17,-30,-13, -5; /M: SY='F'; M=-16,-25,-24,-30,-25, 40,-25, -8, 5,-22, 9, 3,-20,-29,-25,-17,-20,-10, 0, 8, 37,-25; /I: I=-6; MD=-32; /M: SY='V'; M= -6,-27,-18,-30,-25, 7,-30,-20, 24,-24, 18, 12,-23,-26,-23,-20,-16, -4, 25,-21, -1,-25; D=-6; /I: I=-6; MI=-32; IM=-32; DM=-32; /M: SY='A'; M= 15,-14,-14,-20,-15, -7,-14,-20, 1,-15, -2, -3,-12,-17,-13,-17, 4, 8, 8,-23,-11,-14; /I: I=-5; MD=-25;

INSTITUT PASTEUR

Prosite: profile

Numerical results

- SWISS-PROT release number: **40.17**, total number of sequence entries in that release: **108489**.
- Total number of hits in SWISS-PROT: **68 hits in 59 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **68 hits in 59 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **0 hits in 0 different sequences**
- Number of false hits (on unrelated proteins): **0 hits in 0 different sequences**
- Number of known missed hits: **0**
- Number of partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences: **0**
- Precision (true hits / (true hits + false positives)): **100.00 %**
- Recall (true hits / (true hits + false negatives)): **100.00 %**

Comments

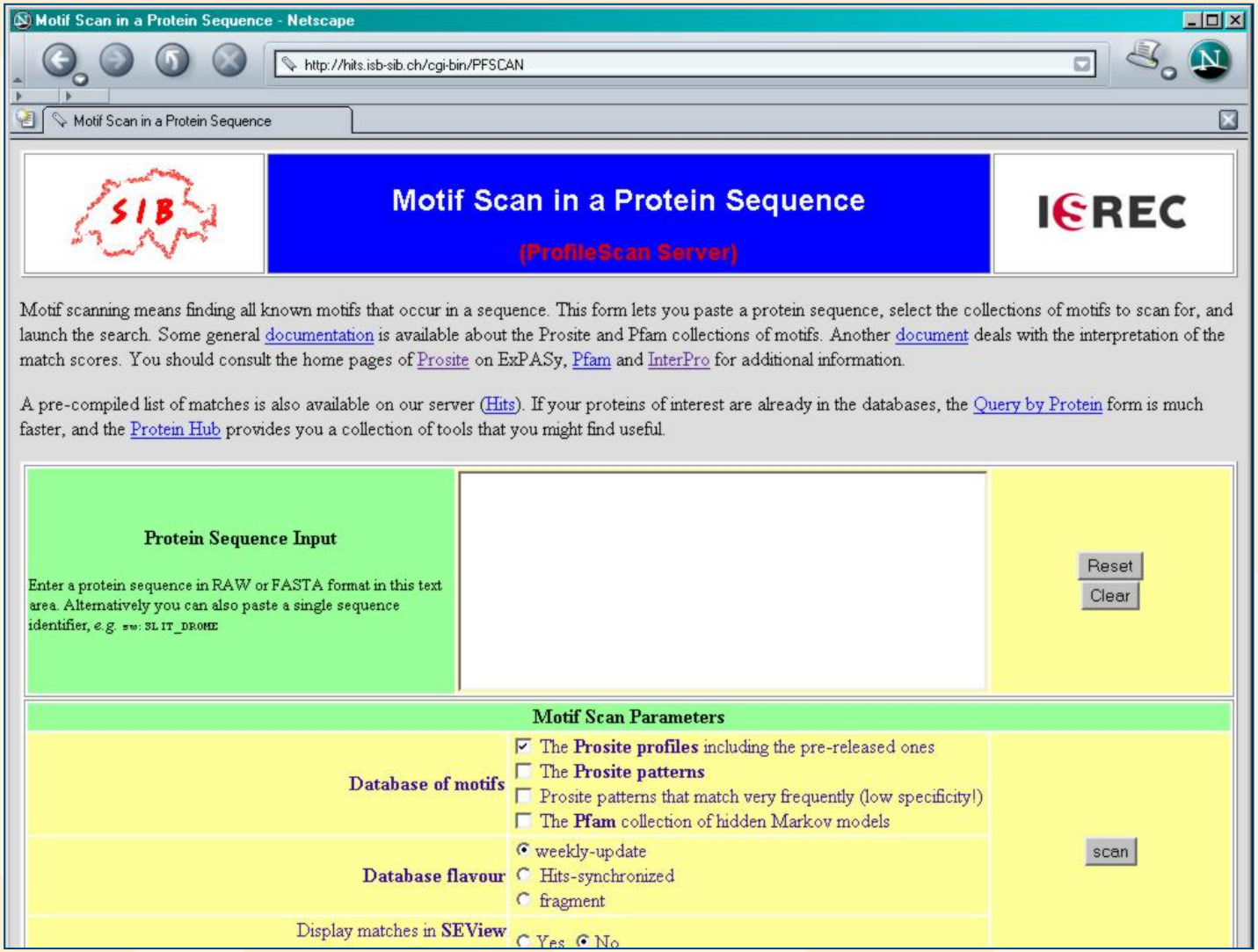
- MATRIX_TYPE: **protein_domain**
- Scaling database: **SWISS-PROT**: **reversed**
- Matrix author: **CJA_Sigrist**
- Taxonomic range: **Archaebacteria, Eukaryotes, Prokaryotes (Bacteria), Eukaryotic viruses**
- Maximum known number of repetitions of the pattern in a single protein: **3**

Cross-references

	True positive hits: DNAB_GUITH (O78411), DNAB_MYCTU (P71715), DNAB_PORPU (P51333),
	Y832_METJA (Q58242), YA54_METJA (Q58454), YE20_METJA (Q58815), YE61_MYCLE (Q49689), YE61_MYCTU (O53152) **Retrieve an alignment of SWISS-PROT true positive hits:** [Clustal format, color, condensed view] [Clustal format, color] [Clustal format, plain text] [Fasta format]
PDB new [Detailed view]	1AM2;

INSTITUT PASTEUR

MotifScan/ProfileScan



INSTITUT PASTEUR

InterPro


EMBL-EBI
European Bioinformatics Institute
Get Nucleotide sequences for Go ? Site search Go ?
EBI Home
About EBI
Research
Services
Toolbox
Databases
Downloads
Submissions
INTERPRO DATABASE
InterPro
InterPro Index
Text Search
Sequence Search
Databases
Documentation
FTP Site
InterPro
InterPro is a useful resource for whole genome analysis and has already been used for the proteome analysis of a number of completely sequenced organisms including preliminary analyses of the mouse and human genomes.
InterPro
Further information on InterPro can be found in the Documentation page, which includes links to the release notes, the user manual, a list of deleted InterPro entries, the dataflow scheme of the database, a fully annotated sample entry and references for the member databases.
InterPro is headed by Rolf Apweiler.
Updated Documents and New Links
Announcement: InterPro release 5.1 is out with new data and updated files.
News: InterPro has a new SRS-based text search which allows users to search a combination of InterPro and protein features.
List of all InterPro entries of each type
Proteome Analysis
Proteome Analysis
Statistical and comparative analysis of the predicted proteomes of fully sequenced organisms.
QuickGO
QuickGO
GO Browser

InterPro

INSTITUT PASTEUR

wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam

keyword search Go

Pfam 24.0 (October 2009, 11912 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM FAMILY	View Pfam family annotation and alignments
VIEW A CLAN	See groups of related families
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	enter any accession or ID Go Example

Enter any type of accession or ID to jump to the page for a Pfam family or clan, UniProt sequence, PDB structure, etc.

Or view the help pages for more information

Citing Pfam

If you find Pfam useful, please consider citing the reference that describes this work:

The Pfam protein families database: R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunesekaran, G. Ceric, K. Forslund, L. Holm, E.L. Sonnhammer, S.R. Eddy, A. Bateman
Nucleic Acids Research (2010) Database Issue 38:D211-222

Mirrors

The following are official Pfam mirror sites:

- WTSI, UK
- SBC, Sweden
- JFRC, USA
- INRA, France
- CCBB, South Korea

INSTITUT PASTEUR

wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam

keyword search Go

Protein: *SRC_MOUSE* (P05480)

1 architecture | 1 sequence | 0 interactions | 1 species | 0 structures

- Summary
- Features
- Sequence
- Interactions
- Structures
- TreeFam

Jump to...

enter ID/acc Go

Summary

SRC_MOUSE

This is the summary of UniProt entry SRC_MOUSE (P05480).

Description:	Neuronal proto-oncogene tyrosine-protein kinase Src EC=2.7.10.2
Source organism:	Mus musculus (Mouse) (NCBI taxonomy ID 10090) View Pfam proteome data.
Length:	541 amino acids

Please note: when we start each new Pfam data release, we take a copy of the UniProt sequence database. This snapshot of UniProt forms the basis of the overview that you see here. It is important to note that, although some UniProt entries may be removed *after* a Pfam release, these entries will not be removed from Pfam until the *next* Pfam data release.

Pfam domains

This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will take you to the page describing that Pfam entry. The table below gives the domain boundaries for each of the domains.

Source	Domain	Start	End
Pfam A	SH3_1	89	142
Pfam A	SH2	156	238
Pfam A	Pkinase_Tyr	275	524

INSTITUT PASTEUR

wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

Family: *SH3_1* (PF00018)

357 architectures | 5643 sequences | 11 interactions | 213 species | 391 structures

- Summary
- Domain organisation
- Alignments
- HMM logo
- Trees
- Curation & models
- Species
- Interactions
- Structures

Jump to... enter ID/acc Go

Summary

SH3 domain

Add annotation

SH3 (Src homology 3) domains are often indicative of a protein involved in signal transduction related to cytoskeletal organisation. First described in the Src cytoplasmic tyrosine kinase P12931. The structure is a partly opened beta barrel.

Example structure

PDB entry 2enm: Solution structure of the SH3 domain from mouse sorting nexin-9

View a different structure: 2enm

Literature references

1. Kami K, Takeya R, Sumimoto H, Kohda D; , EMBO J 2002;21:4268-4276.: Diverse recognition of non-PxxP peptide ligands by the SH3 domains from p67(phox), Grb2 and Pex13p. PUBMED:12169629

InterPro entry IPR001452

SH3 (src Homology 3) domains are small protein modules containing approximately 50 amino acid residues PUBMED:15335710, PUBMED:11256992. They are found in a great variety of intracellular or membrane-associated proteins PUBMED:1639195, PUBMED:14731533, PUBMED:7531822 for example, in a variety of proteins with enzymatic activity, in adaptor proteins that lack catalytic sequences and in cytoskeletal proteins, such as fodrin and yeast actin binding protein ABP-1.

The SH3 domain has a characteristic fold which consists of five or six beta-strands arranged as two tightly packed anti-parallel beta sheets. The linker regions may contain short helices PUBMED:. The surface of the SH3-domain bears a flat, hydrophobic ligand-binding pocket which consists of three shallow grooves defined by conservative aromatic residues in which the ligand adopts an extended left-handed helical arrangement. The ligand binds with low affinity but this may be enhanced by multiple interactions. The region bound by the SH3 domain is in all cases proline-rich and contains PXXP as a core-conserved binding motif. The function of the SH3 domain is not well understood but they may mediate many diverse processes such as increasing local concentration of proteins, altering their subcellular location and mediating the assembly of large multiprotein complexes PUBMED:7953536.

Clan

This family is a member of clan **SH3** (CL0010), which contains the following 5 members:

SH3_1 SH3_2 SH3_3 SH3_4 SH3_5

Internal database links

SCOOP:	3D SH3_2 SH3_5
Similarity to PfamA using HHSearch:	SH3_2

External database links

INSTITUT PASTEUR

INSTITUT PASTEUR

Alignements multiples, recherche et découverte de motifs

I. Alignements multiples

II. Définition de motifs

III. Recherche de motifs

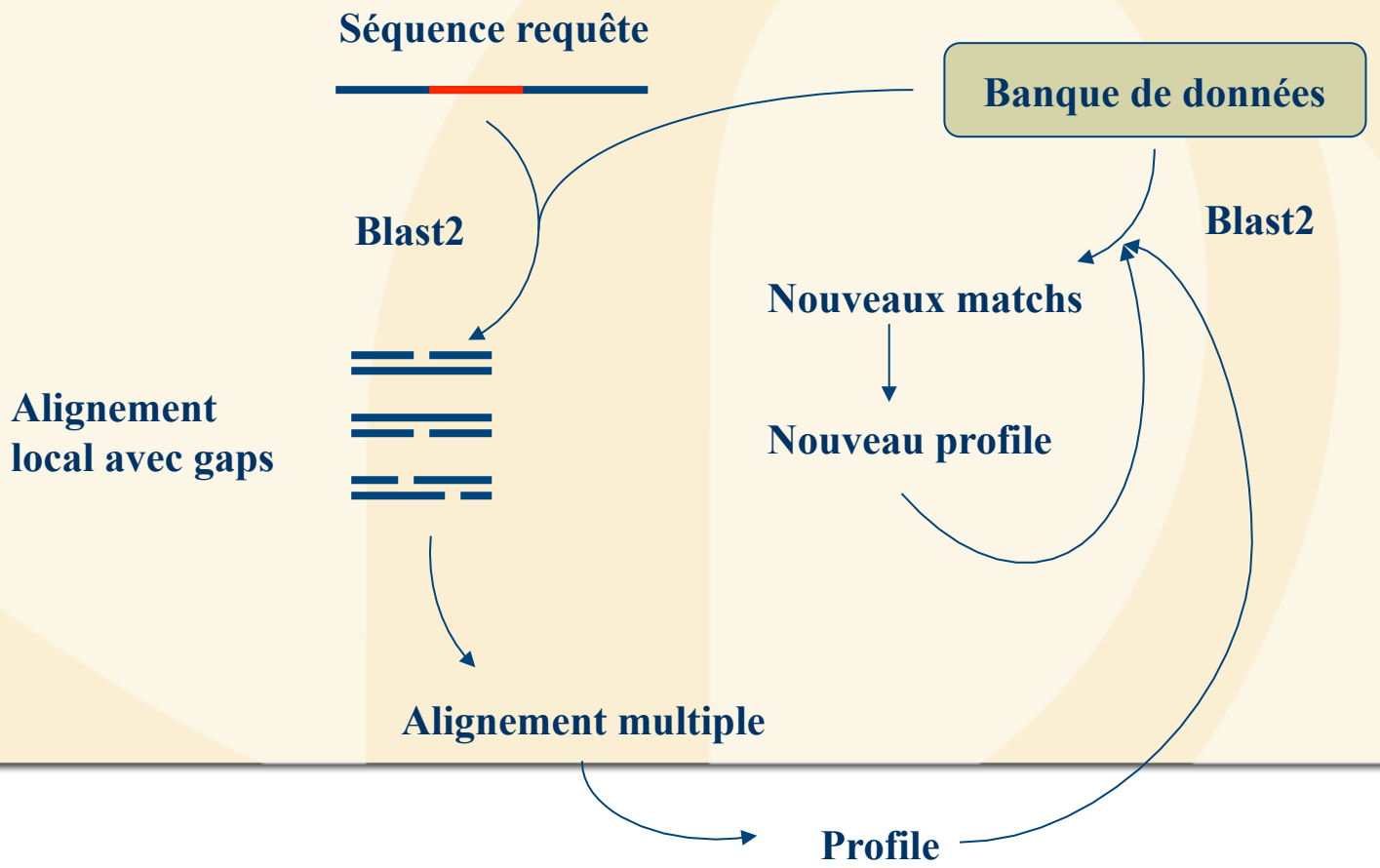
IV. Autre utilisation

PSI et PHY BLAST

Psi-Blast et **Phy-Blast** utilisent un motif et/ou un alignement multiple pour augmenter la sensibilité/sélectivité d'une recherche de similitude dans les banques de données.

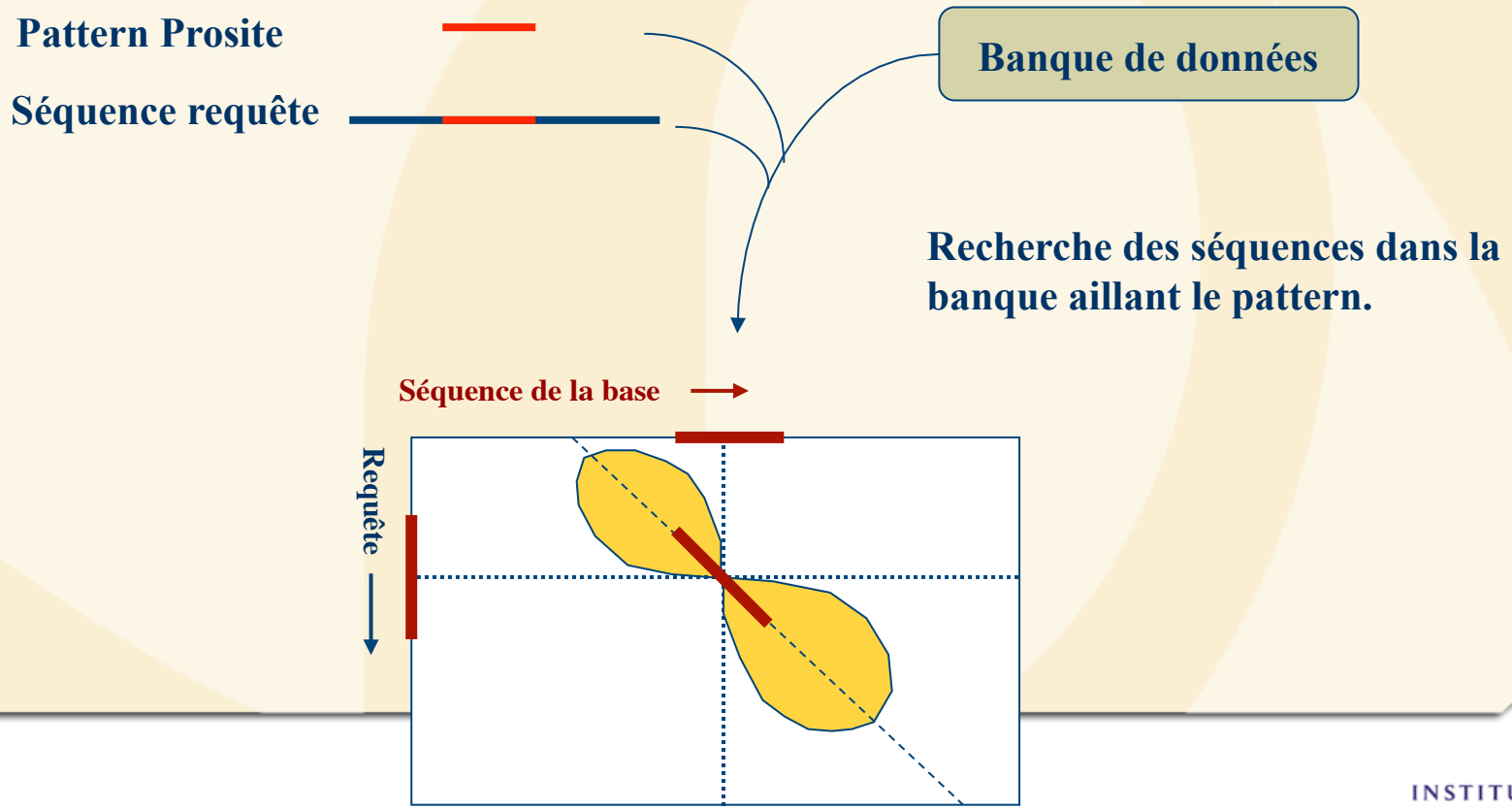
PSI-BLAST: Position Specific Iterated

- **Les séquences extraites d'une recherche Blast2 sont alignées et un profil statistique est construit. La recherche est réitérée avec ce profil comme requête, autant de fois que le souhaite l'utilisateur.**



PHI-BLAST: Pattern Hit Initiated

- L'espace de recherche est limité aux séquences contenant un motif spécifié par l'utilisateur et qui doit également être contenu dans la séquence requête. L'alignement final est ancré sur ce motif.



INSTITUT PASTEUR

Fin