Initiation à la bio-informatique
Module 2 : Alignement de séquences

Partie 2 :

Alignements splicés et Alignements Multiples

Ségolène Caboche

Université de Lille - TAG

(segolene.caboche@pasteur-lille.fr)

22 et 23 février 2022

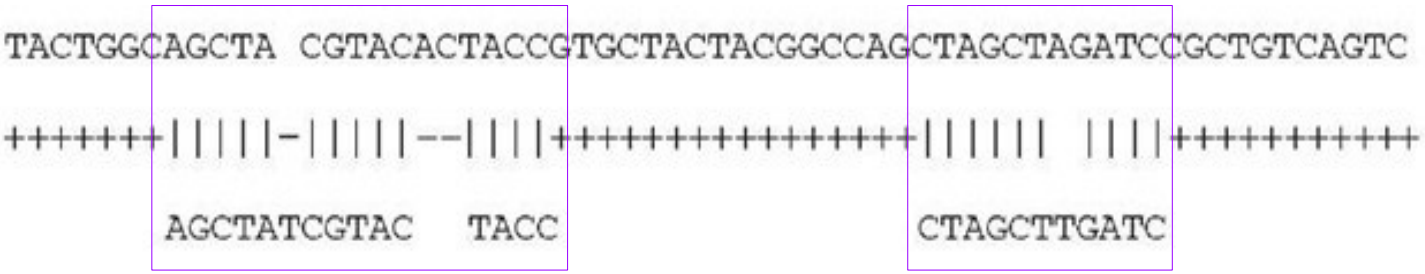
Les alignements splicés

Alignements splicés

- But : aligner des gènes multi-exons avec un cDNA similaire ou une séquence protéique
  - Soit une sous-séquence génomique
  - Soit un génome complet
- Une séquence de cDNA peut être complète ou partielle, par exemple les EST expressed sequence tag
- Les coordonnées des exons sont identifiées par homologie de séquence et la présence de sites d'épissage
- Approche très fiable si les séquences (cDNA ou protéine) sont très similaires à la séquence génomique

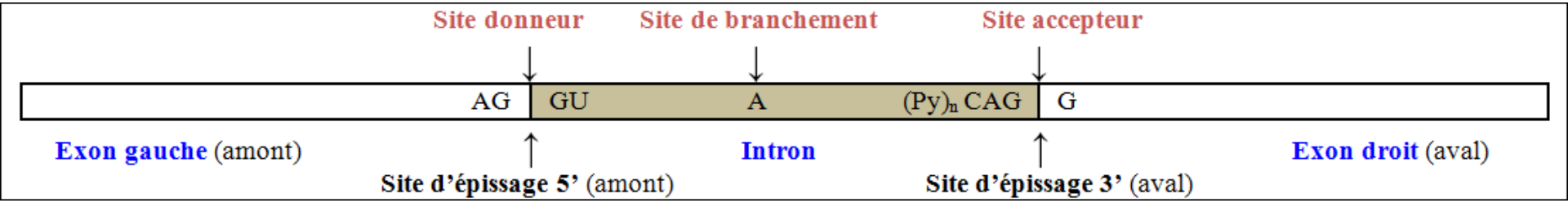
Alignements splicés : spécificités

- 2 types de gaps:
  - Les gaps exoniques (-): représentent des différences mineures entre les exons des 2 séquences
  - Les gaps introniques (+) : représentent des différences majeures dans les introns ou entre les 2 séquences



Alignements splicés : spécificités

- Présence de sites canoniques d'épissage :
  - Site donneur : présence du dinucléotide GT
  - Site accepteur : présence du dinucléotide AG



=> Utilisation de ces spécificités pour construire un alignement splicé

Alignements splicés : définition

- Un alignement splicé optimal entre 2 séquences A et B :
  - est un alignement de score maximum
  - les exons de A doivent être alignés avec des régions de B
- Les introns et les régions intergéniques de A doivent êtres traités comme des gaps introniques
- Les limites exons-introns doivent être exactes notamment grâce aux sites d'épissage donneurs et accepteurs

Algorithmes pour l'alignement splicé

- Par programmation dynamique (avec des matrices supplémentaires pour la gestion des différents types de gaps et les sites d'épissage)
- Heuristiques pour la comparaison rapide de séquences génomiques et de cDNA : méthode générale
  - Etape1 : matches de mots courts entre la séquence génomique et le cDNA (BLAST-like) => identification de chaînes de HSPs (=approximation d'un alignement)
  - Etape2 : pour chaque chaîne de score élevé entre une région génomique et le cDNA, calcul de la matrice de programmation dynamique couvrant tous les HSP et obtention de l'alignement splicé
  - Variation de cette stratégie générale utilisée dans différents programmes

BLAT

- BLAT peut être utilisé pour aligner des séquences nucléiques ou protéiques ou traduites (mRNA) contre une séquence génomique
- Développé pour travailler avec des séquences très similaires
- Les séquences protéiques ou traduites sont plus efficaces pour identifier des matches distants et pour une analyse inter-espéces que les séquences nucléiques
- Méthode :
  - Table contenant tous les mots non-chevauchants de taille k au sein des séquences génomiques (8<=k<=16)
  - L'ensemble des séquences de cDNA est scanné pour localiser les matches exacts ou similaires
  - Si il existe un nombre de matches suffisant, les mots sont étendus pour former des HSP. Les HSP proches sont liés ensemble et alignés.
- BLAT est développé pour trouver des matches entre séquences de longueur supérieure à 40 bases qui partagent ≥95% d'identité ou ≥80% d’identité pour les séquences traduites en protéines

est2genome

- Méthode en 3 étapes :
  - Une séquence génomique est comparée avec un ensemble de séquences de cDNA avec BLAST
  - Pour chaque hit, les positions de début et de fin d'un alignement local exact (Smith-Waterman) sont calculées
  - Les régions correspondantes de la séquence génomique et du cDNA sont ensuite extraites et alignées, et un alignement splicé optimal est calculé par programmation dynamique. Les dinucléotides GT-AG sont utilisés pour les sites d'épissage
- est2genome est bon pour les alignements inter-espèces

Sim4

- Le programme sim4 est divisé en 4 étapes
  - Étape 1 : matches exacts de mots de longueur 12 entre les séquences génomiques et cDNA qui sont étendues en HSP
  - Étape 2 : les HSP sont combinés en chaînes.
  - Étape 3 : les limites des régions exoniques sont déterminées par une **méthode rapide** basée sur la similarité et les dinucleotides GT-GA.
  - Étape 4 : pour chaque paires de régions exoniques dans la séquences génomique et le cDNA, un alignement est produit
- sim4 est capable de faire **rapidement** des comparaison inter-espèces

GeneSeqer

- Les paires de séquences similaires sont identifiées en étendant des mots exacts de taille 12 en HSP qui sont combinées en chaînes
- Les sites d'épissage sont scorés en utilisant une **méthode statistique** (Brendel and Kleffe, 1998) plutôt que les dinucléotides GT-AG.
- Le gros avantage de GeneSeqer est qu'il peut **identifier des exons courts** sur la base des sites d'épissage

DDS/GAP2

- Étape 1 identique aux autres programmes
- Le programme GAP2 calcule un alignement pour chaque paire de régions
- Le programme DDS/GAP2 est capable de générer des alignements inter-espèces

Bilan

- 4 programmes (DDS/GAP2, est2genome, sim4 et GeneSeqer) ont été comparés (Haas et al. 2002)
  - 5016 séquences de cDNA sequences (Arabidopsis)
- Sites d'épissage identiques pour 4918
- Les programmes donnent des résultats divergents pour moins de 2 % des cDNA
- Cependant, les programmes montrent des avantages différents :
  - DDS/GAP2 est meilleur pour aligner le cDNA complet sur le génome
  - GeneSeqer est excellent pour identifier des exons court (3-25bp)
  - SIM4 est le plus rapide

Séquence protéique vs. Séquence nucléique

- Cas spécial de l'alignement splicé : alignement d'une protéine sur une séquence nucléique
- Traduction de la séquence nucléique dans les 6 phases de lecture
- Plusieurs programmes disponibles
  - BLAT (entrée : protéines ou séquences nucléiques)
  - GeneWise (entrée : protéines)
  - Exonerate (comparaison de séquences, similaire à GeneWise pour la comparaison séquences protéiques/nucléiques)
  - Spaln (entrée : protéines ou séquences nucléiques)
  - Scipio (entrée : protéines)

GeneWise

- GeneWise compare directement une protéine à une séquence d'ADN génomique, en prenant en compte les propriétés statistiques des structures de gènes et la présence d'erreurs de séquençage
- Basé sur des chaînes de Markov cachées (HMM)
- Programme très utilisé

Exonerate

- Logiciel de comparaison de séquences 2 à 2
- Inclue la comparaison de séquences protéiques vs. Séquences nucléiques
- Algorithme similaire à GeneWise avec des heuristiques

Exonerate

GAPPED ALIGNMENT OPTIONS

-m | --model <alignment model>

Specify the alignment model to use. The models currently supported are:

ungapped

The simplest type of model, used by default. An appropriate model with be selected automatically for the type of input sequences provided.

ungapped:trans

This ungapped model includes translation of all frames of both the query and target sequences. This is similar to an ungapped tblastx type search.

affine:global

This performs gapped global alignment, similar to the Needleman-Wunsch algorithm, except with affine gaps. Global alignment requires that both the sequences in their entirety are included in the alignment.

affine:bestfit

This performs a best fit or best location alignment of the query onto the target sequence. The entire query sequence will be included in the alignment, but only the best location for its alignment on the target sequence.

affine:local

This is local alignment with affine gaps, similar to the Smith-Waterman-Gotoh algorithm. A general-purpose alignment algorithm. As this is local alignment, any subsequence of the query and target sequence may appear in the alignment.

affine:overlap

This type of alignment finds the best overlap between the query and target. The overlap alignment must include the start of the query or target and the end of the query or the target sequence, to align sequences which overlap at the ends, or in the mid-section of a longer sequence.. This is the type of alignment frequently used in assembly algorithms.

est2genome

This model is similar to the affine:local model, but it also includes intron modelling on the target sequence to allow alignment of spliced to unspliced coding sequences for both forward and reversed genes. This is similar to the alignment models used in programs such as EST_GENOME and sim4.

ner

NERs are non-equivalenced regions - large regions in both the query and target which are not aligned. This model can be used for protein alignments where strongly conserved helix regions will be aligned, but weakly conserved loop regions are not. Similarly, this model could be used to look for co-linearly conserved regions in comparison of genomic sequences.

Exonerate

protein2dna

This model compares a protein sequence to a DNA sequence, incorporating all the appropriate gaps and frameshifts.

protein2dna:bestfit

NEW: This is a bestfit version of the protein2dna model, with which the entire protein is included in the alignment. It is currently only available when using exhaustive alignment.

protein2genome

This model allows alignment of a protein sequence to genomic DNA. This is similar to the protein2dna model, with the addition of modelling of introns and intron phases. This model is simliar to those used by genewise.

protein2genome:bestfit

NEW: This is a bestfit version of the protein2genome model, with which the entire protein is included in the alignment. It is currently only available when using exhaustive alignment.

coding2coding

This model is similar to the ungapped:trans model, except that gaps and frameshifts are allowed. It is similar to a gapped tblastx search.

coding2genome

This is similar to the est2genome model, except that the query sequence is translated during comparison, allowing a more sensitive comparison.

cdna2genome

This combines properties of the est2genome and coding2genome models, to allow modeling of an whole cDNA where a central coding region can be flanked by non-coding UTRs. When the CDS start and end is known it may be specified using the --annotation option (see below) to permit only the correct coding region to appear in the alignemnt.

genome2genome

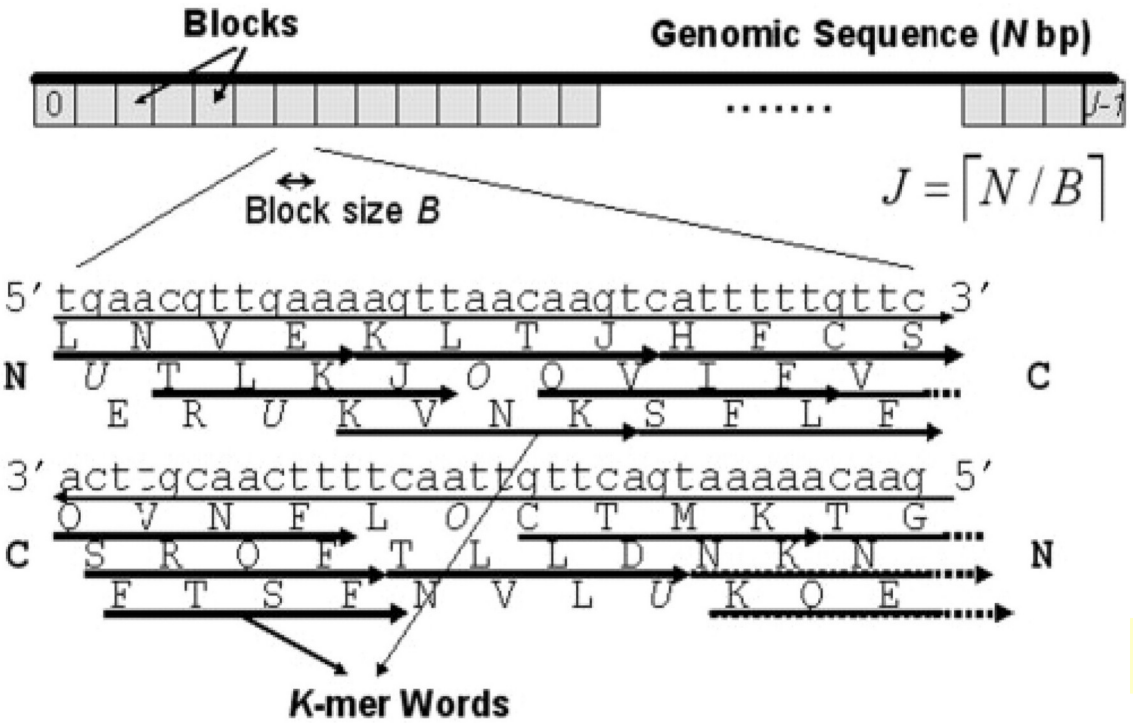
This model is similar to the coding2coding model, except introns are modelled on both sequences. (not working well yet)

Spaln

- Entrée : cDNA et protéines
- Particularité : l'étape 1 est différente des autres méthodes
  - Basée sur des blocs pour identifier les paires de régions similaires
- Alignement efficace basée sur la programmation dynamique

Spaln

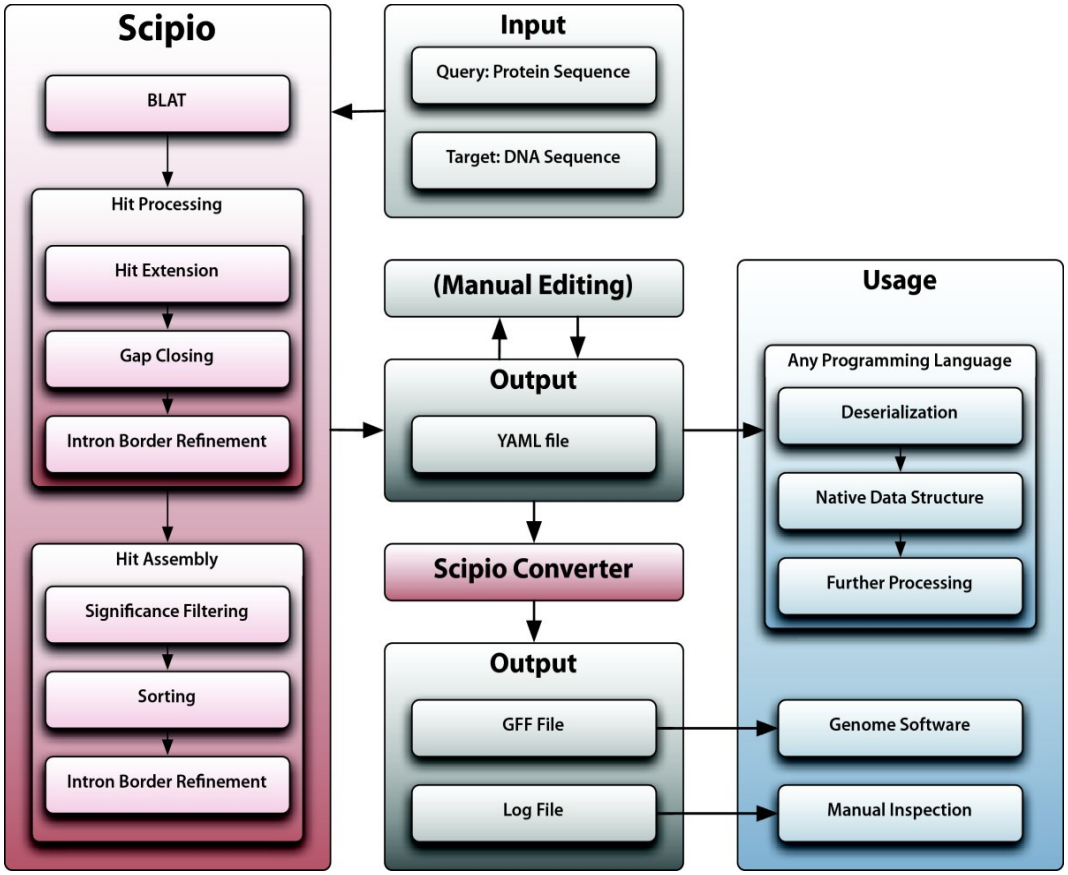
- La séquence génomique est divisée en blocs de longueur fixe B et traduction dans les 6 phases de lecture
- Pour chaque bloc, les k-mers non-chevauchants ne contenant pas de codons de terminaison (O ou U) ou de N sont stockés
- Ce sont les blocs qui sont comparés entre la protéine et le génome



Spicio

- Basé sur BLAT
- Au lieu de produire un ensemble de hits, le programme fournit un ensemble cohérent de positions possibles pour une protéine sur un génome
- La sortie contient aussi des informations sur les erreurs de séquençage

Spicio


Scipio
BLAT
Hit Processing
Hit Extension
Gap Closing
Intron Border Refinement
Hit Assembly
Significance Filtering
Sorting
Intron Border Refinement
Input
Query: Protein Sequence
Target: DNA Sequence
(Manual Editing)
Output
YAML file
Scipio Converter
Output
GFF File
Log File
Usage
Any Programming Language
Deserialization
Native Data Structure
Further Processing
Genome Software
Manual Inspection

Les alignements splicés : Exercices

Exercice 9 partie 2

Les alignements multiples

Définition de l'alignement multiple

- entrée : k séquences

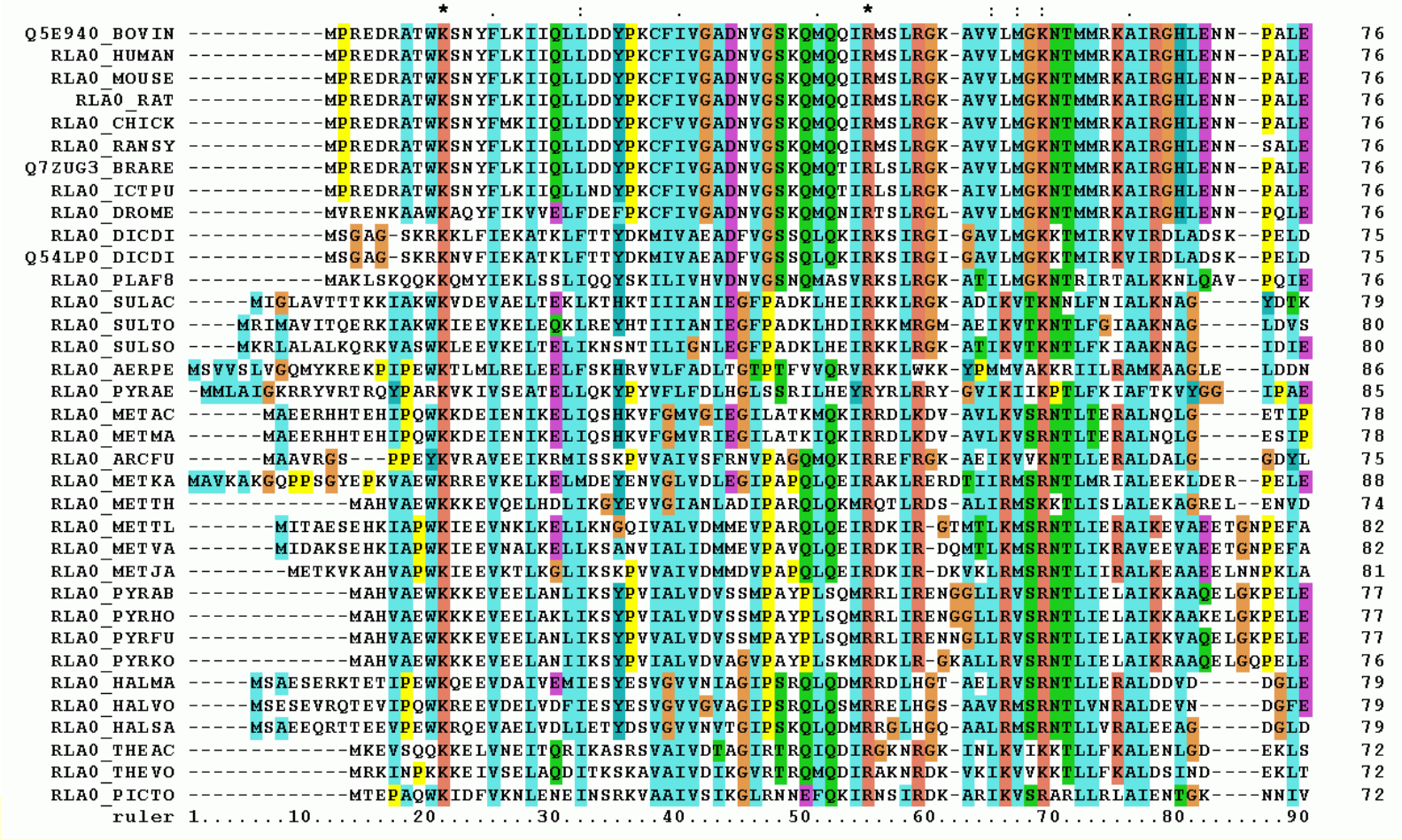
```
C A T G C G A G T A G T A G
C A T G G T A G T A G
C C T G G A G T A C G T A G
C A T G A G C G T A G
```

- sortie : un tableau contenant les k séquences, avec des indels

```
C A T G C G A G T A - G T A G
C A T G - - - G T A - G T A G
C C T G - G A G T A C G T A G
C A T G - - A G - - C G T A G
```

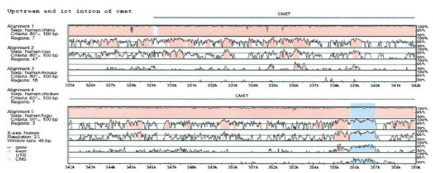
Définition de l'alignement multiple

- Une représentation d'un ensemble de séquences, dans lesquelles les résidus équivalents (d'un point de vue fonctionnel ou structural) sont alignés en colonnes (un site)

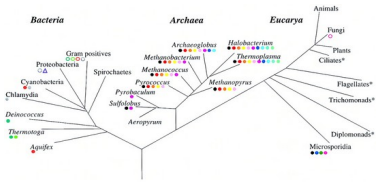


Alignement multiple : Pourquoi ?

Comparative genomics

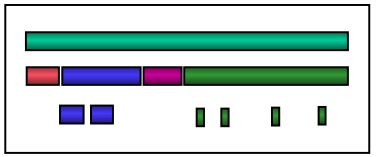


Phylogenetic studies

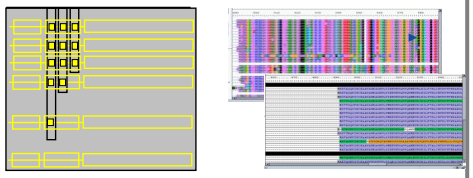


Hierarchical function annotation:

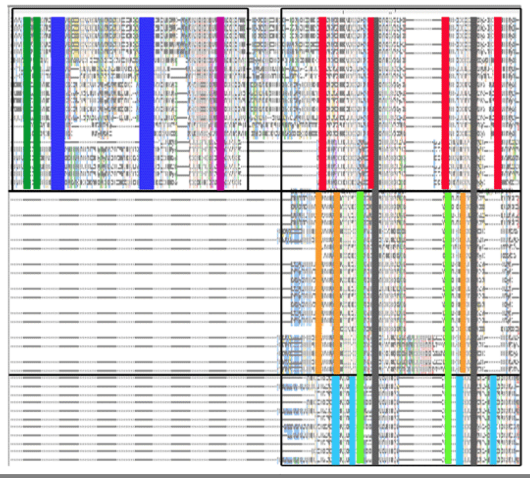
homologs, domains, motifs



Gene identification, validation

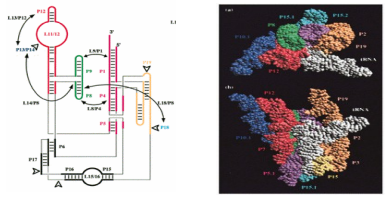


Multiple alignment

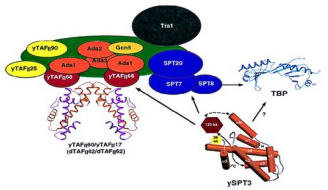


Structure comparison, modelling

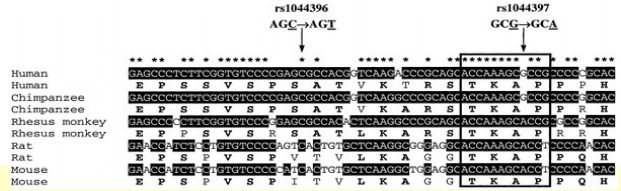
RNA sequence, structure, function



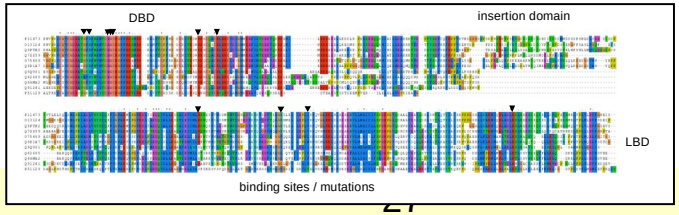
Interaction networks



Human genetics, SNPs



Therapeutics, drug design



Score d'un alignement multiple

- doit rendre compte de la qualité de l'alignement multiple
- habituellement les colonnes sont considérées indépendantes

$$s(x,x) = 1, s(x,y) = -1, s(x,-) = s(-,x) = -2, s(-,-) = 0$$

```
          A  A  C  G  T  A  C  G  A  T  A
          A  -  C  G  T  A  -  A  A  T  G
          G  T  C  G  T  A  -  -  T  T  A
          ------------------------------
(1-2)     1 -2  1  1  1  1 -2 -1  1  1 -1
(1-3)    -1 -1  1  1  1  1 -2 -1 -1  1  1
(2-3)    -1 -2  1  1  1  1  0 -2 -1  1 -1
          =  =  =  =  =  =  =  =  =  =  =
         -1 -5  3  3  3  3 -4 -5 -1  3 -1 = -2
```

Somme des paires

- Définition alternative mais équivalente

$$s(x,x) = 1, s(x,y) = -1, s(x,-) = s(-,x) = -2, s(-,-) = 0$$

```
          A   A   C   G   T   A   C   G   A   T   A
          A   -   C   G   T   A   -   A   A   T   G
          G   T   C   G   T   A   -   -   T   T   A
          ---------------------------------------
(1-2)     1  -2   1   1   1   1  -2  -1   1   1  -1 =  1
(1-3)    -1  -1   1   1   1   1  -2  -1  -1   1   1 =  0
(2-3)    -1  -2   1   1   1   1   0  -2  -1   1  -1 = -3
                                                    =
                                                   -2
```

Les outils les plus populaires

EMBL-EBI Services Research Training About us

Multiple Sequence Alignment

Share Feedback

Tools > Multiple Sequence Alignment

Multiple Sequence Alignment (MSA) is generally the alignment of three or more biological sequences (protein or nucleic acid) of similar length. From the output, homology can be inferred and the evolutionary relationships between the sequences studied.

By contrast, Pairwise Sequence Alignment tools are used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences.

Clustal Omega

New MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments. Suitable for medium-large alignments.

Launch Clustal Omega

Kalign

Very fast MSA tool that concentrates on local regions. Suitable for large alignments.

Launch Kalign

MAFFT

MSA tool that uses Fast Fourier Transforms. Suitable for medium-large alignments.

Launch MAFFT

MUSCLE

Accurate MSA tool, especially good with proteins. Suitable for medium alignments.

Launch MUSCLE

MView

Transform a Sequence Similarity Search result into a Multiple Sequence Alignment or reformat a Multiple Sequence Alignment using the MView program.

Launch MView

T-Coffee

Consistency-based MSA tool that attempts to mitigate the pitfalls of progressive alignment methods. Suitable for small alignments.

Launch T-Coffee

WebPRANK

The EBI has a new phylogeny-aware multiple sequence alignment program which makes use of evolutionary information to help place insertions and deletions.
Try it out at WebPRANK.

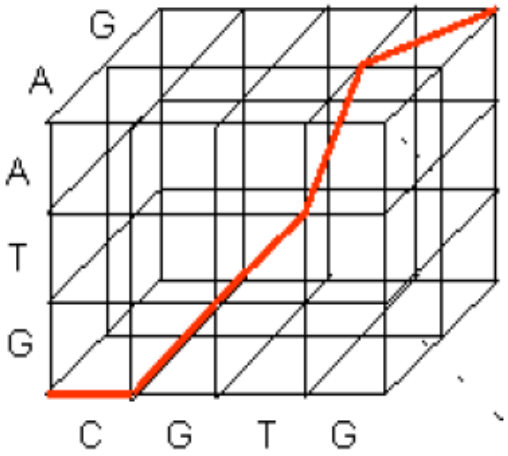
The tools described on this page are provided using **the EMBL-EBI bioinformatics web and programmatic tools framework**. If you have any feedback or encounter any issues please let us know via EMBL-EBI support.

Algorithmes d'alignement multiple

- 3 grandes approches
  - Alignement multiple optimal
  - Alignement multiple progressif
  - Alignement multiple itératif
- Développement de méthodes qui mélangent les approches ou basées sur des approches différentes

Alignement Multiple Optimal

- **Exact**, par programmation dynamique
- Alignement 2 à 2 => chemin dans une matrice de dimension 2
- Alignement multiple de n séquences => chemin dans une matrice de dimension n




```
C  G  T  -  G
-  G  T  A  -
-  -  -  A  G
```

Environ 140 aa

2 Globines => 1 sec

3 Globines => 2 min

4 Globines => 5 hr

5 Globines => 3 semaines

6 Globines => 9 ans

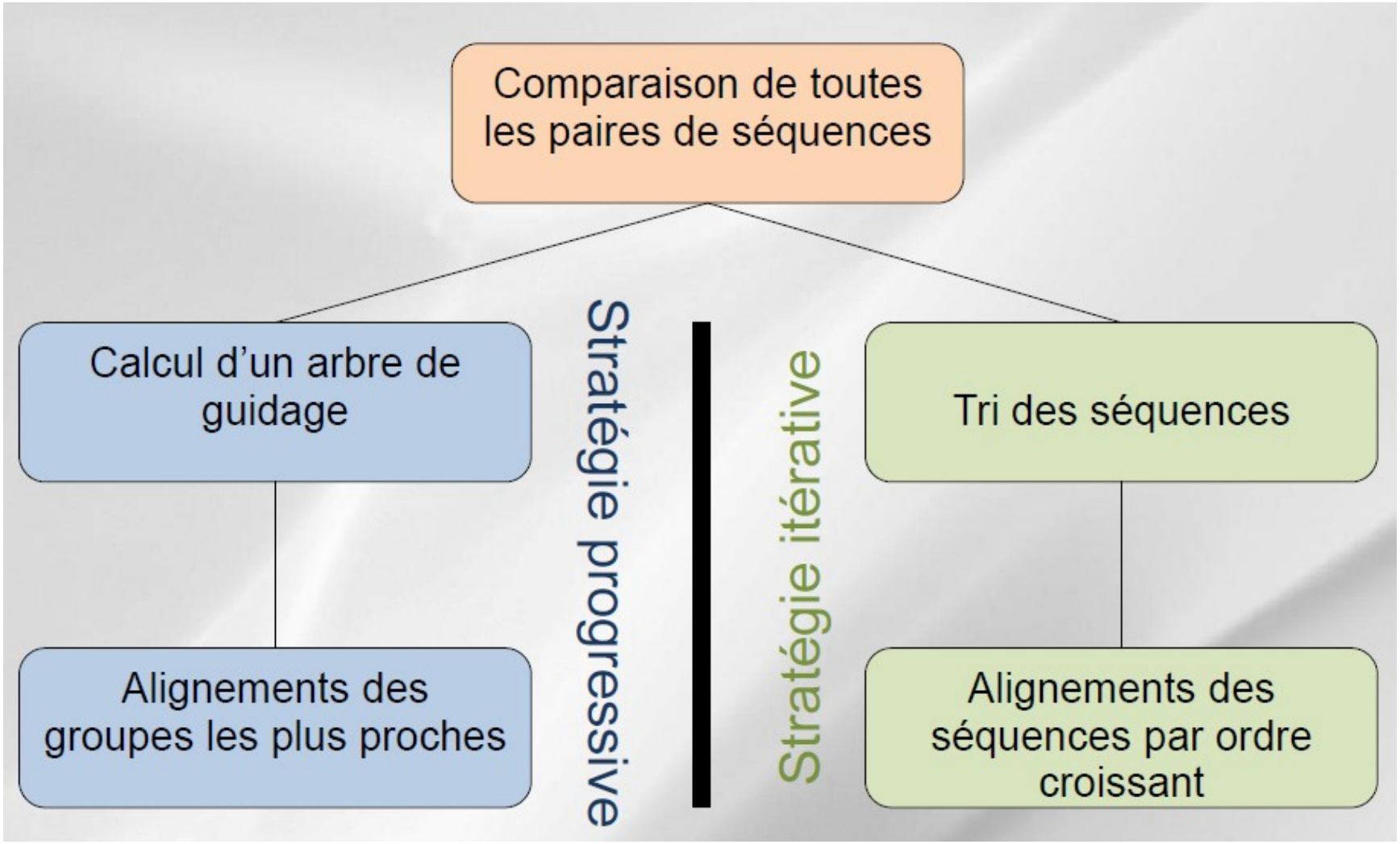
7 Globines => 1000 ans

- Impossible de l'utiliser en pratique => heuristiques

Heuristique : définition

- Algorithme utilisant des règles simples pour diminuer l’espace de recherche des solutions (mais ne donnant pas forcément la meilleure solution)
- Beaucoup de programmes ont été développés => autant d'alignements différents produits

Les grandes approches



Alignement multiple progressif

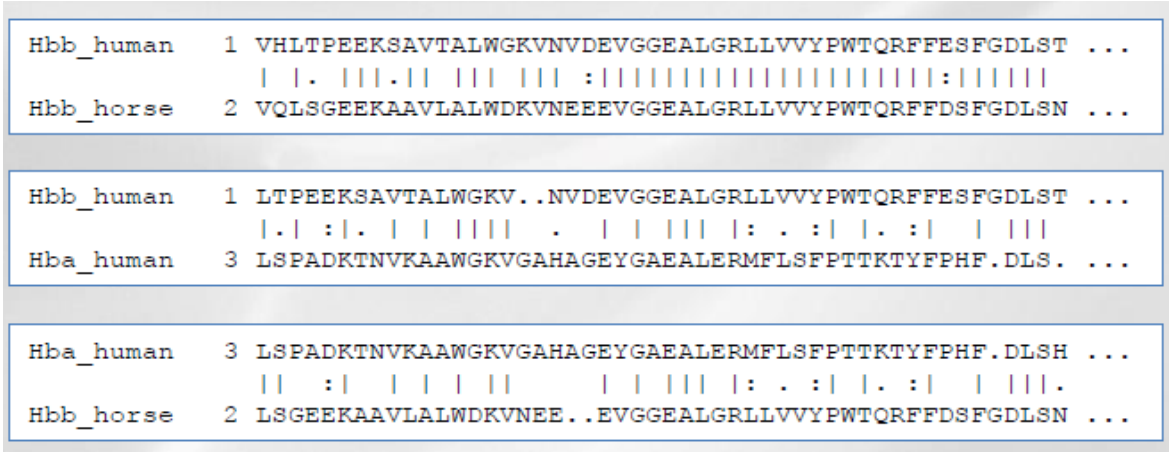
- Évite le calcul de l'ensemble des alignements possibles
- Pas garantie d'obtenir l'alignement optimal
- Principe : Les séquences (ou groupe de séquences) sont alignées progressivement par paires

Alignement multiple progressif

- Problématique :
  - Quelles sont les deux premières séquences à aligner?
  - Dans quel ordre aligner les séquences ?
  - Comment estimer la distance entre deux séquences ?

Alignement multiple progressif

- Étape 1: alignement par paire de toutes les séquences



- Les alignements peuvent être obtenus:
  - Par des méthodes globales ou locales
  - Par programmation dynamiques ou des méthodes heuristiques (non optimales)

Alignement multiple progressif

- Étape 2: construction de la matrice de distance

Dans Clustalw:

Distance entre deux séquences = 1 - $\frac{\text{Nb de résidus identiques}}{\text{Nb de résidus comparés}}$

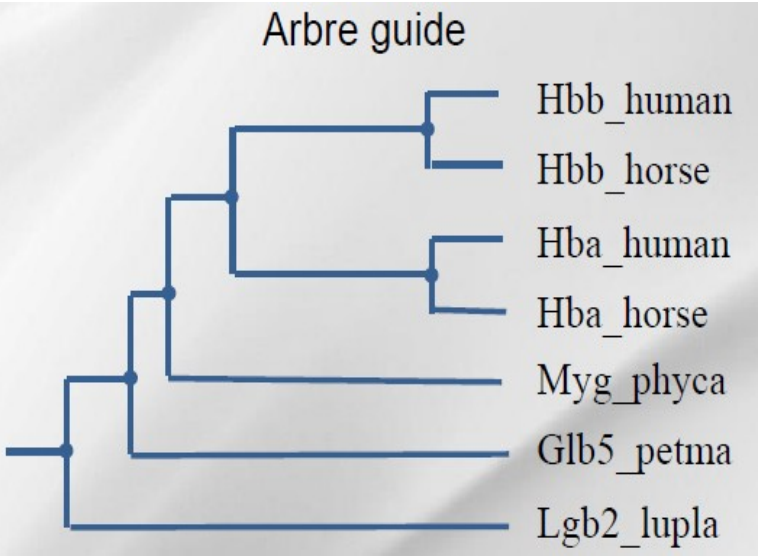
Ex : Hbb_human vs Hbb_horse = 83% identité = distance de 17%

Hbb_human	1	-					
Hbb_horse	2	.17	-				
Hba_human	3	.59	.60	-			
Hba_horse	4	.59	.59	.13	-		
Myg_phyca	5	.77	.77	.75	.75	-	
Glb5_petma	6	.81	.82	.73	.74	.80	-
Lgb2_lupla	7	.87	.86	.86	.88	.93	.90	-
		1	2	3	4	5	6	7

Alignement multiple progressif

- Étape 3: construction de l'arbre guide

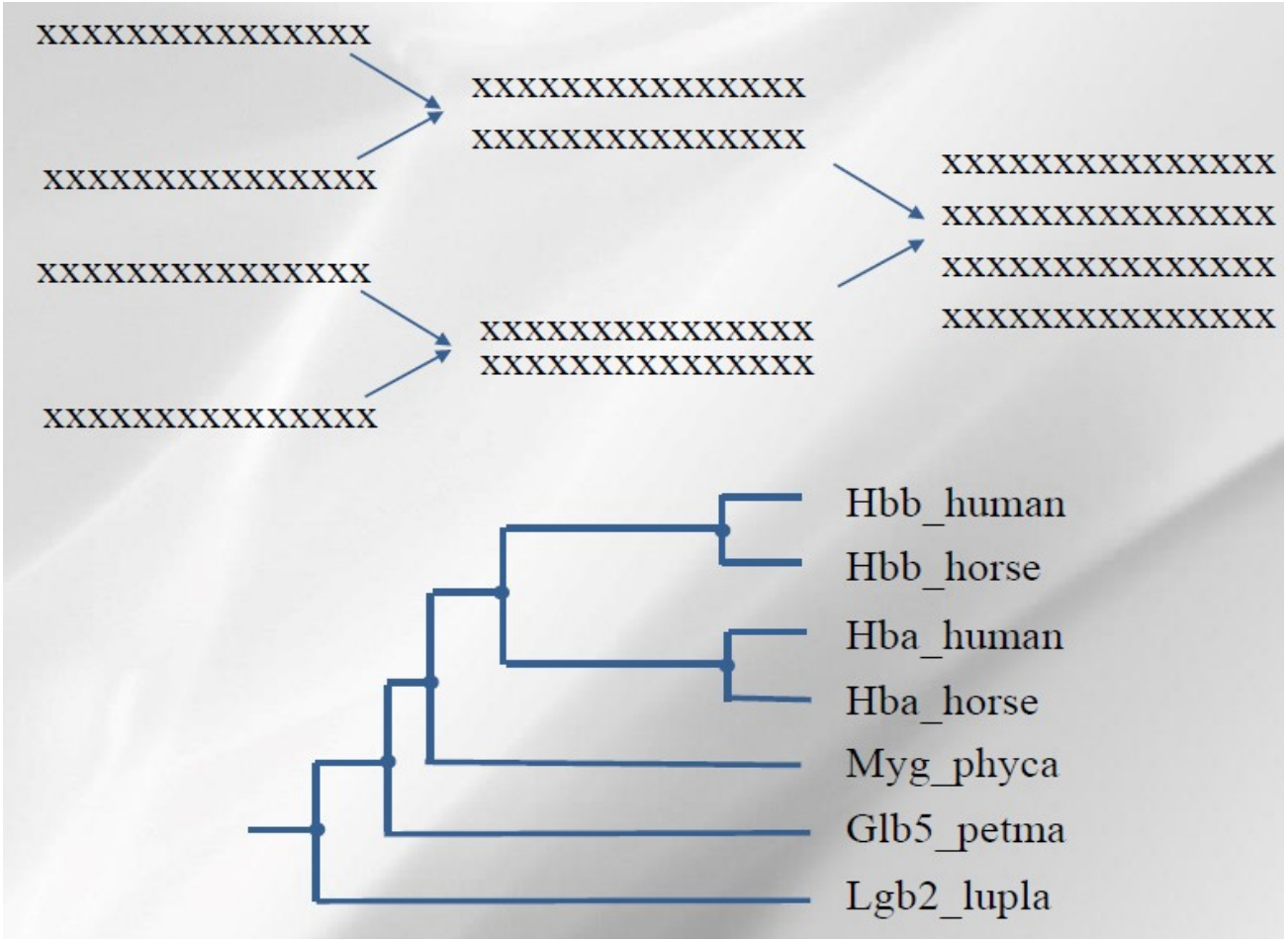
Hbb_human	1	-					
Hbb_horse	2	.17	-				
Hba_human	3	.59	.60	-			
Hba_horse	4	.59	.59	.13	-		
Myg_phyca	5	.77	.77	.75	.75	-	
Glb5_petma	6	.81	.82	.73	.74	.80	-
Lgb2_lupla	7	.87	.86	.86	.88	.93	.90	-
		1	2	3	4	5	6	7



1. Joint les deux séquences les plus proches
2. Calcul à nouveau les distances et joint les deux séquences les plus proches
3. Répétition de l'étape 2 jusqu'à ce que toutes les séquences soient jointes

Alignement multiple progressif

- Étape 4: Alignement progressif selon l'ordre des branches de l'arbre guide



MultAlin

- F. Corpet, 1988
- Principe :

  1- calcule une matrice de similarité des paires

  2- construit un arbre de clustering hiérarchique (UPGMA)

  3- construit l'alignement multiple en suivant l'arbre

  4- reconstruit un arbre de clustering hiérarchique avec les nouveaux alignements paire à paire issus de l'alignement trouvé

  5- réitère le processus jusqu'à stabilisation de l'arbre de clustering

MultAlin : exemple

- Soient 4 séquences à aligner

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

1 - calcul des meilleurs alignements 2 à 2 : scores (Mach = 1, Mismatch =-1, Indel = -1)

2 - construction d'un arbre de clustering :

	①	②	③	④
①	.	6	0	2
②	.	.	0	3
③	.	.	.	4
④	.	.	.	.

MultAlin : exemple

- Soient 4 séquences à aligner

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

1 - calcul des meilleurs alignements 2 à 2 : scores (Mach = 1, Mismatch =-1, Indel = -1)

2 - construction d'un arbre de clustering :

	①	②	③	④
①	.	6	0	2
②	.	.	0	3
③	.	.	.	4
④	.	.	.	.

① TACCATGA
② TACCAT-A

MultAlin : exemple

- Soient 4 séquences à aligner

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

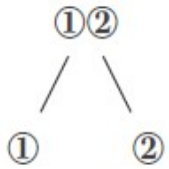
1 - calcul des meilleurs alignements 2 à 2 : scores (Mach = 1, Mismatch =-1, Indel = -1)

2 - construction d'un arbre de clustering :

	①	②	③	④
①	.	6	0	2
②	.	.	0	3
③	.	.	.	4
④	.	.	.	.

	①②	③	④
①②	.	0	2.5
③	.	.	4
④	.	.	.

① TACCATGA
② TACCAT-A



MultAlin : exemple

- Soient 4 séquences à aligner

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

1 - calcul des meilleurs alignements 2 à 2 : scores (Mach = 1, Mismatch =-1, Indel = -1)

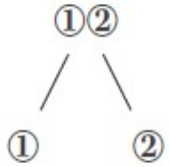
2 - construction d'un arbre de clustering :

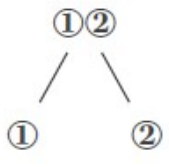
	①	②	③	④
①	.	6	0	2
②	.	.	0	3
③	.	.	.	4
④	.	.	.	.

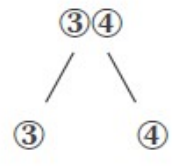
① TACCATGA
② TACCAT-A

	①②	③	④
①②	.	0	2.5
③	.	.	4
④	.	.	.

③ GACGA-C-CA
④ GACCATCTCA







MultAlin : exemple

- Soient 4 séquences à aligner

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

1 - calcul des meilleurs alignements 2 à 2 : scores (Mach = 1, Mismatch =-1, Indel = -1)

2 - construction d'un arbre de clustering :

	①	②	③	④
①	.	6	0	2
②	.	.	0	3
③	.	.	.	4
④	.	.	.	.

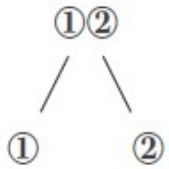
① TACCATGA
② TACCAT-A

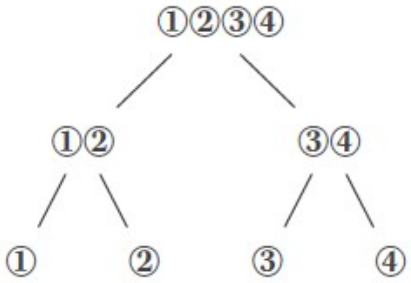
	①②	③	④
①②	.	0	2.5
③	.	.	4
④	.	.	.

③ GACGA-C-CA
④ GACCATCTCA

	① ②	③④
①②	.	1.25
③④	.	.

① TACCAT--GA
② TACCAT---A
③ GACGA-C-CA
④ GACCATCTCA





MultAlin : exemple

- Soient 4 séquences à aligner

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

3- nouvelle matrice des scores et on recommence:

① TACCAT--GA
② TACCAT---A
③ GACGAC--CA
④ GACCATCTCA

ClustalW

- Thompson et al., 1994
- Le plus populaire
- Principe :

  1- calcule une matrice de similarité des paires par programmation dynamique

  2- converti les similarités en distances

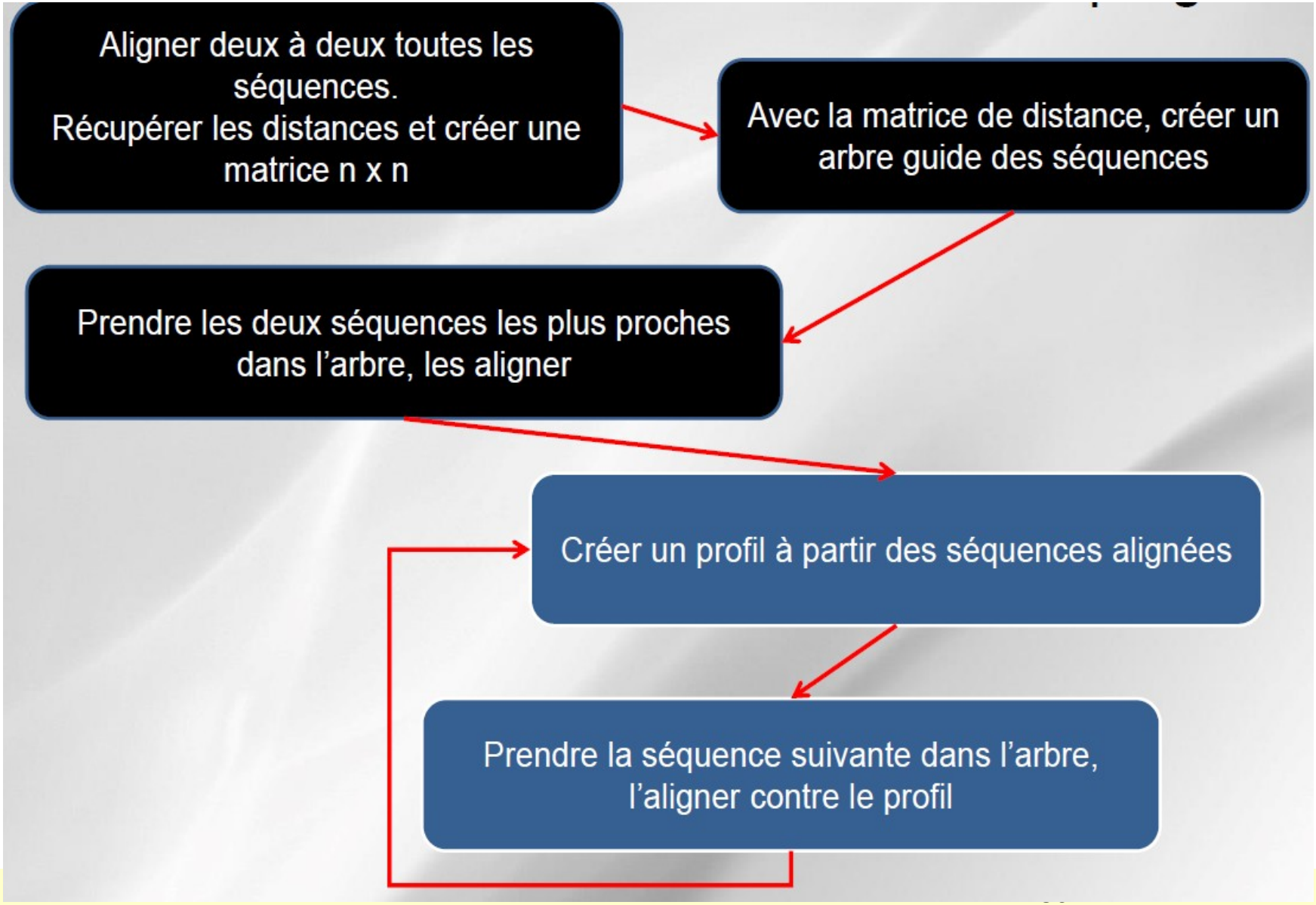
  3- construit l'arbre guide (méthode du Neighbor-Joining)

  4- aligne progressivement les nœuds de l'arbre par ordre décroissant de similarité

ClustalW

- ClustalW utilise les profils
- Les séquences déjà alignées servent de profil pour diriger la suite de l'alignement
- Un profil est représenté sous forme de tableau dans lequel sont données pour chaque position la fréquence observée de chaque lettre
- Chaque nouvelle séquence est alignée contre le profil des séquences déjà alignées

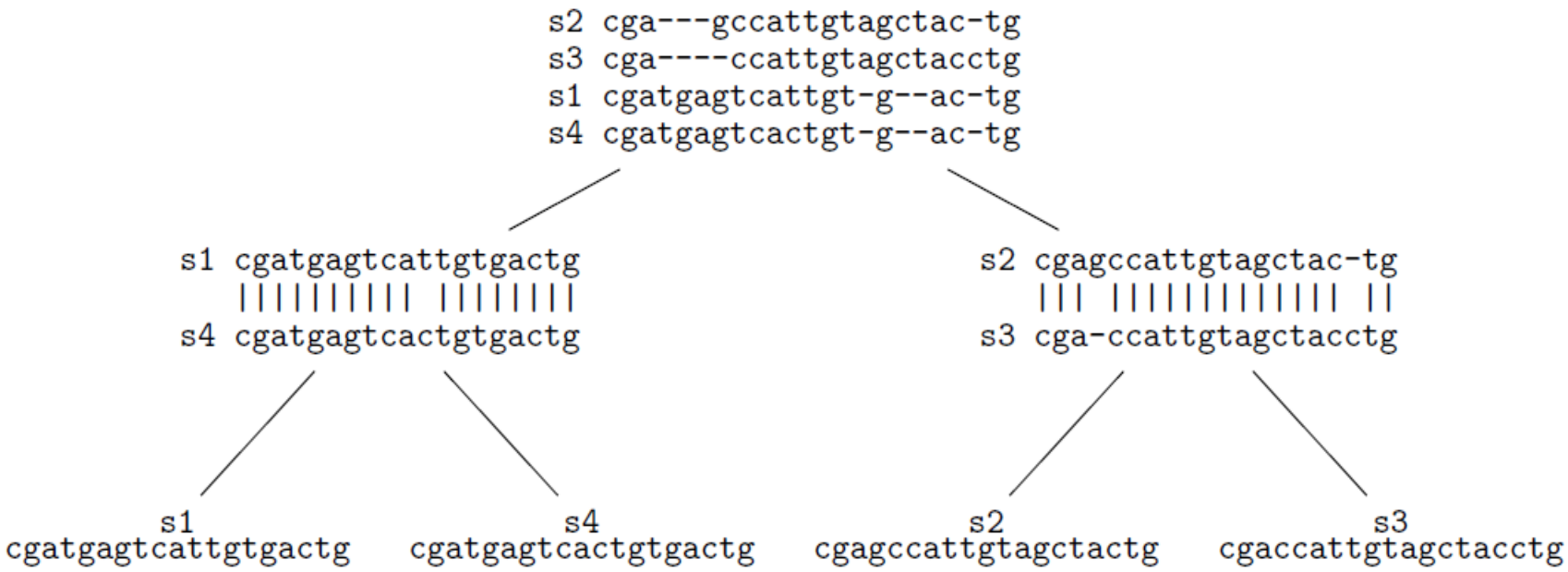
ClustalW


Aligner deux à deux toutes les séquences.
Récupérer les distances et créer une matrice n x n
Avec la matrice de distance, créer un arbre guide des séquences
Prendre les deux séquences les plus proches dans l'arbre, les aligner
Créer un profil à partir des séquences alignées
Prendre la séquence suivante dans l'arbre, l'aligner contre le profil

ClustalW

construction de l'alignement multiple final

"Once a gap, always a gap."



ClustalW est optimisé pour les protéines

- Pondération des séquences en fonction de leur sur- ou sous représentation (éviter que les groupes de séquences proches biaisent l'alignement)
- Adaptation des matrices de similarité au fil de l'algorithme en fonction de la divergence des séquences à aligner
  - BLOSUM 80 pour aligner les séquences proches,
  - BLOSUM 50 pour aligner des séquences distantes, par exemple.
- Pénalités de gaps spécifiques à chaque résidu
  - Par exemple, les glycines sont davantage susceptibles d'avoisiner un gap que les valines.
- Pénalités de gaps réduites dans les régions hydrophiles
  - Encourage la formation de gaps dans des boucles plutôt que dans des régions structurées.
- Pénalités de gaps augmentées dans le voisinage d'autres gaps
  - Évite la formation de petits gaps voisins, au profit de longs gaps

ClustalW : paramètres

- Les scores d'alignement calculés par clustalW utilisent 2 méthodes pour les alignements par paires:

  1- la programmation dynamique qui a pour avantage d'être optimale mais lente (option SLOW)

  => La programmation dynamique sera plutôt utilisée pour des jeux de courtes séquences mais deviendra extrêmement lente pour des jeux de données supérieures à 100 séquences de 1000 acides aminés chacune.

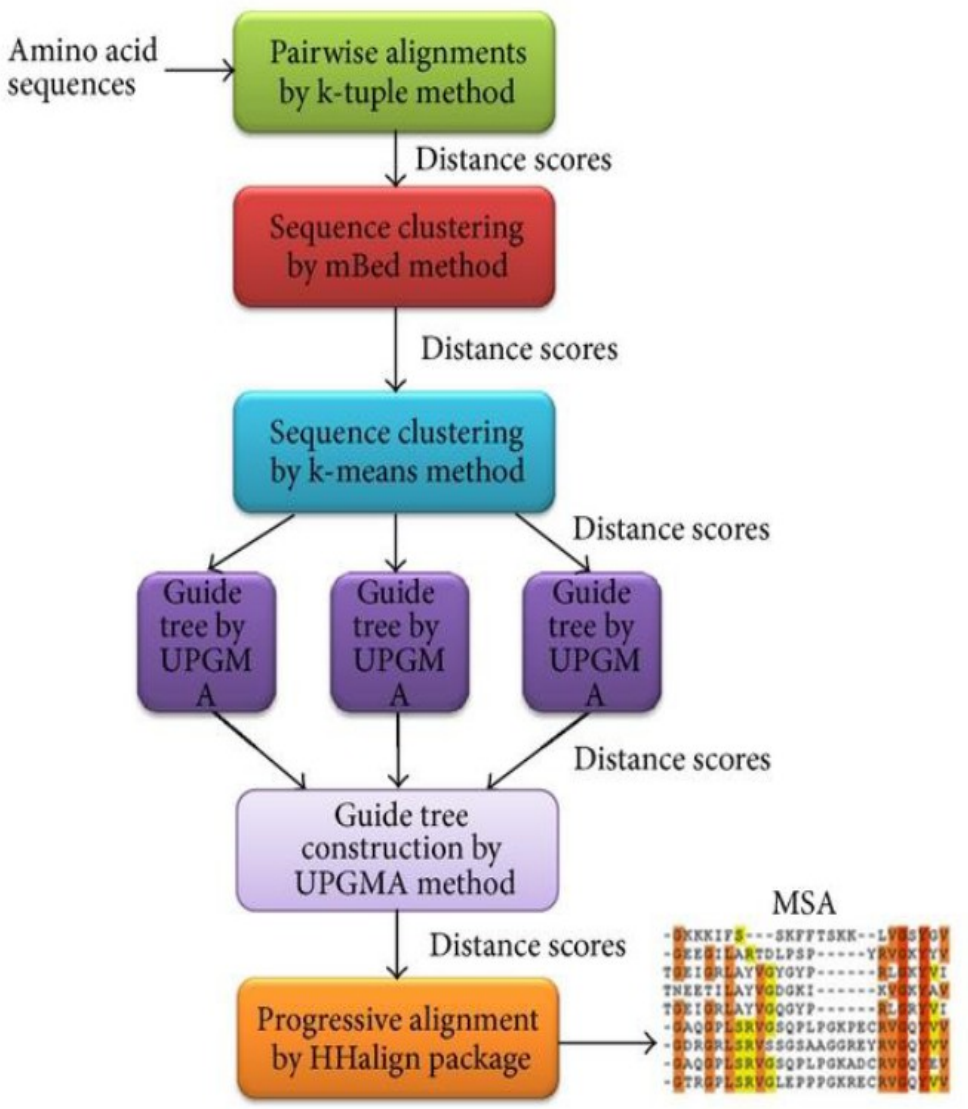
  2- l'algorithme de Wilburet Lipman, qui est très rapide mais plus approximatif (option FAST)

- Le programme n'utilise que les meilleures diagonales, c'est à dire celles présentant le plus de fragments d'appariements exacts.

Conclusions sur ClustalW

- Grand succès de ClustalW
- D’autres programmes fondés sur d’autre algorithmes ou heuristiques donnent souvent de meilleurs résultats dans les cas difficiles
- Si vos séquences sont difficiles à aligner (peu de similarités, longueurs différentes), il est impératif d’essayer d’autres programmes basés sur des algorithmes d'alignement différents (par exemple méthode itérative)

Nouvelle version : Clustal Omega

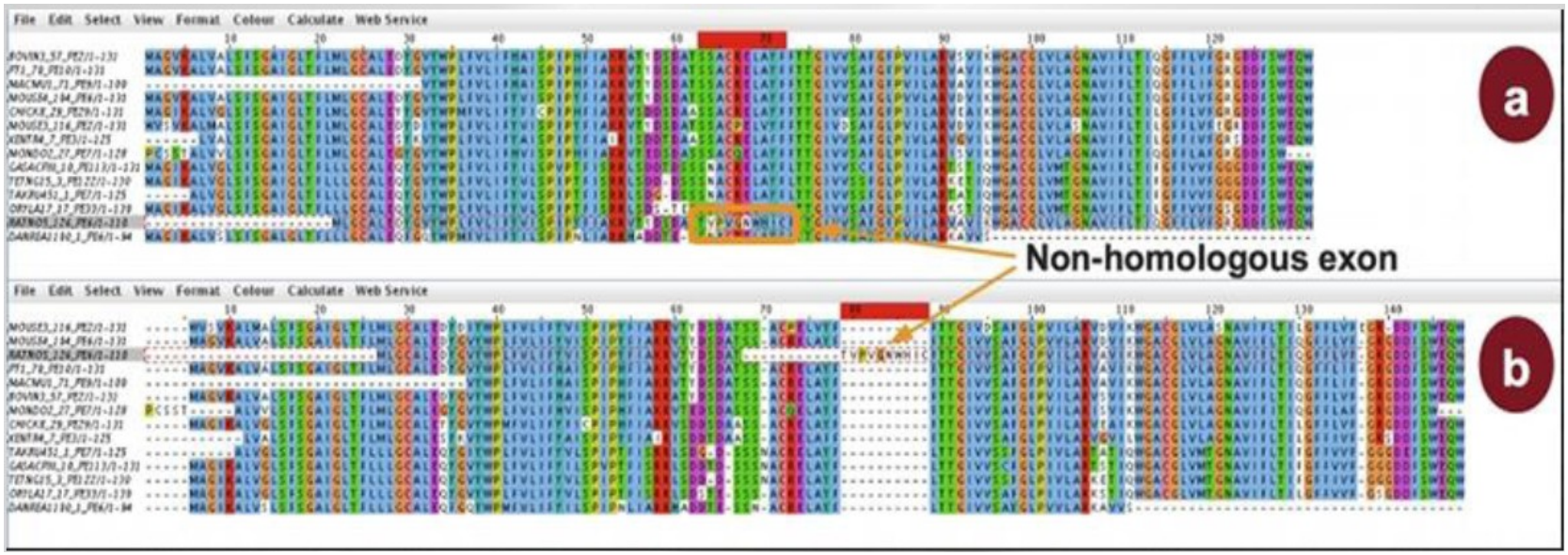


- CLUSTAL Omega utilise des arbres guides avec des graines et des techniques de profile-profile HMM pour générer les alignements
- Changement de plusieurs heuristiques (k-tuple, clustering de séquence avec les méthodes Bed et k-means, méthode UPGMA pour la construction de l'arbre guide suivie d'un alignement progressif avec le package Hhalign pour produire un l'alignement multiple

  => Clustal Omega est précis et permet l'alignement d'un nombre « infini » de séquences

PRANK

- L’exon non homologue ne doit pas être aligné or seules les méthodes fondées sur la phylogénie ne l’alignent pas



(a) alignement des exons non-homologues

(b) alignement basé sur la phylogénie

PRANK

- Les différences avec les autres méthodes d'alignement :
  - PRANK utilise les informations phylogénétiques pour le placement des gap et pour modéliser le processus de substitution
- Quand la phylogénie est correcte, PRANK retourne des alignements de qualité supérieure comparé aux autres méthodes progressives
- Quand la phylogénie est incorrecte, les performances sont très affectées

  => Si votre arbre phylogénétique est faux, l'alignement multiple sera faux

PRANK

- La reconstruction des homologies, incluant le placement correct des indels, est possible pour des séquences provenant d'espèces proches
- PRANK n'est pas utilisable pour l'alignement de séquences protéiques très divergentes
- Si les séquences sont très différentes, l'homologie ne peut pas être construire correctement et PRANK peut refuser de les aligner

Alignement multiple itératif

- Principe

  1- calcul un score de similarité entre toutes les paires de séquences par comparaison des séquences deux à deux; ensemble de scores d'alignement qui sont regroupés dans une matrice de similarités

  2- Cette matrice est utilisée pour **trier les séquences**, généralement des plus similaires aux plus éloignées

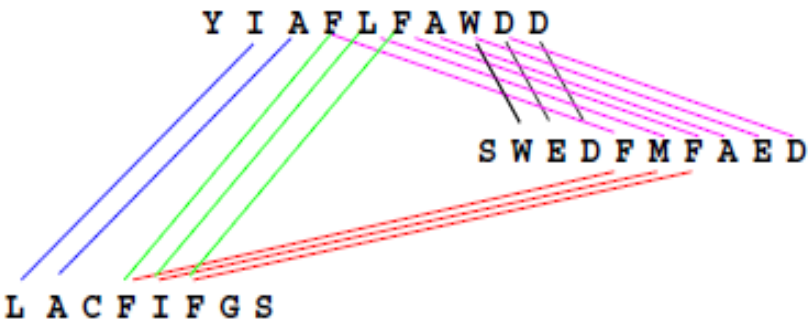
  3- Cette liste est parcourue **itérativement** pour construire l'alignement multiple final (**pas d'arbre guide**) : les deux séquences les plus proches sont alignées (itération 1). A partir de cet alignement, on calcule un «profil» (# une séquence consensus) puis on aligne la troisième séquence avec ce profil (itération 2). Un nouveau profil est calculé avec ces 3 séquences, et la quatrième séquence est alignée...

DIALIGN

- DIALIGN est un programme d'alignement multiple qui repose sur une méthode très différente de celle employée par ClustalW. Il s'agit ici d'un algorithme itératif utilisant une **approche locale** pour calculer les alignements.
- La première étape consiste à comparer toutes les paires de séquences
  - cette étape consiste à rechercher tous les **fragments** (=mots, k-mer) pour ne retenir que ceux qui sont compatibles.Un fragment consiste en une suite (la plus grande possible) de résidus consécutifs, similaires entre deux séquences. Selon cette définition, on constate qu'un fragment ne peut pas contenir d'indels.
- Ensuite, on ne retient que ceux qui sont compatibles, c'est-à-dire des fragments qui ne se croisent pas

DIALIGN

- Etape 1: détection des fragments dans les paires de séquences



- Etape 2: sélection d'un ensemble cohérent de fragments pour construire l'alignement
  - pas de croisements
  - pas de chevauchements
  - score maximal

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

ClustalW vs DIALIGN

Exemple

GARFIELD THE LAST FAT CAT
GARFIELD THE FAT CAT
GARFIELD THE VERY FAST CAT
THE FAT CAT

Alignement fourni par Clustal

```
seq1        GARFIELDTHELASTFA-TCAT
seq2        ----GARFIELDTHEFA-TCAT
seq3        GARFIELDTHEVERYFASTCAT
seq4        ------------THEFA-TCAT
```

Alignement fourni par Dialign2

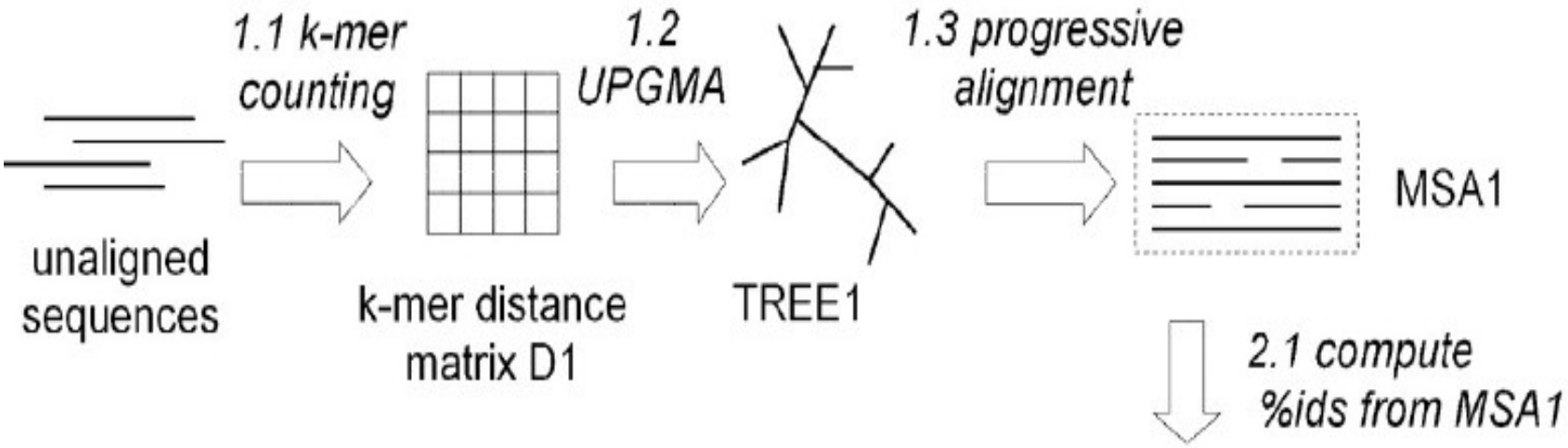
```
seq1        GARFIELD THE LAST FA-T CAT
seq2        GARFIELD THE ---- FA-T CAT
seq3        GARFIELD THE VERY FAST CAT
seq4        -------- THE ---- FA-T CAT
```

Autres approches et approches mixtes

- Développement de nouveaux algorithmes
- Développement de méthodes basées sur des étapes progressives et itératives

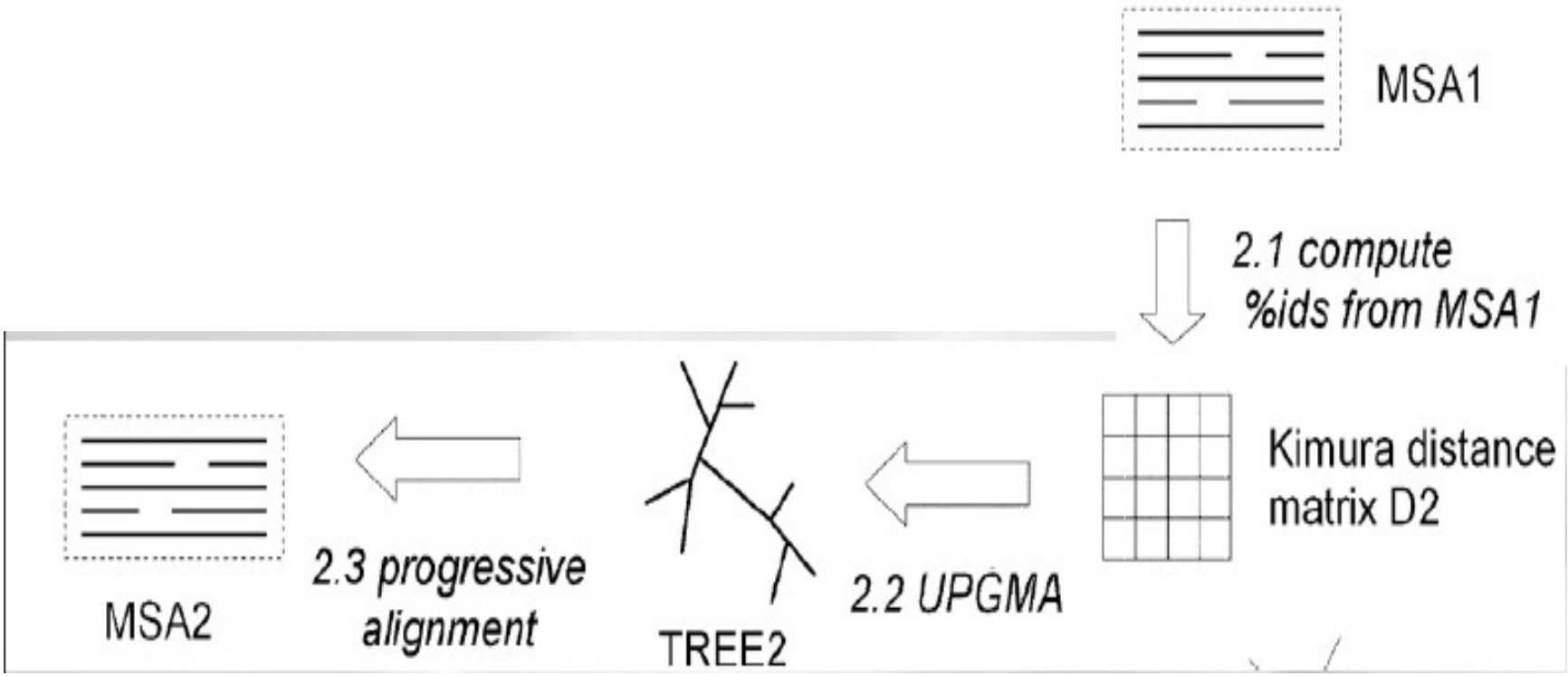
MUSCLE

- multiple sequence alignment by log-expectation



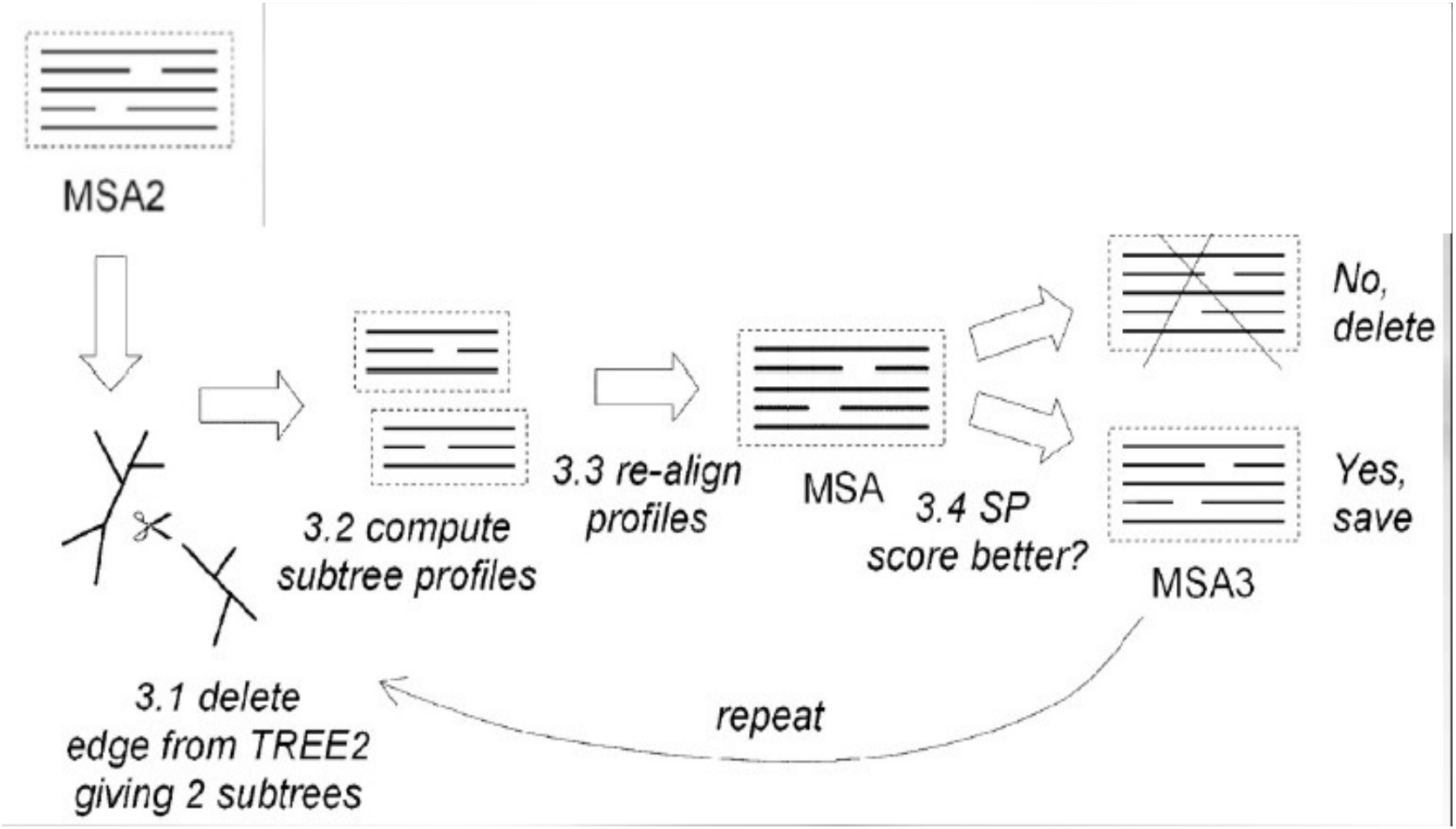
- Un arbre guide est construit à partir de la matrice de distance qui a été calculé par UPGMA ou NJ (NeighborJoinning), et une racine est identifiée.
- Un alignement progressif est construit en suivant l'ordre des branches de l'arbre guide, produisant un alignement multiple de l'ensemble des séquences

MUSCLE



- La deuxième étape tente d'améliorer l'arbre guide et construit un nouvel alignement progressif selon cet arbre
- Cette étape peut être répétée (méthode itérative)

MUSCLE



- Troisième étape : raffinement

MUSCLE

- MUSCLE utilise une méthode très rapide mais parfois plus approximative pour calculer les distances: il compte le nombre de k-mers partagés par 2 séquences, sans construire d'alignement
- Typiquement 3,000 fois plus rapide que la méthode utilisée par ClustalW mais les arbres guides sont généralement moins fiables

T-coffee

- Tree-based Consistency Objective Function for alignment Evaluation (Notredame, et al. JMB,302(205-217),2000)
- Mélange de résultats d'alignement global (ClustalW) et local (Lalign)
- M-Coffee permet de combiner les sorties de plusieurs méthodes d'alignement (Clustal, Mafft, Probcons, Muscle...) en un unique alignement
- T-coffee est capable de combiner la séquence avec :
  - des structures (3D-Coffee/Expresso)
  - des profiles (PSI-Coffee)
  - des structures secondaires d'ARN (R-Coffee)

MAFFT

- MAFFT: Katoh et al, 2002
- MAFFT a été développé dans le but d'accélérer considérablement le processus d'alignement multiple
- Permet d'aligner un grand nombre de séquences sans pour autant sacrifier à la qualité de l'alignement
- Algorithme divisé en 3 grandes étapes

MAFFT

ETAPE 1

- Ré-encodage des séquences : Suite de lettres → une suite de valeurs numériques
  - Chaque acide aminé est décrit par sa polarité et son volume, les séquences sont réécrites dans ce système
  - Les nucléotides sont recodés en utilisant les fréquences locales des quatre bases
- Les segments de similarité entre chaque paire de séquences sont repérés au moyen d'un algorithme de calcul appelé **transformée de Fourier rapide**
- Les paires de séquences sont ensuite alignées sur la base de ces segments de similarité (cf.DIALIGN). Sauf pour des séquences très divergentes, ce procédé permet d'aligner toutes les paires de séquences environ 10 fois plus vite que ClustalW

MAFFT

ETAPE 2 :

- Un arbre de guidage (cf. ClustalW et MUSCLE) est ensuite calculé à partir des alignements précédents
- Le calcul des distances entre les séquences est simplifié et accéléré en recodant les séquences protéiques dans un alphabet réduit à 6 lettres: par exemple, les acides aminés hydrophobes
- La distance entre 2 séquences est estimée à partir du nombre de mots de 6 lettres (k-mers) que ces séquences partagent dans ce nouvel alphabet (cf.MUSCLE)

ETAPE 3 :

- Les séquences sont ensuite alignées progressivement en suivant l'ordre indiqué par l'arbre de guidage

MAFFT

- Plusieurs programmes sont proposés le site Web de MAFFT
- Ce que nous venons de décrire correspond à l'option **FFT-NS-1** (FastFourierTransform-New Scoring matrix-1step)
- Contrairement à ClustalW mais comme MUSCLE, MAFFT peut occasionnellement procéder à un deuxième passage.
- Dans ce dernier, l'alignement réalisé précédemment sert à recalculer la distance entre chaque paire de séquences, un nouvel arbre de guidage et un nouvel alignement multiple

  => Cette option s'appelle **FFT-NS-2**

MAFFT

- De manière similaire à MUSCLE, MAFFT peut procéder à un raffinement de l'alignement
- Dans ce cas, l'arbre de guidage est scindé en deux, puis les deux moitiés sont réalignées
- On recommence ainsi tant que le score d'alignement s'améliore
- On procède alors à un nombre i d'itérations i étant inconnu *a priori*
- Cette option porte le nom de **FFT-NS-i**
- On peut limiter à deux le nombre d'itérations («two cycles only»)

MAFFT

- Il faut par ailleurs noter que le site de MAFFT propose des programmes fondés non pas sur la transformées de Fourier rapide, mais sur l'algorithme de programmation dynamique.
- Ainsi le programme nommé **G-INS-i** aligne les paires de séquences suivant l'algorithme global de Needlman-Wunsch, comme ClustalW, calcule un arbre de guidage, aligne toutes les séquences suivant cet arbre et procède enfin à un raffinement de l'alignement comme décrit ci-avant.
- Les programmes **L-INS-i** et **E-INS-i** procèdent de la même façon, mais avec l'algorithme d'alignement local de Smith-Waterman.
- Bien entendu, ces programmes, nettement plus lents, ne conviennent pas pour un grand nombre de séquences
- Enfin, le programme **Q-INS-i** est spécifiquement dédié à l'alignement de séquences d'ARN

MAFFT : résumé des options

- 1.Mode basique, rapide —juste progressif
  - a)FFTNS1 (fftns--retree1)
  - b)FFTNS2 (fftns) (sameas mafft--retree2)

  OK jusqu'à 1 000 séquences facilement alignables
- 2. Mode intermédiaire —progressif + itérations
  - a)FFTNSI (fftnsi) default twocycles, or e.g. fftnsi--maxiterate1000
  - b)NWNSI (nwnsi) sameas FFTNSI, but no FFT, Needleman-Wunschonly.

  OK entre 100 et 500 séquences
- 3. Mode avancé —progressif + itérations + consistance (cf. T-Coffee)
  - a) EINSI (einsi) Smith-Waterman (plusieurs régions similaires même ordre)
  - b) LINSI (linsi) Smith-Waterman stricte (1 région similaires)
  - c) GINSI (linsi) global Needleman-Wunsch

MACSE

- Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons
- MACSE aligne des séquences nucléiques codantes
  - Préserve la structure en codons
  - Prend en compte la traduction en AA
  - Tolère des délétions non multiples de 3 (frameshifts) et codons stop
- MACSE est la première solution pour aligner des datasets de gènes codant contenant des séquences non-fonctionnelles (pseudogènes) sans perturber la structure des codons
- Très utilisé en phylogénie

Quelle méthode utiliser ?

- cela dépend du type de séquences à aligner
- plus les séquences sont divergentes, moins le résultat est fiable
  - quand le taux d'identité est supérieur à 35%, toutes les méthodes sont satisfaisantes : alignements corrects à plus de 90%
  - twilight zone : 10-20 % identité : **Aucune méthode n'assure un alignement avec plus de 50% de correction**
- Clustal a tendance à autoriser moins de gaps que Dialign2
  - similarité locale : Dialign2
  - similarité globale : Clustal
- Pas de méthode universelle
- Pas de confiance aveugle vis-à-vis du résultat obtenu

Quelle méthode utiliser ?

- BaliBASE : base de données d'alignements multiples et de benchmarks contenant plus de 150 familles de protéines
- alignements basés sur la structure secondaire
  - Référence 1 séquences équidistantes avec différents niveaux de conservation
  - Référence 2 protéines homologues + 1 séquence orpheline
  - Référence 3 sous-groupes avec moins de 25% d'identité entre les groupes
  - Référence 4 extensions N/C-terminales
  - Référence 5 insertions internes
  - Référence 6 répétitions internes
  - Référence 7 protéines transmembranaires
  - Référence 8 permutations de domaines
- Réf. 1, 2 et 3: préférer Clustal à Dialign2
- Réf. 4 et 5: préférer Dialign2 à Clustal

Quelle méthode utiliser ?

Exemple : domaine SH3

SH3 (Src homology 3) domains are often indicative of a protein involved in signal transduction related to cytoskeletal organization. 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.

Prosite PS50002

Séquences à aligner		longueur
1aboA	P00520	57
1ycsB	P04637	60
1pht	P27986	80
1ihvA	P00383	49
1vie	P12497	51

- séquences courtes
- similarité faible ($< 25\%$) et diffuse

Quelle méthode utiliser ?

SH3 - Véritable Alignement

basé sur l'alignement des éléments de structure secondaire

```
1aboA   -NLFVALYDfvasgdntlsitkGEKLRVLgynhn------
1ycsB   kGVIYALWDyepqnddelpmkeGDCMTIIhrede------
1pht    gYQYRALYDykkereedidlhlGDILTVNkgslvalgfsd
1ihvA   -NFRVYYRDsrd------pvwkGPAKLLWkg---------
1vie    -drvrkksga---------awqGQIVGWYctnlt------

1aboA   --------gEWCEAQt--kngqGWVPSNYITPVN------
1ycsB   ------deiEWWWARl--ndkeGYVPRNLLGLYP------
1pht    gqearpeeiGWLNGYnettgerGDFPGTYVEYIGrkkisp
1ihvA   --------eGAVVIQd--nsdiKVVPRRKAKIIRd-----
1vie    -------peGYAVESeahpgsvQIYPVAALERIN------
```

Quelle méthode utiliser ?

SH3 - Alignement fourni par Clustal

```
1aboA   -NLFV-ALYDFVASGDNTLSITKGEKLRV-------LGYNHNG
1ycsB   KGVIY-ALWDYEPQNDDELPMKEGDCMTI-------IHREDED
1pht    -GYQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQ
1ihvA   ------NFRVYYRDSRD--PVWKGPAKLL---------WKGEG
1vie    ---------DRVRKKSG--AAWQGQIVGW---------YCTNL

1aboA   -------EWCEA--QTKNGQGWVPSNYITPVN------
1ycsB   EI-----EWWWA--RLNDKEGYVPRNLLGLYP------
1pht    EARPEEIGWLNGYNETTGERGDFPGTYVEYIGRKKISP
1ihvA   -------AVVIQ---DNSDIKVVPRRKAKIIRD-----
1vie    TP----EGYAVESEAHPGSVQIYPVAALERIN------
```

Quelle méthode utiliser ?

SH3 - Alignement fourni par Dialign2

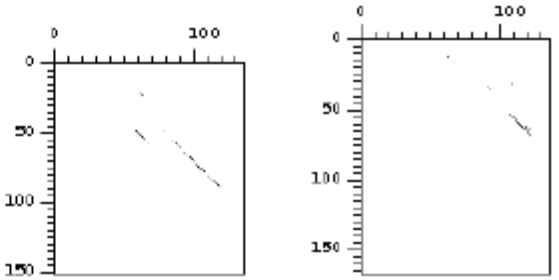
Quelle méthode utiliser ?

Exemple : 5 protéines, domaine HLH

domaine *helix - loop - helix*

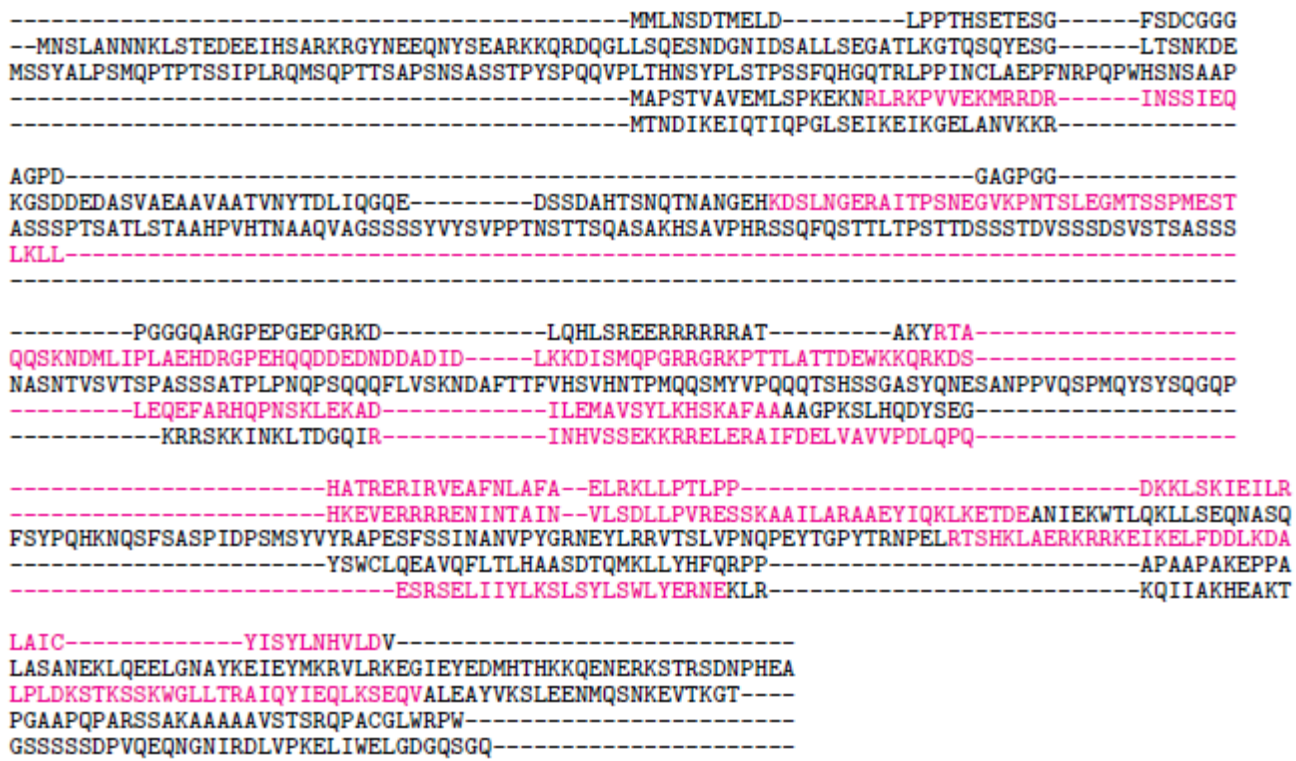
Séquences à aligner:	longueur:
1) HEN1-Human	133
2) CBF1-Yeast	351
3) HES5-Mouse	167
4) INO4-Yeast	151
5) ESC1-Yeast	413

- longueurs dissemblables
- similarité locale



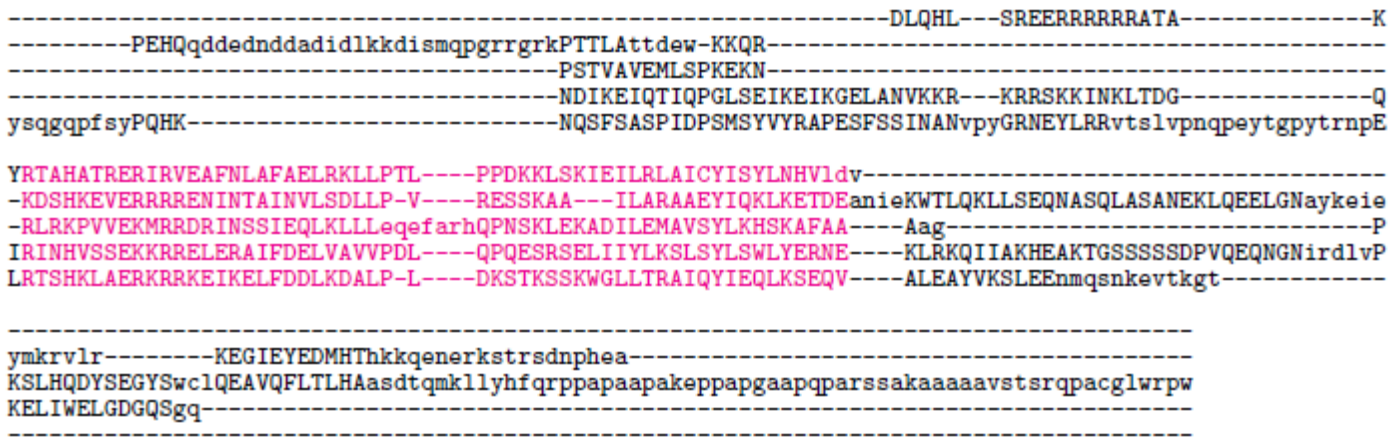
Quelle méthode utiliser ?

helix-loop-helix, alignement Clustal

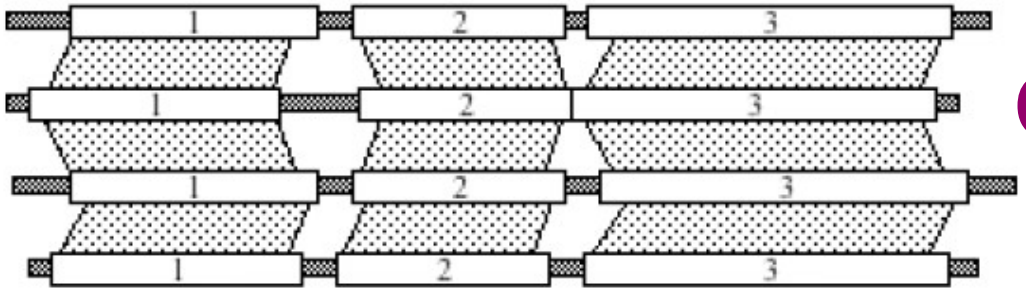


Quelle méthode utiliser ?

helix-loop-helix, alignement Dialign2 (2/2)

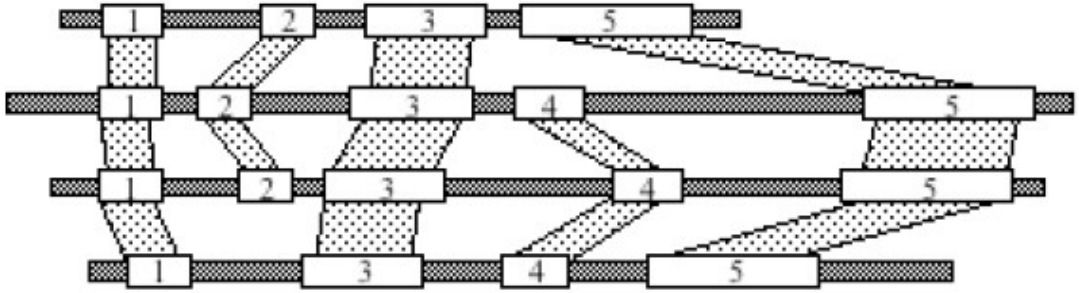


Quelle méthode utiliser ?



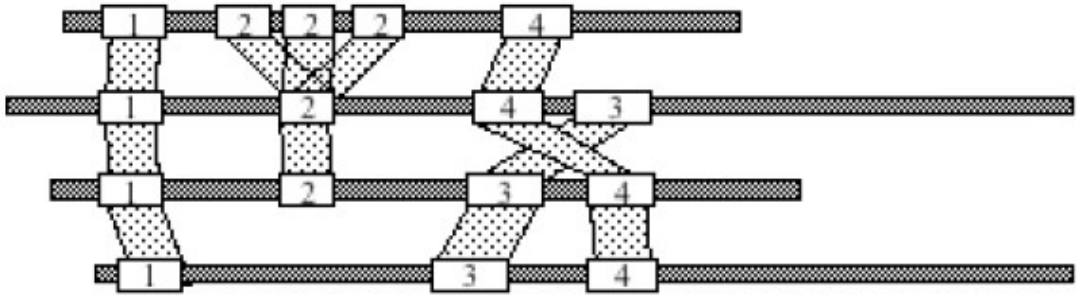
Global

- Muscle
- ClustalW



Blocs

- Dialign
- T-coffee



Local

- MEME

collection d'outils pour l'analyse de motifs biologiques

Quelle méthode utiliser ?

Tableau 5.6 – **Quelques programmes d'alignement multiple.**

	Rapidité	Séquences proches	Séquences éloignées	Qualité
Multalin	++	+++	+	++
Clustal W	+	++	++	+++
Muscle	+++	+++	+	+++
MAFFT	++	++	+	+++
T-Coffee	+	+	+++	+++
DIALIGN	+	+	+++	+

Bioinformatique: Cours et cas pratique

Par Gilbert Deléage,Manolo Gouy

Quelle méthode utiliser ?

- Articles qui comparent les performances des logiciels d'alignement multiple
- Par exemple : Evaluating the Accuracy and Efficiency of Multiple Sequence Alignment Methods, Pervez et al.2014
- Multalin donne généralement de bons résultats même avec des taux importants d'insertions et de délétions sauf dans le cas de séquences longues avec de grands indels
- Clustal Omega et Dialign-TX ne sont pas bons avec de grandes séquences et des grands indels
- T-Coffee, MUSCLE et MAFFT(FFT-NS-2) sont généralement cohérents

Changement d'échelle : comparaison de génomes

- taille de séquences bien plus longue (1000bp → > 1000000bp)
- présence de réarrangements/duplications (combinées)
- autant de programmes qui produisent des alignements différents !
- Différentes écoles :
  - MGA (Bielefeld)
  - MUMmer (Baltimore/Celera genomics)
  - Lagan, Multilagan (Standford)
  - MAUVE (Wisconsin-Madison)
  - GLASS AVID (Berkeley)
  - et bien d'autres ...

Formats de sortie

Formats de sortie

- Chaque logiciel d'alignement propose généralement son propre format de sortie

  => beaucoup de formats de sortie disponibles
- FASTA, Clustal, MSF, PIR, MAF, HSSP ...

Formats de sortie : FASTA

```
>TRY2_RAT/24-239
-----------------------IVGGYTCQENSVPYQVSLNSGY------------HFC
GGSLI------NDQ-WV-VSAAHCYKS----------RIQVRLGE-HNINVLEGN-----
-----EQFVNAAKIIKHPNFDRKT-L------------------------NNDIMLIKLS
SP--VKLNARVATVALPS---SCA---PAGTQCLISGWGN----------TLSSGV----
-------NEPDLLQ-CLDAP-LLPQADCEAS---YPGK--------ITDNMVCVGFL---
-EGG-KDSCQGDSGGPVVCNGE--------LQGIVSWG-YGCALPDN---PGVYTKVCNY
VDWI--------------

>Q16LB2_AEDAE/136-374
-----------------------ILNGIEADLEDFPYLGALALLDNYT-----STVSYRC
GANLI------SDR-FM-LTAAHCLFG--------KQAIHVRMGTLSLTDNPDED-----
----APVIIGVERVFFHRNYTRRPIT-----------------------RNDIALIKLN
RT---VVEDFLIPVCLYT---EQNDP-LPTVPLTIAGWGG----------NDSAS-----
-------LMSSSLM-KASVT-TYERDECNSL---LAKKI-----VRLSNDQLCALGRSEF
NDGLRNDTCVGDSGGPLELSIGR----RKYIVGLTSTG-IVCGNE-F---PSIYTRISQF
IDWI--------------
```

Formats de sortie : clustal

- Format standard très utilisé (extension ".aln")
- La première ligne commence avec le mot "CLUSTAL W" ou "CLUSTALW"
- Une ou plusieurs lignes vides
- Un ou plusieurs blocs de séquences
- Chaque bloc contient :
  - Une ligne pour chaque séquence, pour laquelle :
    - Nom de la séquence, espace , maximum 60 lettres
    - Optionnel : espace suivi d'un comptage des résidus dans la séquence
  - Une ligne montrant le degré de conservation pour les colonnes de l'alignement dans le bloc
  - Une ou plusieurs lignes vides

Formats de sortie : clustal

```
CLUSTAL format alignment by MAFFT (v7.058b)


Dmel_RpS16      MQQKRREPVQAVQVFGRKKTATAVAYCKRGNGLLKVNGRPLEQIEPKVLQYKLQEPLLLL
Pbar_RpS16      MQKKQKEPIHSVQVFGRKKSATAVAYCKRGRGNLRVNGRPLELVEPRVLQYKLQEPILLL
                **:*::**:::********:**********.* *:******* :**:**********:***

Dmel_RpS16      GKEKFAGVDIRVRVSGGGHVAQIYAIRQAISKALVAFYQKYVDEASKKEIKDILVQYDRT
Pbar_RpS16      GKEKFSGVDIRVRVSGGGHVAQIYAIRQAISKALVAYYQKYVDEASKKEVKDILIQYDRT
                *****:**********************************:************:****:*****

Dmel_RpS16      LLVGDPRRCEPKKFGGPGARARYQKSYR
Pbar_RpS16      LLVADPRRCEPKKFGGPGARARYQKSYR
                ***.****************************
```

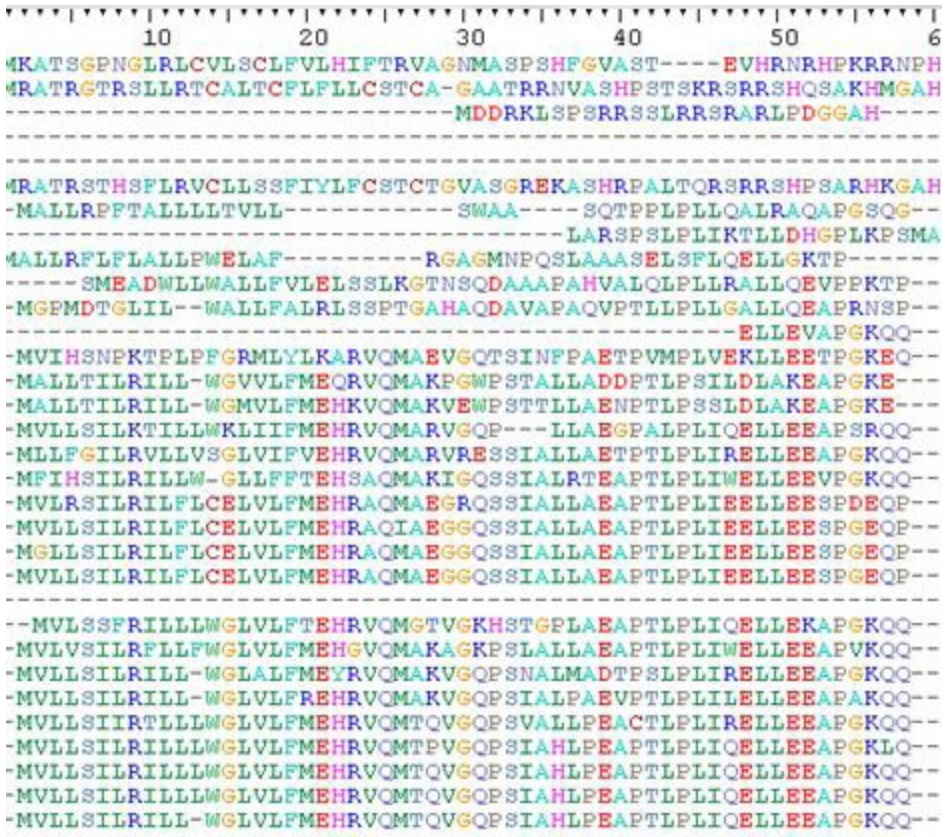
Éditer un alignement multiple

Édition des alignements multiples

- La qualité de l'alignement est essentielle
- Chaque colonne de l'alignement (site) = résidus homologues (nucléotides, acides aminés)
- Parties non «fiables» de l'alignement supprimées de l'analyse phylogénétique
- La plupart des méthodes de reconstruction d'arbres ne tiennent compte que des substitutions, les brèches ou gaps (événements d'insertion/délétion) ne sont pas utilisées

Édition manuelle

- Vérification de l'alignement et comparaison de méthodes (MUSCLE et Clustal par ex.)
- Suppression des sites où il y a au moins un gap
- Suppression des zones trop divergentes (étape difficile à la main)



Edition automatique

- GBLOCKS
- http://molevol.cmima.csic.es/castresana/Gblocks.html
- Le programme élimine les régions avec gaps et les régions trop divergentes
- Conservation de blocs comme on peut le faire à la main mais ici fait de manière reproductible

GUIDANCE : aligneur/éditeur

- GUIDANCE: a web server for assessing alignment confidence scores
- http://guidance.tau.ac.il/
- Guidance prend en entrée des séquences non alignées et le serveur les alignent (ClustalW, MAFFT, MUSCLE, PAGAN et PRANK).
- Il multiplie la production d'alignements multiples et score ensuite la «récurrence» des sites alignés.

Quelques éditeurs

- Un bon éditeur de séquences:
  - Prendre en compte l'affichage coloré de l'alignement : les couleurs aident à la compréhension
  - Reconnaissance des formats d'alignement et maintient de ce format après l'édition
  - Interface graphique ergonomique
- Beaucoup d'éditeurs disponibles : CINEMA (Colour INteractive Editor for Multiple Alignment), GDE (Genetic Data Environment), GeneDoc (pour Windows), SEAVIEW, Jalview, BioEdit (pour Windows uniquement)...
- Certains comme seaview combinent plusieurs fonctionnalités : édition, construction et nettoyage d'alignements (et reconstruction phylogénétique)
- La plupart des logiciels doivent être installés

Alignements multiples : Exercices

Exercices 10, 11 et 12

Modélisation des motifs biologiques

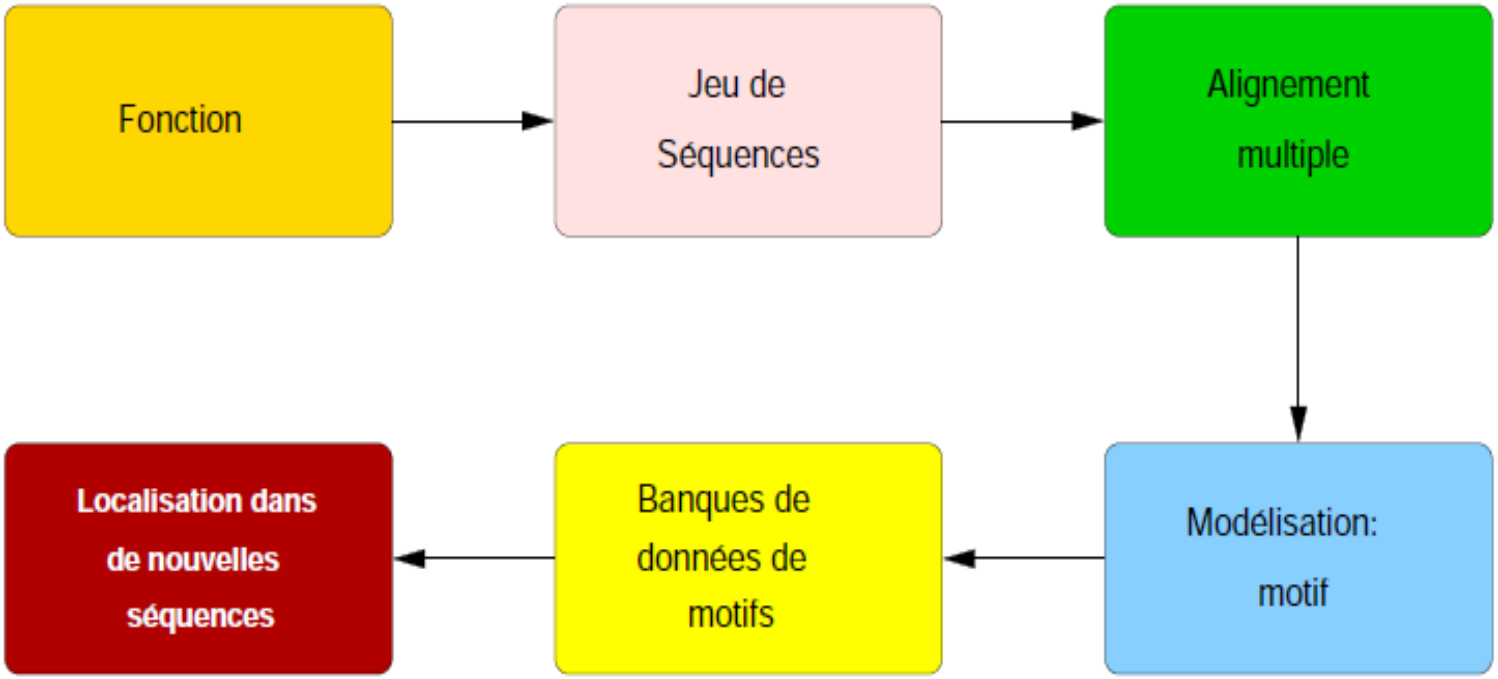
Motifs biologiques : définition

- Une suite non-aléatoire de nucléotides/acides aminés. Il est la conséquence de plusieurs facteurs biologiques :
  - son processus d'évolution (construction et transformation)
  - ses contraintes spatiales
  - sa fonction
- Caractériser ces motifs permet d'étudier les séquences qui leurs sont associées
- Par exemple pour l'ADN, la recherche de gènes dans le génome (début, fin, phase) peut être réalisée en déterminant les motifs associés à ces gènes, comme le codon d'initiation, le codon de terminaison, les signaux de transcription (TATA box, etc.)

Motifs biologiques : définition

- Pour les protéines :
  - domaine protéique: unité structurale (et fonctionnelle) indépendante, évolutivement conservée (doigt de zinc, boucle,...)
  - motifs protéiques: plus courts
    - site de modification post-traductionnelle
    - site de liaison (ADN, métal,...)
    - site actif d'enzyme
  - famille protéique: ensemble de protéines évolutivement reliées; un ou plusieurs domaines protéiques communs

Motifs biologiques : problématique

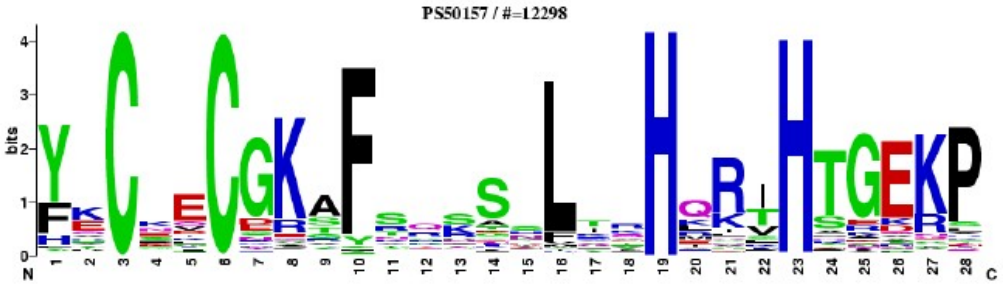


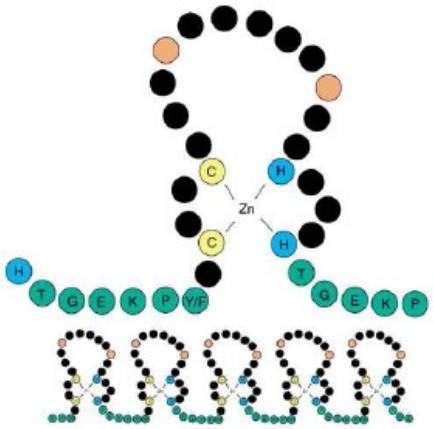
- **Problème 1:** trouver une représentation des motifs à partir des alignements multiples
- **Problème 2:** concevoir des algorithmes pour localiser les occurrences des motifs dans une nouvelle séquence

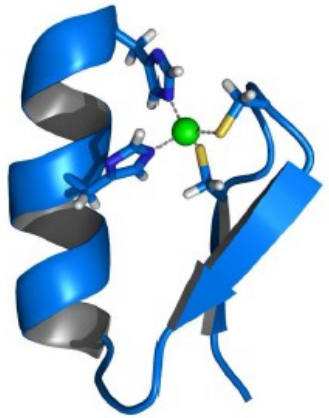
Modélisation des motifs biologiques

Motif doigt de zinc (*C2H2-type*)

```
TYY1_HUMAN YVCPFDGCNKKFAQSTNLKSHILT--H
YKQ8_CAEEL YKCT--VCRKDISSSESLRTHMFKQHH
BASO_HUMAN FQCD--ICKKTFKNACSVKIHHKN-MH
ZG2-9_XENL FVCT--VCGKTYKYKHGLNTHLHS--H
P43_XENBO  LKCSVPGCKRSFRKKRALRIHVSE--H
IKAR_MOUSE FECN--MCGYHSQDRYEFSSHITRGEH
TRA1_CAEEL YKCEFADCEKAFSNASDRAKHQNR-TH
ZN10_HUMAN YKCN--QCGIIFSQNSPFIVHQIA--H
XFIN_XENLA FRCS--ECSRSFTHNSDLTAHMRK--H
TF3A_BUFAM CKCETENCNLAFTTASNMRLHFKR-AH
ZG58_XENLA FVCT--ECNLSFAGLANLRSHQHL--H
P43_XENBO  YRCSYEDCQTVSPTWTALQTHLKK--H
TSH_DROME  FRCV--WCKQSFPTLEALTTHMKDSKH
ZN76_HUMAN FRCGYKGCGRLYTTAHHLKVHERA--H
TF3A_BUFAM YRCPRENCDRTYTTKFNLKSHILT-FH
SUHW_DROAN YACK--ICGKDFTRSYHLKRHQKYSSC
ZN76_HUMAN YTCPEPHCGRGFTSATNYKNHVRI--H
SRYC_DROME FKCN--YCPRDFTNFPNWLKHTRR-RH
EVI1_HUMAN YRCK--YCDRSFSISSNLQRHVRN-IH
```







modélisation : motif Prosite

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

Modélisation des motifs biologiques

- Plusieurs modélisations possibles pour les motifs biologiques :
  - Séquences consensus / Expressions régulières
  - Matrices / Profiles
  - Chaînes de Markov cachées (HMM)

Séquences consensus

Code IUPAC (International Union of Pure and Applied Chemistry)

A	adenine
C	cytosine
G	guanine
T	thymine
U	uracile
R	G A (purine)
Y	T C (pyrimidine)
K	G T (groupe keto)

M	A C (groupe amino)
S	G C (strong)
W	A T (weak)
B	G T C (pas A)
D	G A T (pas C)
H	A C T (pas G)
V	G C A (pas T)
N	A G C T

Séquence consensus : à chaque position de l'alignement multiple, on retient la lettre majoritaire

Séquences consensus

Exemple : modélisation du site de fixation du facteur de transcription *c-Ets-1* chez les murins (15 séquences)

G	C	C	G	G	A	A	G	T	G
A	C	C	G	G	A	A	G	C	A
G	C	C	G	G	A	T	G	T	A
A	C	C	G	G	A	A	G	C	T
A	C	C	G	G	A	T	A	T	A
C	C	C	G	G	A	A	G	T	G
A	C	A	G	G	A	A	G	T	C
G	C	C	G	G	A	T	G	C	A
T	C	C	G	G	A	A	G	T	A
A	C	A	G	G	A	A	G	C	G
A	C	A	G	G	A	T	A	T	G
T	C	C	G	G	A	A	A	C	C
A	C	A	G	G	A	T	A	T	C
C	A	A	G	G	A	C	G	A	C
T	C	T	G	G	A	C	C	C	T
Séquence consensus →	N	C	M	G	G	A	W	G	Y	N

Source :TRANSFAC M00032

Méthode très simple mais très limitée pour modéliser les motifs biologiques

Les expressions régulières

C-C-{P}-{P}-x-C-[STDNEKPI]-x(3)-[LIVMFS]-x(3)-C

```
IGF1B_HUMAN      APQTGIVDECCFRSCDLRRLEMYCAPLKPAKSAR
IGF1_PIG         APQTGIVDECCFRSCDLRRLEMYCAPLKPAKSAR
IGF1_CANFA       APQTGIVDECCFRSCDLRRLEMYCAPLKPAKSAR
IGF2_HORSE       -RSRGIPEECCFRSCDLALLETYCATPAKSERDV
INS_CHIBR        -----IVPQCCTSICTLYQLENYCN--------
INS_ORNAN        -----IWEECCKGVCSMYQLENYCN--------
INS_AOTTR        MQKRGVVDQCCTSICSLYQLQNYCN--------
                      :   :**   * :  *: **
```

{P}	tout sauf P
x	n'importe quel *AA*
x(3)	3 *AA* quelconques
[LIVM]	L, I, V ou M

Les expressions régulières

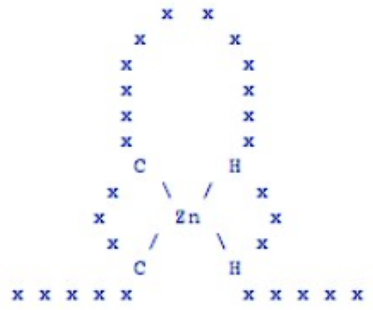
Example: C2H2 Zinc-finger motif

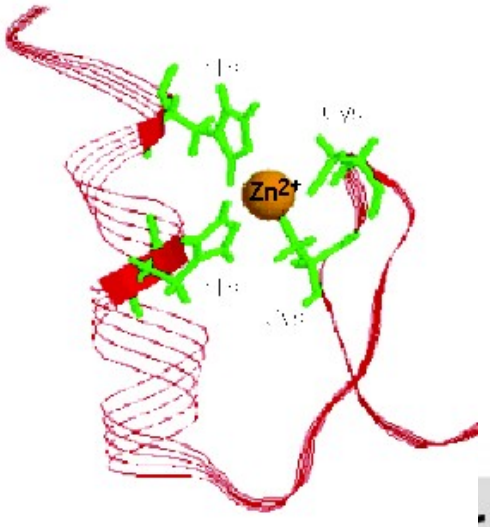
```
                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    665        675        685        695        705        715
Sp1              ACTCPYCKDS EGRGSG---- DPGKKKQHIC HIQGCGKVYG KTSHLRAHLR WHTGERPFMC
Sp2              ACTCPNCKDG EKRS------ GEQGKKKHVC HIPDCGKTFR KTSLLRAHVR LHTGERPFVC
Sp3              ACTCPNCKEG GGRGTN---- -LGKKKQHIC HIPGCGKVYG KTSHLRAHLR WHSGERPFVC
Sp4              ACSCPNCREG EGRGSN---- EPGKKKQHIC HIEGCGKVYG KTSHLRAHLR WHTGERPFIC
DrosBtd          RCTCPNCTNE MSGLPPIVGP DERGRKQHIC HIPGCERLYG KASHLKTHLR WHTGERPFLC
DrosSp           TCDCPNCQEA ERLGPAGV-- HLRKKNIHSC HIPGCGKVYG KTSHLKAHLR WHTGERPFVC
CeT22C8.5        RCTCPNCKAI KHG------- DRGSQHTHLC SVPGCGKTYK KTSHLRAHLR KHTGDRPFVC
Y40B1A.4         PQISLKKKIF FFIFSNFR-- GDGKSRIHIC HL--CNKTYG KTSHLRAHLR GHAGNKPFAC
```

Regular expression (as found in PROSITE) :

C-x(2,4)-C-x(12)-H-x(3)-H

C = cysteine
H = histidine
x(i) = any aa (i times)





Transcription factor Sp1 binding to DNA

http://www.expasy.ch/cgi-bin/prosite-search-ac?PDOC00028

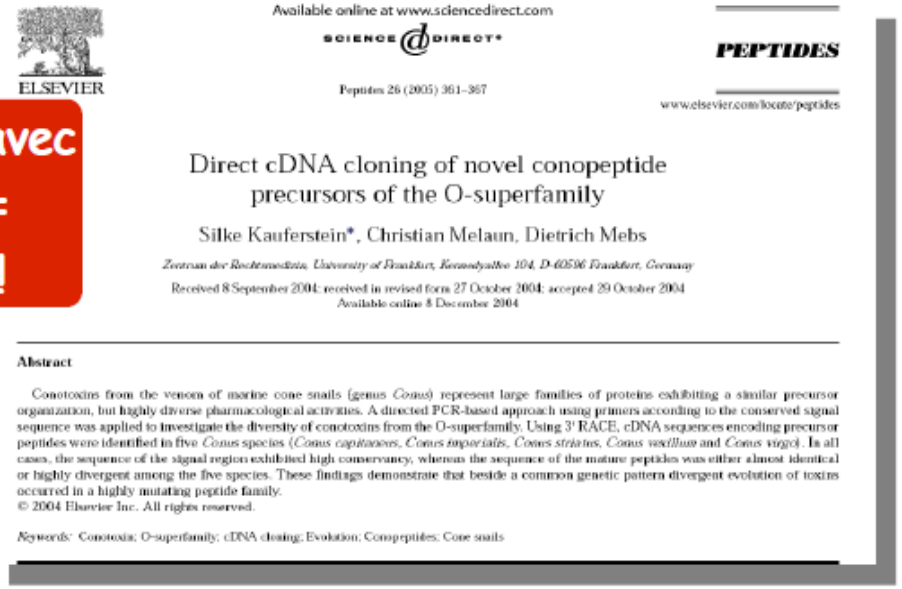
Les expressions régulières

```
C-C-{P}-{P}-x-C-[STDNEKPI]-x(3)-[LIVMFS]-x(3)-C
```

- la protéine suivante **est** détectée:

```
>CXO10_CONCE Conotoxin-10 precursor - Conus capitaneus
MKLTCVLIVVVLFLTACQLITTDDSTGKQRYQAWKLRSKMQNSVLSRLSKRCDEEGT
GCSSDSECCSGRCTPEGLFEFCE
```

aucun rapport avec les insulines= faux-positif!



ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Peptides 26 (2005) 351–367

PEPTIDES

www.elsevier.com/locate/peptides

Direct cDNA cloning of novel conopeptide precursors of the O-superfamily

Silke Kauferstein*, Christian Melaun, Dietrich Mebs

Zentrum der Rechtsmedizin, University of Frankfurt, Kennedyallee 104, D-60596 Frankfurt, Germany

Received 8 September 2004; received in revised form 27 October 2004; accepted 29 October 2004
Available online 8 December 2004

Abstract

Conotoxins from the venom of marine cone snails (genus *Conus*) represent large families of proteins exhibiting a similar precursor organization, but highly diverse pharmacological activities. A directed PCR-based approach using primers according to the conserved signal sequence was applied to investigate the diversity of conotoxins from the O-superfamily. Using 3′ RACE, cDNA sequences encoding precursor peptides were identified in five *Conus* species (*Conus capitaneus*, *Conus imperialis*, *Conus striatus*, *Conus vexillum* and *Conus virgo*). In all cases, the sequence of the signal region exhibited high conservancy, whereas the sequence of the mature peptides was either almost identical or highly divergent among the five species. These findings demonstrate that beside a common genetic pattern divergent evolution of toxins occurred in a highly mutating peptide family.

Keywords: Conotoxin; O-superfamily; cDNA cloning; Evolution; Conopeptides; Cone snails

Les expressions régulières

C-C-{P}-{P}-x-C-[STDNEKPI]-x(3)-[LIVMFS]-x(3)-C

- la protéine suivante **n'est pas** détectée:

>BXB11_BOMMO Bombyxin B-11 precursor (BBX-B11) Bombyx mori (Silk moth).
MMKTAVMFILVVVISLTYSSEEQEVARTYCGRHLANILAYVCFGVEKRGGAQYAPYWQ
ETYLRSRKGPGVVDE**CCFRPCKLEVLKSFF**FFFCD

Faux-négatif!

J. Mol. Biol. (1996) 259, 926–937

JMB

Multiple Gene Copies for Bombyxin, an Insulin-related Peptide of the Silkmoth *Bombyx mori*: Structural Signs for Gene Rearrangement and Duplication Responsible for Generation of Multiple Molecular Forms of Bombyxin

Hidehiko Kondo[1], Masaya Ino[2], Akinori Suzuki[3], Hironori Ishizaki[1] and Masafumi Iwami[1,2*]

Les expressions régulières

- Qu'est-ce qu'une bonne expression régulière ?

Suffisamment tolérant	limiter le nombre de *faux positifs*
pas de sur-adaptation limiter le nombre de *faux négatifs*	Suffisamment discriminant

- Construction d'une expression régulière :
  - À la main
  - PRATT
- Détection des séquences correspondant à l'expression régulière
  - Avec un automate très simple

Les matrices de fréquences

```
IGF1B_HUMAN    APQTGIVDECCFRSCDLRRLEMYCAPLKPAKSAR
IGF1_PIG       APQTGIVDECCFRSCDLRRLEMYCAPLKPAKSAR
IGF1_CANFA     APQTGIVDECCFRSCDLRRLEMYCAPLKPAKSAR
IGF2_HORSE     -RSRGIVEECCFRSCDLALLETYCATPAKSERDV
INS_CHIBR      -----IVDQCCTSICTLYQLENYCN---------
INS_ORNAN      -----IVEECCKGVCSMYQLENYCN---------
INS_AOTTR      MQKRGVVDQCCTSICSLYQLQNYCN---------
```

il faut prendre en compte cette répartition

expression régulière: **[LIVMFS]**, mais en fait:

- L : 6 fois sur 7 (86%)
- M : 1 fois sur 7 (14%)

matrice de fréquence

Les matrices de fréquences (PFM)

- Lignes => positions de l'alignement
- Colonnes => les acides nucléiques/acides aminés

Exemple : *c-Ets-1*

GCCGGAAGTG
ACCGGAAGCA
GCCGGATGTA
ACCGGAAGCT
ACCGGATATA
CCCGGAAGTG
ACAGGAAGTC
GCCGGATGCA
TCCGGAAGTA
ACAGGAAGCG
ACAGGATATG
TCCGGAAACC
ACAGGATATC
CAAGGACGAC
TCTGGACCCT

A	C	G	T	
7	2	3	3	N
1	14	0	0	C
5	9	0	1	M
0	0	15	0	G
0	0	15	0	G
15	0	0	0	A
8	2	0	5	W
4	1	10	0	G
1	6	0	8	Y
5	4	4	2	N

→

Position Frequency Matrix

0.47	0.13	0.2	0.2
0.07	0.93	0	0
0.33	0.6	0	0.07
0	0	1	0
0	0	1	0
1	0	0	0
0.53	0.13	0	0.33
0.27	0.07	0.67	0
0.07	0.4	0	0.53
0.33	0.27	0.27	0.13

Des PFM aux PWM

- PWM : Position Weight Matrix
- Poids positif : les bases qui apparaissent plus que la moyenne
- Poids négatif : les bases qui apparaissent moins que la moyenne
- Poids de la base x dans une colonne de l'alignement :

$$\log_2\left(\frac{f(x)}{0.25}\right)$$

  - f(x) est la fréquence de x dans la colonne considérée
  - 0,25 suppose que les quatre bases sont la même probabilité d'apparition
- Le problème des 0 : ajout d'un pseudo-compte pour éviter qu'il y ait sur-adaptation

$$\log_2\left(\frac{f(x)+0.05}{0.25}\right)$$

Des PFM aux PWM

Suite de l'exemple :

Position Frequency Matrix

0.91	-0.94	0.2	0.2
0.07	0.93	0	0
0.33	0.6	0	0.07
0	0	1	0
0	0	1	0
1	0	0	0
0.53	0.13	0	0.33
0.27	0.07	0.67	0
0.07	0.4	0	0.53
0.33	0.27	0.27	0.13

$\longrightarrow$

Position Weight Matrix

0.91	-0.94	-0.32	-0.32	T
-1.8	1.9	-2.3	-2.3	A
0.4	1.26	-2.3	-1.8	C
-2.3	-2.3	2	-2.3	G
-2.3	-2.3	2	-2.3	G
2	-2.3	-2.3	-2.3	A
1.1	-0.94	-2.3	0.4	T
0.11	0.07	1.42	-2.3	A
-1.8	0.4	0	1.1	C
0.4	0.11	0.11	-0.94	G

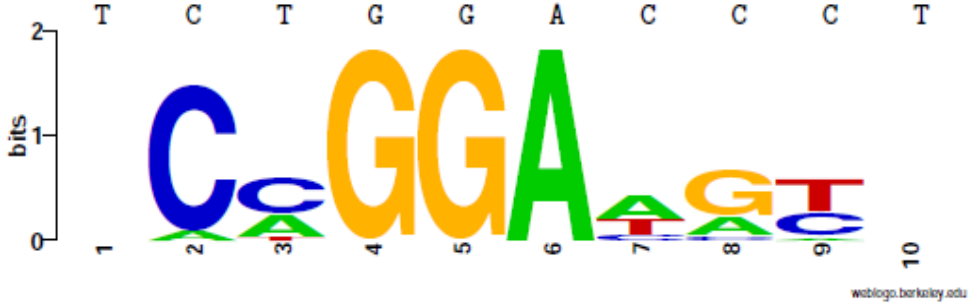
Score d'un mot dans le modèle: somme des poids

T A C G G A T A C G $\rightarrow$ 6.16

WebLogo

Exemple pour le site de C-ets-1:

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



http://weblogo.berkeley.edu/logo.cgi

Les profiles

- Point de départ : matrice des positions 20 colonnes

1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	2	0	0	0	0	0
0	0	0	0	7	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	3	4	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	7	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
3	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	3	1	0	0	0	0	0	0	0	0	0
0	0	0	1	0	1	0	0	0	0	2	0	0	0	0	0	2	2	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0
1	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	1	0	0	0	0	0	0	0	0	4	0	0
3	0	0	1	0	0	0	0	0	0	1	1	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0
3	0	0	1	2	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	6	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2	5	0	0	0	0	0	0	0	0	0
0	0	0	0	2	0	0	3	0	0	0	0	0	1	0	0	2	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0
0	0	0	0	0	1	0	0	3	0	0	2	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	
2	0	0	0	0	0	0	0	0	0	0	1	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	2	0	0	4	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	4	0	0	1	1	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0
0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	4	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0

Les profiles

- Apporter plus de souplesse . . .
- autoriser des insertions, des délétions
  - Ajouter des pénalités particulières pour l'insertion ou la délétion d'un acide aminé → colonne supplémentaire
- autoriser des substitutions (entre acides aminés voisins)
  - M(j; k) : similarité entre les deux acides aminées j et k

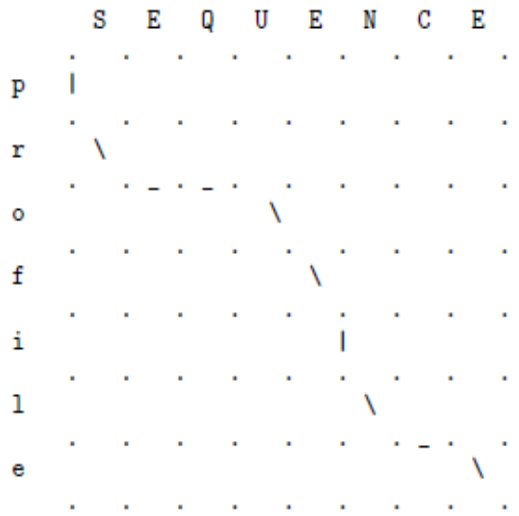
    (PAM, BLOSUM, . . .)
- f_{ik} : fréquence du kième l'acide aminé à la position i

$$\texttt{Profil}_{ij} = \sum_{k=1}^{20} f_{ik} M(j,k)$$

Les profiles

5.2	0	0.8	2.1	...		...	4.55
2.8	0	-5.3	9.1	...		...	4.55
2.6	0	-5.7	10	...	5.2	...	4.55
1.7	0	0.3	0.3	...	$=$	...	4.55
1.7	0	0.3	0.3	...	$\frac{1}{9}M(A,A)+\frac{1}{9}M(A,D)+\frac{5}{9}M(A,P)+\frac{2}{9}M(A,V)$	...	4.55
-1.7	0	-8.2	-5.6	...		...	4.55
4.7	0	0.6	1.7	...		...	4.55
-0.0	0	-3.3	2.8	...		...	4.55
-3.2	0	8.8	-5.6	...		...	4.55
0.3	0	1.7	-2.8	...		...	4.55
5.6	0	-1.1	3.3	...		...	4.55
6.9	0	-1.2	6.4	...		...	4.55
-0.2	0	-4.7	-4	...		...	4.55
-0.8	0	-3.7	2.7	...		...	4.55
0.7	0	-3.7	5.6	...		...	4.55
-3	0	10	-5	...		...	4.55
-1.2	0	-0.2	-3.7	...		...	4.55
4.6	0	-2.2	4.1	...		...	4.55
0	0	-3.9	-2.7	...		...	4.55
0.6	0	0	-2.6	...		...	4.55
3.4	0	1.7	1.6	...		...	4.55
-3.0	0	-3	0	...		...	4.55
2.2	0	-2.4	4	...		...	4.55
-3.0	0	-3	0	...		...	4.55
-3.4	0	7.6	-6.1	...		colonne des indels →	4.55

Recherche avec un profil



- Score : alignement entre le profil et la séquence

  Les pénalités de substitutions et de gaps sont données par le profil

- Seuil d'admission : **E-value**

Modélisation avec les HMM

- HMM = Hidden Markov Model = Modèle de Markov caché
- un ensemble d'états
- des probabilités de transitions entre les états
- un ensemble d'observations
- une probabilité d'émission qui indique pour chaque état la probabilité d'y émettre une telle information

Modélisation avec les HMM

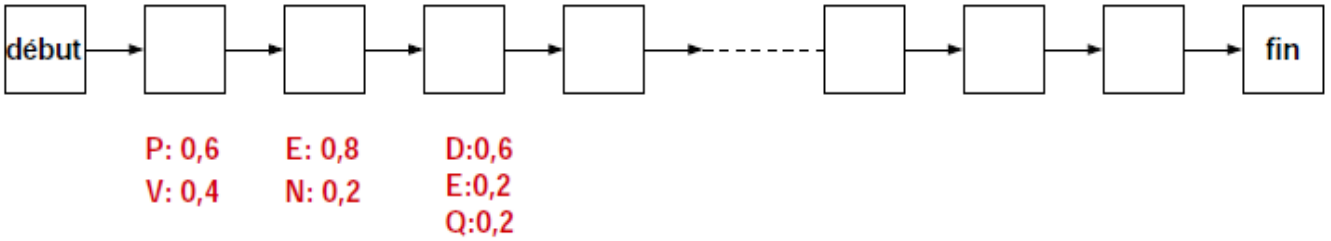
1. Si l'alignement n'a pas d'indels

```
PPY_LOPAM/1-36     PEDWASYQAAVRHYVNLITRQRY
PAHO_BOVIN/30-65   PEQMAQYAAELRRYINMLTRPRY
PAHO_CHICK/26-61   VEDLIRFYNDLQQYLNVVTRHRY
PAHO_ANSAN/1-36    VEDLRFYYDNLQQYRLNVFRHRY
NPF_HELAS/4-39     PNELRQYLKELNEYYAIMGRTRF
```

1 observation = 1 acide aminé

1 état = 1 colonne de l'alignement multiple

émissions = fréquences de chaque a.a.

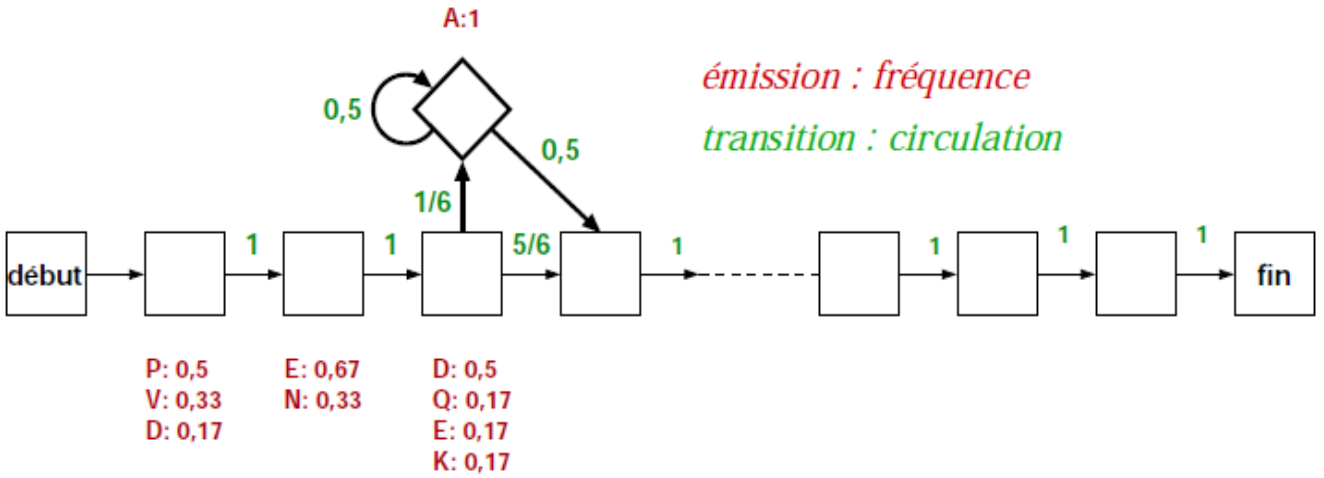


Modélisation avec les HMM

2. Avec des insertions : Une insertion est un fragment de la séquence qui n'apparait pas dans le modèle.

```
PPY_LOPAM/1-36     PED..WASYQAAVRHYVNLITRQRY
PAHO_BOVIN/30-65   PEQ..MAQYAAELRRYINMLTRPRY
PAHO_CHICK/26-61   VED..LIRFYNDLQQYLNVVTRHRY
PAHO_ANSAN/1-36    VED..LRFYYDNLQQYRLNVFRHRY
NPF_HELAS/4-39     PNE..LRQYLKELNEYYAIMGRTRF
NPF_MONEX/1-39     DNKAALRDYLRQINEYFAIIGRPRF
```

un nouvel état = un bloc d'insertion



Modélisation avec les HMM

3. Et finalement, avec les délétions

Une délétion est un fragment du modèle qui ne correspond à aucun acide aminé.

```
PMY_PETMA/1-36    PEE..LSKYMLAVRNYINLITRQRY
PPY_LOPAM/1-36    PED..WASYQAAVRHYVNLITRQRY
PAHO_BOVIN/30-65  PEQ..MAQYAAELRRYINMLTRPRY
PAHO_CHICK/26-61  VED..LIRFYNDLQQYLNVVTRHRY
PAHO_ANSAN/1-36   VED..LRFYYDNLQQYRLNVFRHRY
NPF_HELAS/4-39    PNE..LRQYLKELNEYYAIMGRTRF
NPF_MONEX/1-39    DNKAALRDYLRQINEYFAIIGRPRF
Q9PT97/29-62      AEE..LAKYYSALRHYINLITRQ..
```

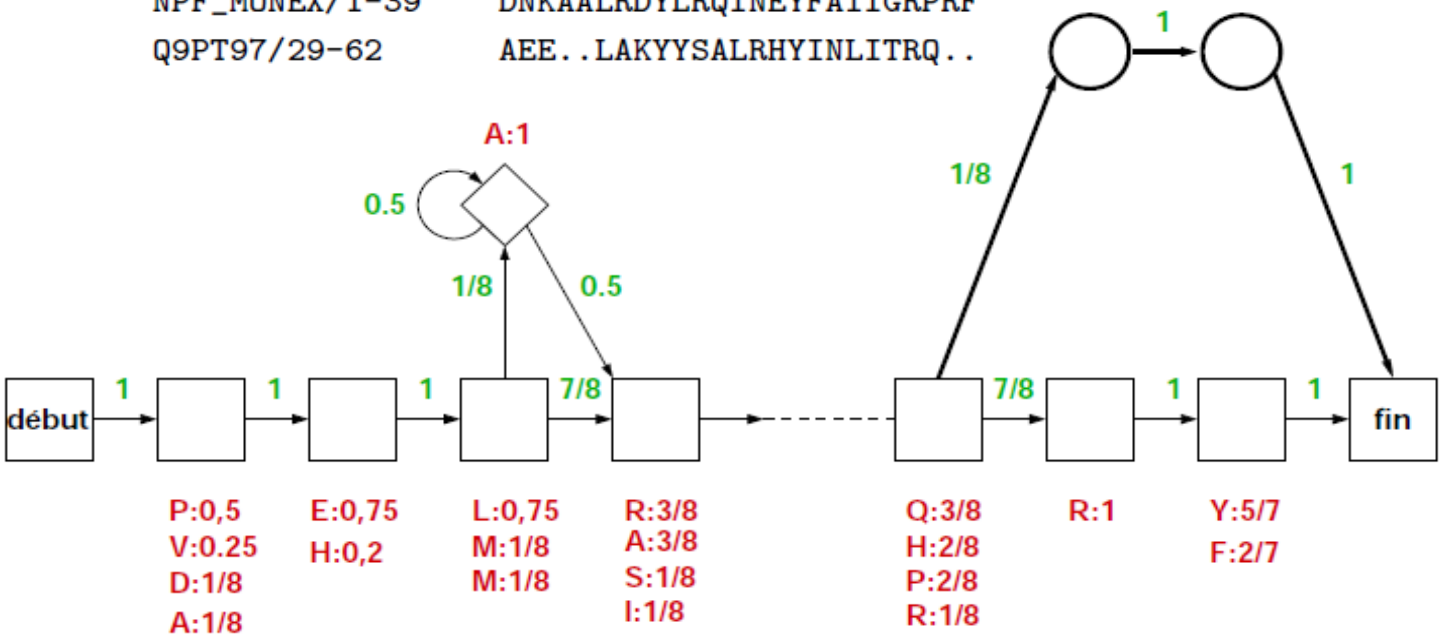
Option 1: Ajouter des arcs entre les états matchants, mais nombre d'arcs quadratique

Option 2: Ajouter des états **silencieux**, qui n'émettent rien

un état = un a.a. délété

Modélisation avec les HMM

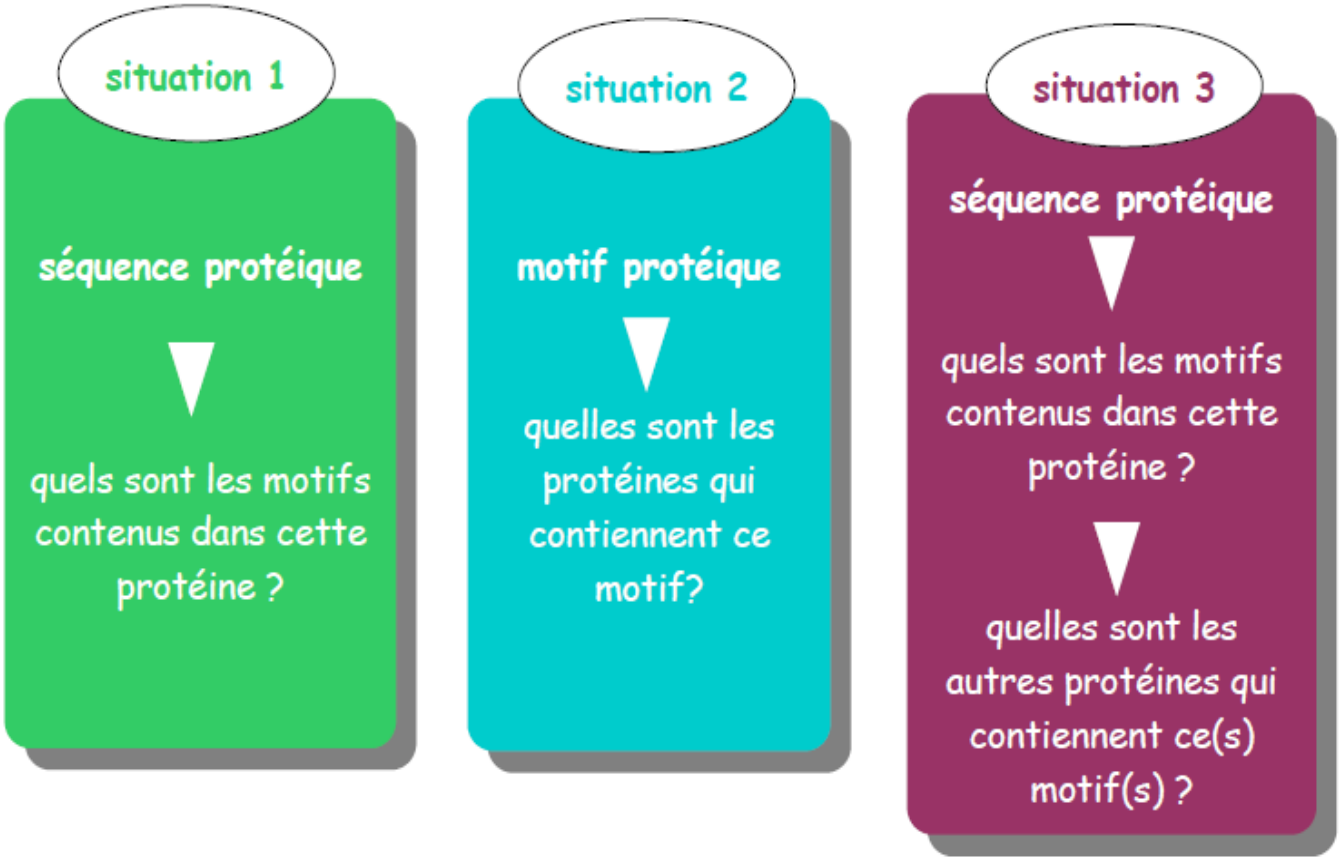
```
PMY_PETMA/1-36     PEE..LSKYMLAVRNYINLITRQRY
PPY_LOPAM/1-36     PED..WASYQAAVRHYVNLITRQRY
PAHO_BOVIN/30-65   PEQ..MAQYAAELRRYINMLTRPRY
PAHO_CHICK/26-61   VED..LIRFYNDLQQYLNVVTRHRY
PAHO_ANSAN/1-36    VED..LRFYYDNLQQYRLNVFRHRY
NPF_HELAS/4-39     PNE..LRQYLKELNEYYAIMGRTRF
NPF_MONEX/1-39     DNKAALRDYLRQINEYFAIIGRPRF
Q9PT97/29-62       AEE..LAKYYSALRHYINLITRQ..
```



Modélisation avec les HMM

- Score : trouver la meilleure interprétation de la séquence dans le modèle
- Algorithme de Viterbi (similaire à l'alignement deux à deux)
- Seuil d'admission : E-value

Retour sur les problématiques



Prosite

- banque de données protéiques
  - collection de motifs protéiques
- expressions régulières (pattern)
- profiles
- outil pour interroger Prosite: ScanProsite
  - https://prosite.expasy.org/scanprosite/

Prosite

ScanProsite tool

This form allows you to scan proteins for matches against the PROSITE collection of motifs as well as against your own patterns.

- (●) **Option 1 - Submit PROTEIN sequences to scan them against the PROSITE collection of motifs.**
- ( ) Option 2 - Submit MOTIFS to scan them against a PROTEIN sequence database.
- ( ) Option 3 - Submit PROTEIN sequences and MOTIFS to scan them against each other.

Reset

STEP 1 - Submit PROTEIN sequences [help]

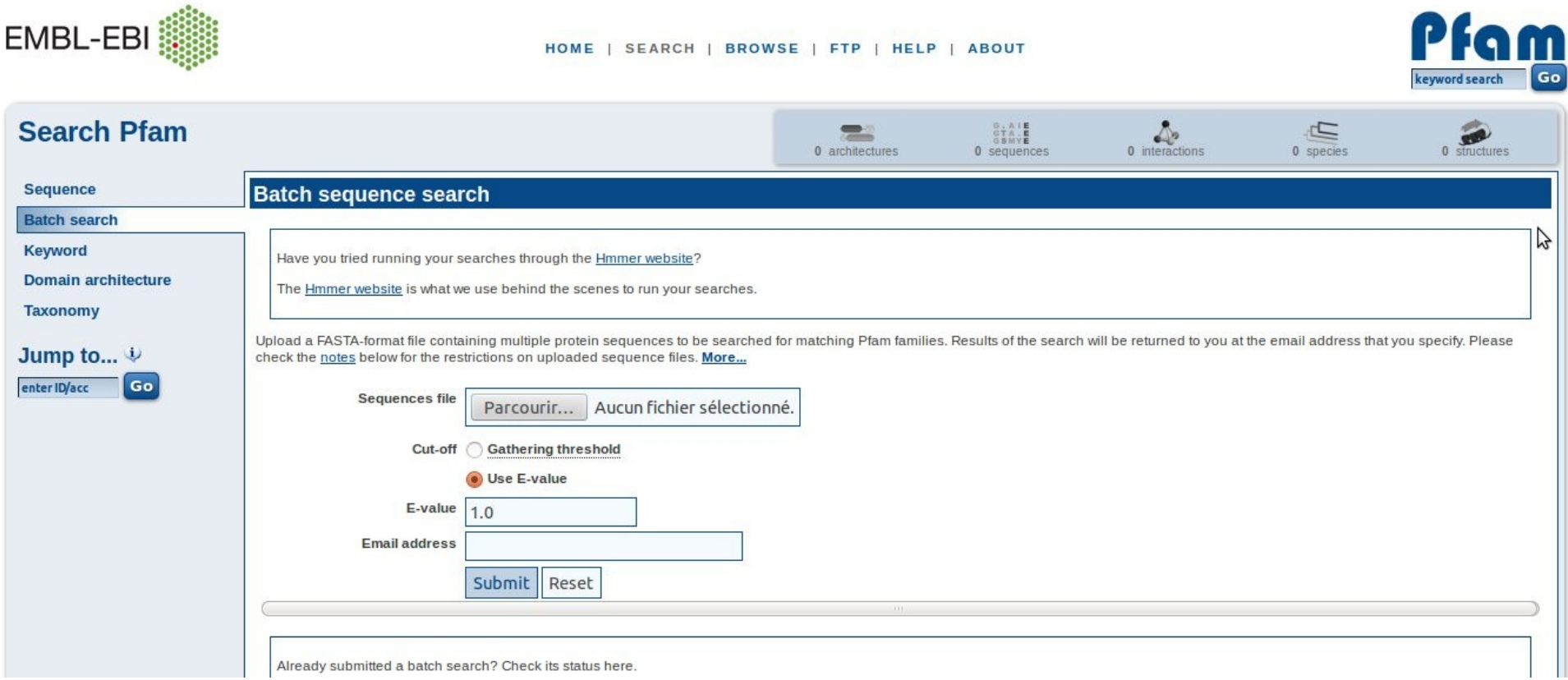
- (●) Submit PROTEIN sequences (max. 10) Examples
- ( ) Submit a PROTEIN database (max. 16MB) for repeated scans (The data will be stored on our server for 1 month).

Enter UniProtKB accessions or identifiers or PDB identifiers or sequences in FASTA format

PFAM

- Pfam utilise des HMM pour décrire des domaines/familles
- deux familles de modèles
  - Pfam-A: HMM de domaines annotés et vérifiés manuellement
  - Pfam-B: générés automatiquement à partir de la base PRODOM (moins fiables!)
- le gros plus: analyse statistique des résultats (E-value)
- Plus fiable

PFAM


EMBL-EBI
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
keyword search
Go
Search Pfam
0 architectures
0 sequences
0 interactions
0 species
0 structures
Sequence
Batch search
Keyword
Domain architecture
Taxonomy
Jump to...
enter ID/acc
Go
Batch sequence search
Have you tried running your searches through the Hmmer website?
The Hmmer website is what we use behind the scenes to run your searches.
Upload a FASTA-format file containing multiple protein sequences to be searched for matching Pfam families. Results of the search will be returned to you at the email address that you specify. Please check the notes below for the restrictions on uploaded sequence files. More...
Sequences file
Parcourir...
Aucun fichier sélectionné.
Cut-off
Gathering threshold
Use E-value
E-value
1.0
Email address
Submit
Reset
Already submitted a batch search? Check its status here.

InterPro

- Ressource qui permet l'analyse fonctionnelle des séquences protéiques en les classant en famille et en prédisant la présence de domaines et de sites importants
- Regroupement de plusieurs banques de données (14 ressources dont PFAM et prosite)
- InterProScan est l'outil qui permet de scanner des séquences contre les signatures d'InterPro

Conclusions

- régions homologues évolutivement conservées > domaines/motifs protéiques
- domaine/motif = unité fonctionnelle/structurale
- identifier des domaines protéiques dans une séquence = étape essentielle de l'annotation fonctionnelle !
- meilleure stratégie: interroger plusieurs banques de données complémentaires (InterPro)
- lecture des fiches de domaines > indications sur la fonction potentielle associée au domaine
- attention aux prédictions faites !!!

Motifs biologiques : Exercices

Exercice 13

Sources

- http://what-when-how.com/bioinformatics/spliced-alignment-bioinformatics/
- D'après les cours de :
  - l'équipe l’équipe Bonsai, CRIStAL UMR 9189
  - Géraldine PASCAL, Genotoul bioinfo
  - Hélène Touzet (équipe bonsai)
  - Carl Herrmann (IBDML, Université de la Méditerranée)
- Recent progress in multiple sequence alignment: a survey, Notredame C., Pharmacogenomics. 2002 Jan;3(1):131-44
- Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features, Iwata *et al.*, Nucl. Acids Res. (2012)