M1 Bioinformatique, Connaissances et Données

Master Sciences et Numérique pour la Santé

Année 2016-2017

HMSN206 - Partie Alignement

Partie II-1 : Alignement multiple global

Anne-Muriel Arigon Chifolleau

http://www.lirmm.fr/~arigon/enseignement/HMSN206/

LIRMM - UM

- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
- La suite ...

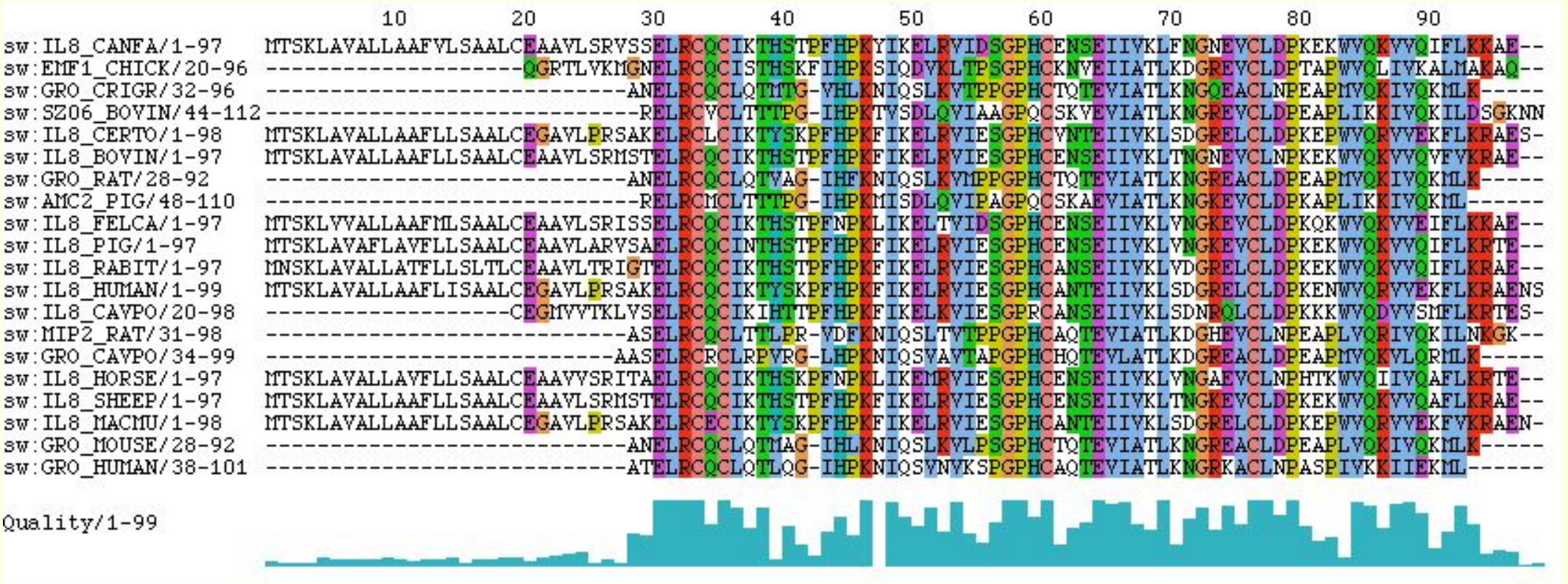
Alignement multiple

- Permet d'aligner plusieurs séquences simultanément
- En général pour les protéines
- Alignements faits à la main par les experts biologistes
- Généralisation naturelle de l'alignement 2 à 2 mais beaucoup plus complexe en terme de calcul

Problème : Étant données k séquences $s_1, s_2, \ldots, s_k$, trouver le meilleur alignement multiple pour ces séquences

- Ce pb est NP-complet ($\Rightarrow \nexists$ de solution exacte en tps raisonnable)

Exemples d'alignement



Intérêt des alignements multiples

- Recherche dans les banques $\Rightarrow$ plrs séquences similaires à la requête

  Il est naturel de vouloir aligner ces séquences entre elles
- MSA détectent les régions qui ont été conservées lors de l'évolution

  Très svt des domaines associés à une fonction clé de la molécule
- Plusieurs protéines de fonctions similaires dans différentes espèces

  $\rightarrow$ Quelles parties semblables ? $\Rightarrow$ Consensus/Profil

  $\rightarrow$ Quelles parties différentes ?
- Permet de trouver d'autres membres d'une famille de protéines
- Séquençage de génomes (assemblage, recouvrement EST)
- Point de départ pour les analyses phylogénétiques

Alignements multiples plus informatifs que les ali. de 2 séq.

Les 2 grands types de MSA

- Un MSA peut être global ou local (comme un alignement de 2 séquences)

- Global : l'alignement 2 à 2 est étendu pour inclure 3 séq. ou plus
  Des protéines de $\neq$ organismes peuvent être conservées sur toute la longueur si elles assurent une fonction biologique importante

  Logiciels : CLUSTALW, CLUSTAL Ω, T-COFFEE, MUSCLE, MAFFT, MULTALIN, DIALIGN, ...

- Local : recherche de domaines/régions conservés
  Les domaines fonctionnels de protéines peuvent être conservés tandis que le reste de la séquence diverge

  Logiciels : DIALIGN, BLOCKS Web site, eMOTIF, GIBBS, HMMER, ...

Comment quantifier la qualité d'un alignement multiple ?

- Généraliser une fonction de score pour un alignement de k séq.
- Le score d'un MSA est la somme des scores de ses colonnes (ici les colonnes sont de hauteur k)
- Les colonnes sont considérées indépendantes
- Fonction à k paramètres ?
  Si longueur des séquences $\alpha \Rightarrow \alpha^k$ colonnes possibles
  ($22^5 > 5$ millions)
  On ne peut pas associer un coût à chaque colonne ! ? !
- Mesure raisonnable S, quelles propriétés ?

1. Même score pour les colonnes contenant les mêmes caractères (indépendamment de l'ordre)

$$S(I,-,I,V) = S(V,I,I,-) = S(V,I,-,I) = S(V,-,I,I) = \cdots$$

2. Récompense les colonnes avec beaucoup de résidus identiques ou similaires

3. Pénalise les colonnes avec des résidus différents et des indels (gaps)

- Plusieurs méthodes de score : méthode SP, méthodes basées sur la phylogénie (arbre ou étoile), les concensus, les profils, le contenu en information (entropie), la concistance, . . .

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

  $$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V
					60

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$

$p(N,C) = -3$

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V
	60	24			

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$

$p(N,C) = -3$

$p(C,C) = 9$

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V
	60	24	9		

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$

$p(N,C) = -3$

$p(C,C) = 9$

$p(I,I) = 4$

$p(I,-) = -2$

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V
	60	24	9	16	

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$
$p(N,C) = -3$
$p(C,C) = 9$
$p(I,I) = 4$
$p(I,-) = -2$
$p(V,V) = 4$

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V
	60	24	9	16	40

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$
$p(N,C) = -3$
$p(C,C) = 9$
$p(I,I) = 4$
$p(I,-) = -2$
$p(V,V) = 4$

1	N	N	N	I	V	
2	N	N	N	I	V	
3	N	N	N	–	V	
4	N	N	C	I	V	
5	N	C	C	I	V	
	60	24	9	16	40	=

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$
$p(N,C) = -3$
$p(C,C) = 9$
$p(I,I) = 4$
$p(I,-) = -2$
$p(V,V) = 4$

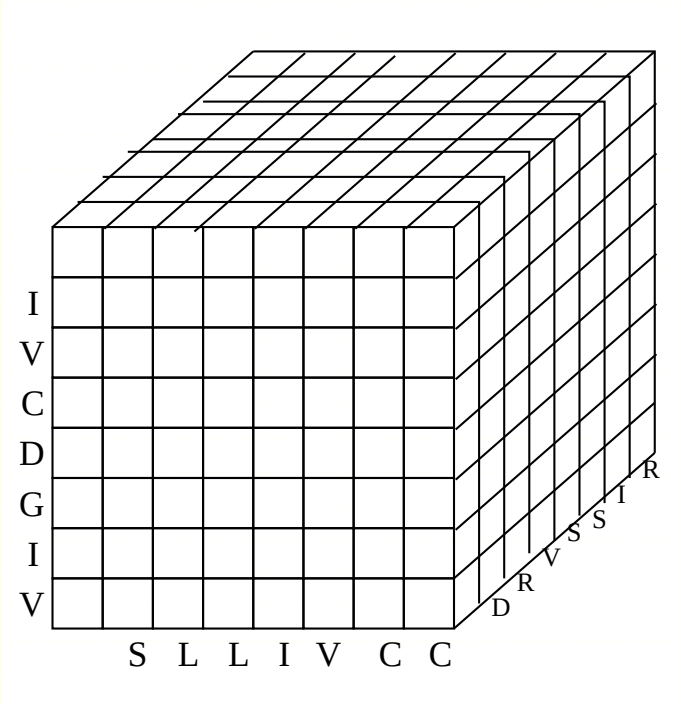
1	N	N	N	I	V		
2	N	N	N	I	V		
3	N	N	N	–	V		
4	N	N	C	I	V		
5	N	C	C	I	V		
	60	24	9	16	40	=	**149**

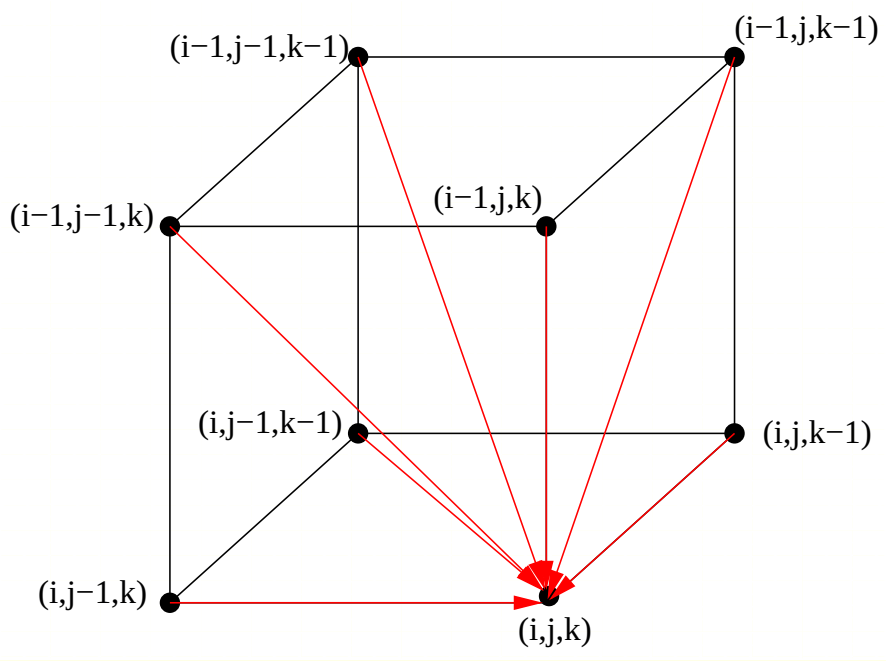
- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

- Le principe de programmation dynamique est généralisable à k séquences de longueur n
- Matrice $\mathcal{A}$ de dimension n^k
- $\mathcal{A}(i_1, i_2, \ldots, i_k)$ contient le score d'alignement optimal entre les préfixes $s_1[1..i_1], s_2[1..i_2], \ldots, s_k[1..i_k]$
- Remplissage des n^k cases de la table $\Rightarrow$ espace mémoire $O(n^k)$
- Chaque entrée ($\sim$ case) dépend de $2^k - 1$ entrées déjà calculées
- Calculer le score SP requiert $O(k^2)$ car il y a $\frac{k(k-1)}{2}$ paires

Temps total d'exécution en $O(k^2 2^k n^k)$

Exemple avec 3 séquences





Exemple avec 3 séquences : la récurrence

- 3 séquences : r, s et t, méthode de score SP et g pénalité de gap
- $\mathcal{A}$ matrice de prog. dyn.

Initialisation : $\mathcal{A}(0,0,0) = 0$; $\mathcal{A}(i,0,0) = i \times 2g$;
$\mathcal{A}(0,j,0) = j \times 2g$; $\mathcal{A}(0,0,k) = k \times 2g$.

Remplissage :

$$\mathcal{A}(i,j,k) = \max \begin{cases} \mathcal{A}(i-1,j-1,k-1) + SP(r_i, s_j, t_k) \\ \mathcal{A}(i,j-1,k-1) + SP(-, s_j, t_k) \\ \mathcal{A}(i-1,j,k-1) + SP(r_i, -, t_k) \\ \mathcal{A}(i-1,j-1,k) + SP(r_i, s_j, -) \\ \mathcal{A}(i,j,k-1) + SP(-, -, t_k) \\ \mathcal{A}(i,j-1,k) + SP(-, s_j, -) \\ \mathcal{A}(i-1,j,k) + SP(r_i, -, -) \end{cases}$$

Méthode exacte : conclusion

- Méthode guère plus complexe pour k séquences que pour 2 (facilement programmable)

- Mais le temps de calcul en $O(k^2 2^k n^k)$, ainsi que la place mémoire nécessaire deviennent prohibitif quand k augmente.

- Illustration :
  $\rightarrow$ 2 séquences de 100 a.a. $\rightarrow$ 1 sec.
  $\rightarrow$ 3 séquences de 100 a.a. $\rightarrow$ 10 min
  $\rightarrow$ 4 séquences de 100 a.a. $\rightarrow \sim$ 3 jours
  $\rightarrow$ à partir de 9 séq. le tps de calcul dépasse l'âge de l'univers
  ...

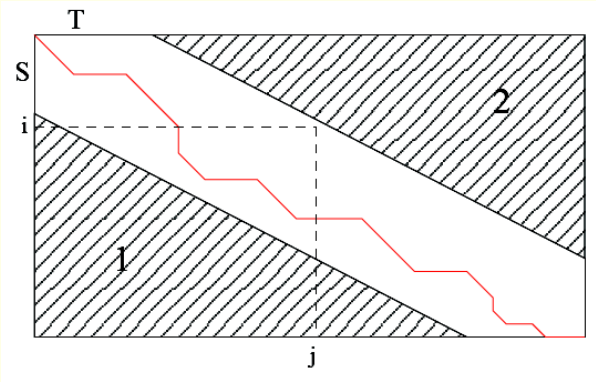
$\Rightarrow$ Mise au point d'algorithmes heuristiques performants et de bonne qualité (pb encore ouvert aujourd'hui)

- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
  - → Alignement exact (ex : MSA, DCA)
  - → Alignement progressif (ex : CLUSTALW)
  - → Amélioration de l'approche progressive (ex : T-COFFEE, MAFFT, DIALIGN,. . .)
  - → Autres méthodes d'alignement multiple
  - → Evaluation de MSA
- La suite ...

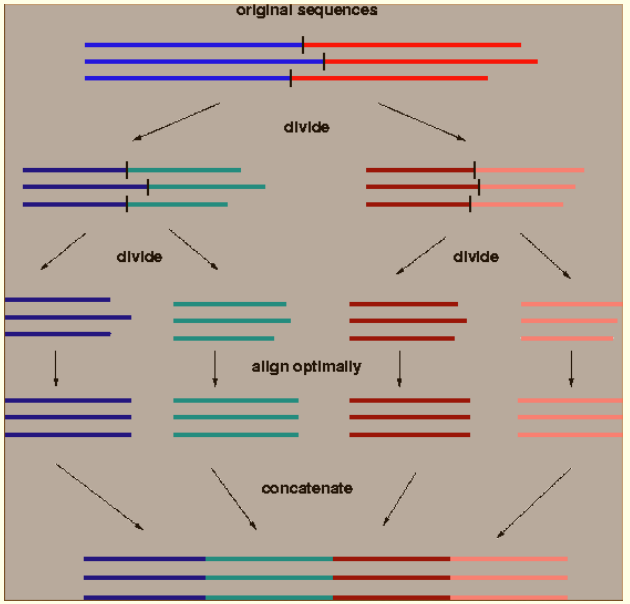
- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
  - → Alignement exact (ex : MSA, DCA)
  - → Alignement progressif (ex : CLUSTALW)
  - → Amélioration de l'approche progressive (ex : T-COFFEE, MAFFT, DIALIGN,. . .)
  - → Autres méthodes d'alignement multiple
  - → Evaluation de MSA
- La suite ...

Algorithmes basés sur une méthode exacte

- Objectif : alignement simultané de séquences 'optimal'
- Variante du principe de programmation dynamique multi-dimensionnelle
- Measurement Systems Analysis (MSA) software : heuristique basée sur l'algorithme complet de N&W
  $\Rightarrow$ restreindre le calcul de l'alignement autour de la diagonale



- Divise Conquer multiple sequence Alignement (DCA) software : heuristique basée sur MSA et l'approche « diviser pour régner »



- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
  - → Alignement exact (ex : MSA, DCA)
  - → Alignement progressif (ex : CLUSTALW)
  - → Amélioration de l'approche progressive (ex : T-COFFEE, MAFFT, DIALIGN,. . .)
  - → Autres méthodes d'alignement multiple
  - → Evaluation de MSA
- La suite ...

Méthode progressive

- Difficile d'aligner simultanément k séquences $\Rightarrow$ construction itérative d'un alignement multiple en regroupant les alignements par paire
  1. on aligne toutes les paires de séquences et un score est donné à chaque alignement par paire
  2. on choisit l'ordre d'alignement des séquences
  3. on aligne les séquences progressivement en se basant sur les alignements par paire

- Plusieurs alogorithmes qui utilisent différents critères à chaque étape :
  $\rightarrow$ l'algorithme d'aln par paire (prog dyn, point d'ancrage, ...) et la fonction de score / distance (score, matrice de substitution, ...)
  $\rightarrow$ l'ordre d'aln des séquences (arbre guide, étoile, ...)
  $\rightarrow$ la méthode pour aligner les groupes de séquenses (aln de profils, aln d'aln, ...)

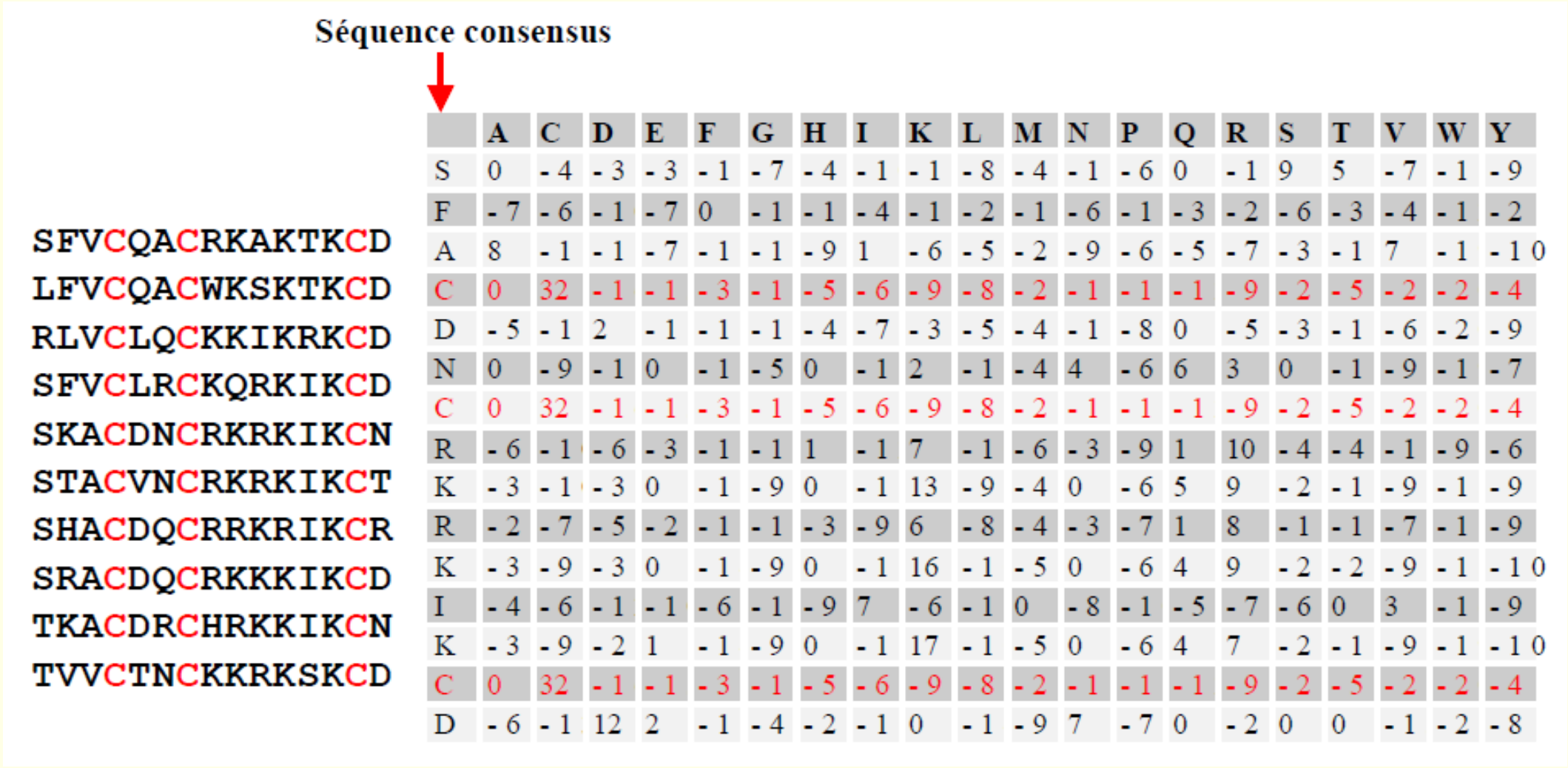
- Approche la + utilisée pour l'alignement multiple global : rapide, peu de mémoire et bonnes performances avec des séquences conservées

- profil $\sim$ alignement intermédiaire dans l'alignement progressif

- Plusieurs méthodes pour progressivement alignées les séquences
  - $\rightarrow$ alignement de 2 séquences
  - $\rightarrow$ alignement d'une séquence et d'un profil
  - $\rightarrow$ alignement de 2 profils

- Méthode du Profil
  - $\rightarrow$ Profil = matrice de scores position-spécifique dans un alignement multiple (Position-Specific Scoring Matrix=PSSM)
  - $\rightarrow$ 1 colonne de l'alignement = 1 ligne dans le profil
  - $\rightarrow$ Exemple de méthode de construction : les scores pour chaque type de résidu sont calculés à partir d'une matrice de substitutions (PAM, Blosum, ...) et des fréquences de variation observées pour une position donnée de l'alignement

Exemple de profil



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
S	0	-4	-3	-3	-1	-7	-4	-1	-1	-8	-4	-1	-6	0	-1	9	5	-7	-1	-9
F	-7	-6	-1	-7	0	-1	-1	-4	-1	-2	-1	-6	-1	-3	-2	-6	-3	-4	-1	-2
A	8	-1	-1	-7	-1	-1	-9	1	-6	-5	-2	-9	-6	-5	-7	-3	-1	7	-1	-10
C	0	32	-1	-1	-3	-1	-5	-6	-9	-8	-2	-1	-1	-1	-9	-2	-5	-2	-2	-4
D	-5	-1	2	-1	-1	-1	-4	-7	-3	-5	-4	-1	-8	0	-5	-3	-1	-6	-2	-9
N	0	-9	-1	0	-1	-5	0	-1	2	-1	-4	4	-6	6	3	0	-1	-9	-1	-7
C	0	32	-1	-1	-3	-1	-5	-6	-9	-8	-2	-1	-1	-1	-9	-2	-5	-2	-2	-4
R	-6	-1	-6	-3	-1	-1	1	-1	7	-1	-6	-3	-9	1	10	-4	-4	-1	-9	-6
K	-3	-1	-3	0	-1	-9	0	-1	13	-9	-4	0	-6	5	9	-2	-1	-9	-1	-9
R	-2	-7	-5	-2	-1	-1	-3	-9	6	-8	-4	-3	-7	1	8	-1	-1	-7	-1	-9
K	-3	-9	-3	0	-1	-9	0	-1	16	-1	-5	0	-6	4	9	-2	-2	-9	-1	-10
I	-4	-6	-1	-1	-6	-1	-9	7	-6	-1	0	-8	-1	-5	-7	-6	0	3	-1	-9
K	-3	-9	-2	1	-1	-9	0	-1	17	-1	-5	0	-6	4	7	-2	-1	-9	-1	-10
C	0	32	-1	-1	-3	-1	-5	-6	-9	-8	-2	-1	-1	-1	-9	-2	-5	-2	-2	-4
D	-6	-1	12	2	-1	-4	-2	-1	0	-1	-9	7	-7	0	-2	0	0	-1	-2	-8

Aln progressif avec arbre guide : CLUSTALW

- Programme original CLUSTAL utilisé depuis 1988, régulièrement amélioré depuis [Higgins & Sharp, 88]
- CLUSTALW ([Thompson *et al*, 94]) = version la plus classique (W pour *Weighting*) ; les séquences et les paramètres sont pondérés
- CLUSTAL Omega ([Sievers *et al*, 11]) = version la plus récente ; amélioration en terme d'échelle et de qualité des aln (cf. partie suivante)
- Fonctionnalités : ajout d'une séq. ou d'un ali. à un ali. déjà fait, production d'un arbre phylogénétique, paramètre *slow/fast*, . . .
- Méthode
  1. Construction de la matrice des distances
  2. Construction d'un arbre guide
  3. Alignement progressif suivant cet arbre

Etapes de ClustalW (1)

- Trois étapes

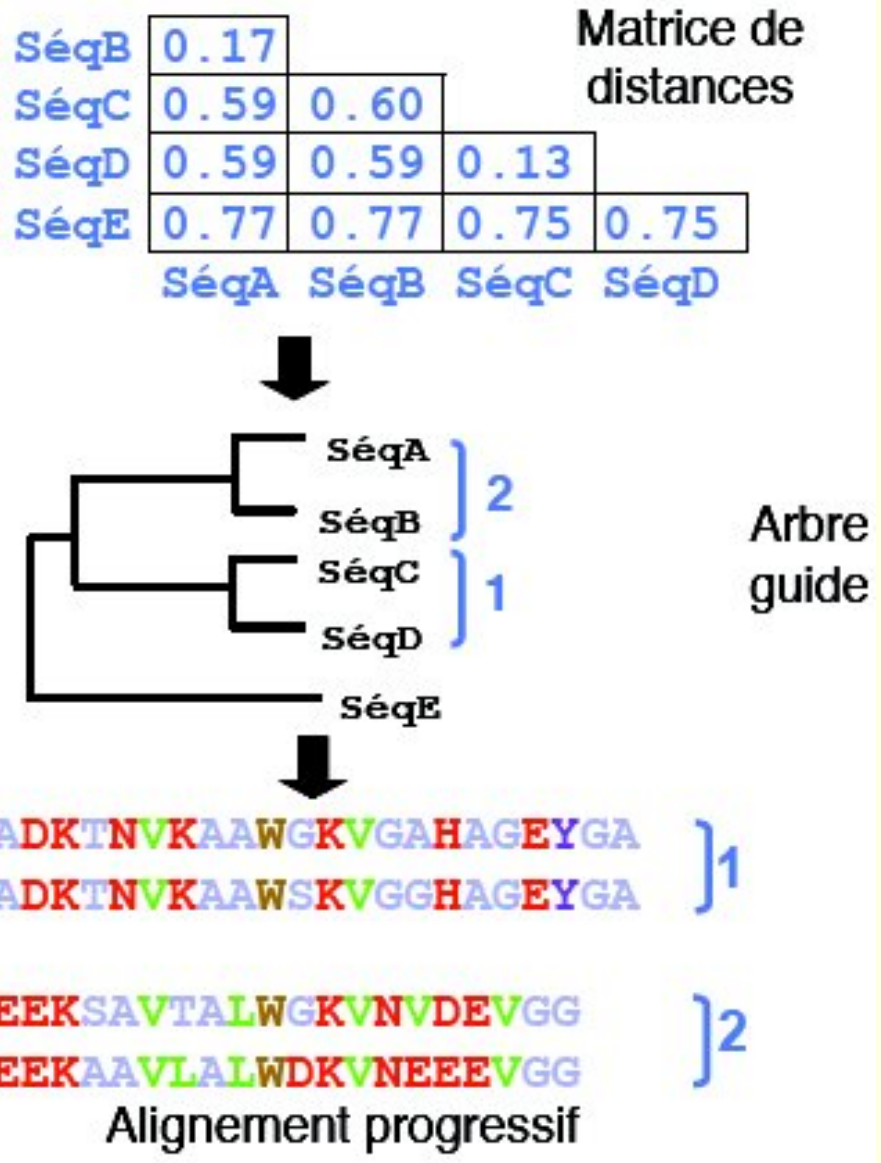
Alignement de toutes les paires possibles et établissement d'une matrice de distances basée sur les scores des alignements

⬇

Construction d'un arbre guide à l'aide des distances d'alignements calculées précédemment

⬇

Alignement des séquences ou des groupes de séquences en suivant l'ordre déterminé par l'arbre guide



Etapes de ClustalW (2)

- Trois étapes

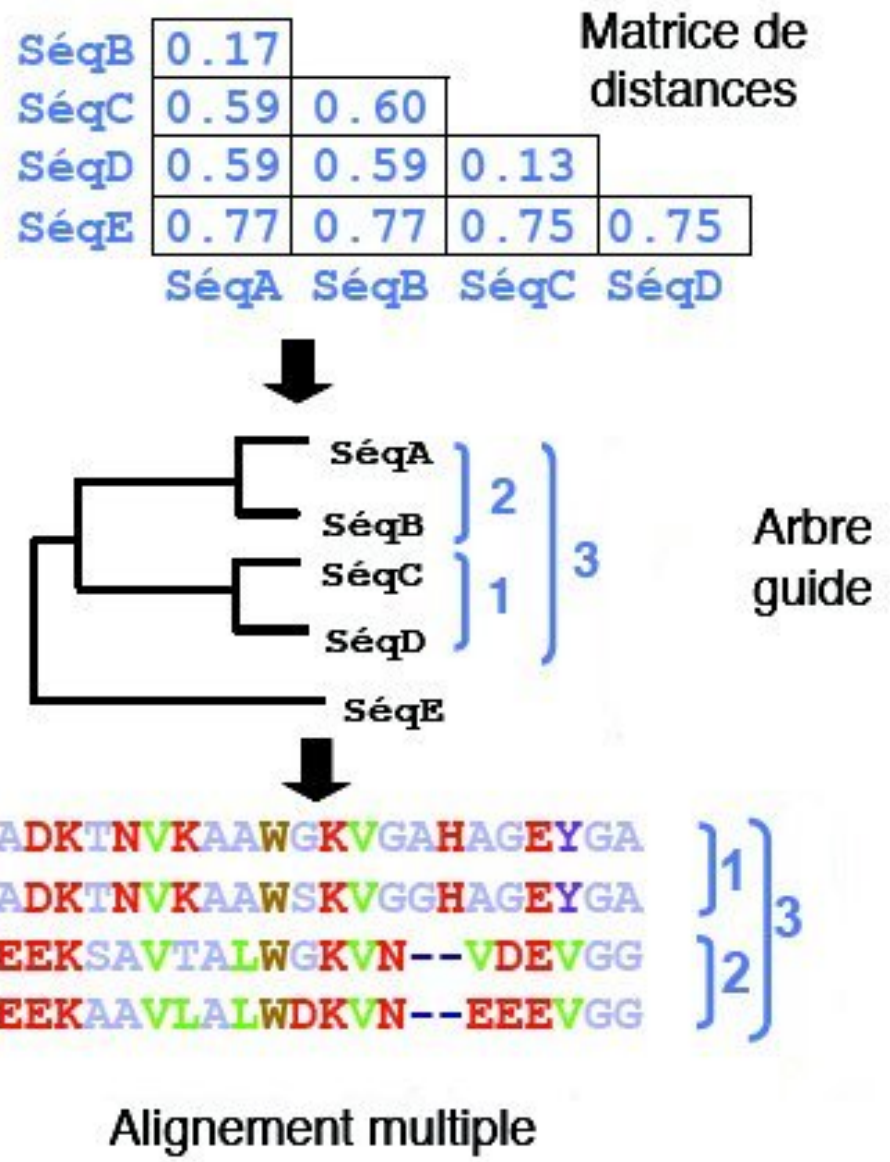
Alignement de toutes les paires possibles et établissement d'une matrice de distances basée sur les scores des alignements

⬇

Construction d'un arbre guide à l'aide des distances d'alignements calculées précédemment

⬇

Alignement des séquences ou des groupes de séquences en suivant l'ordre déterminé par l'arbre guide



Etapes de ClustalW (3)

- Trois étapes

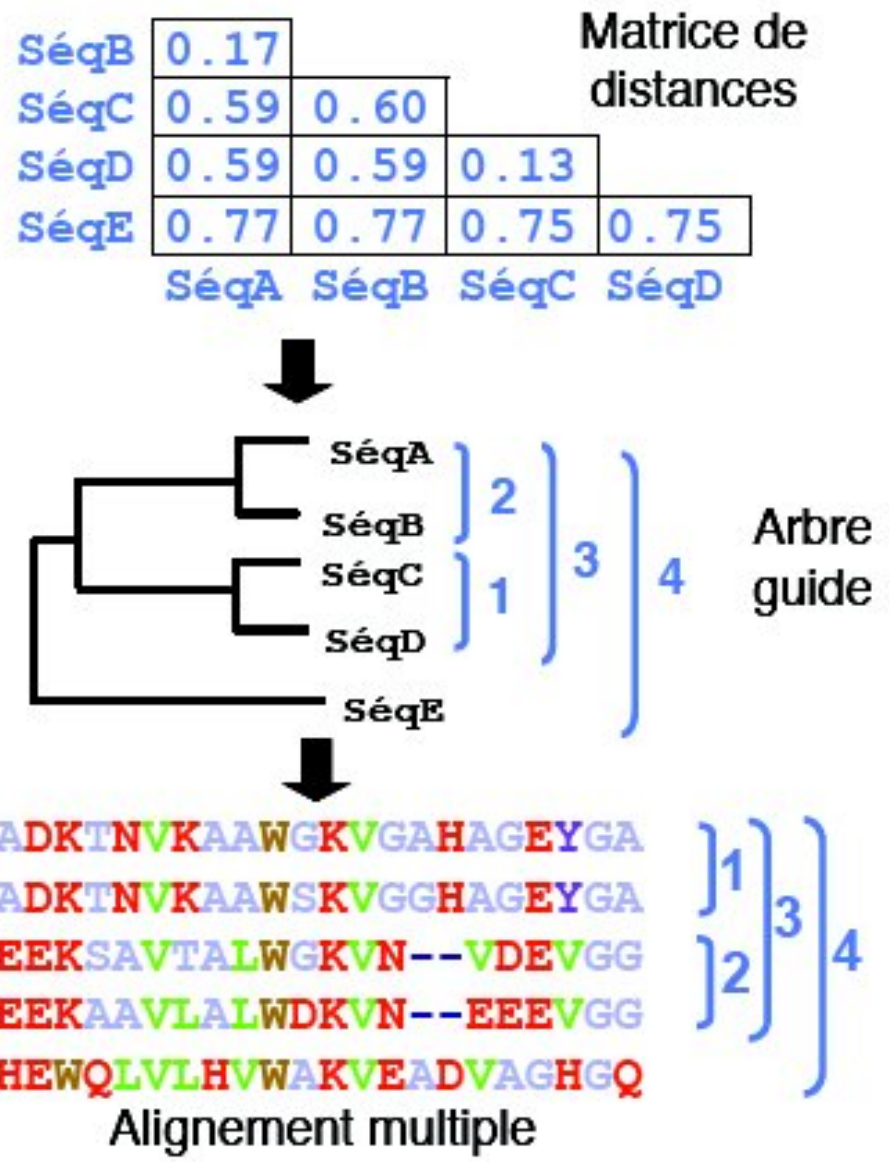
Alignement de toutes les paires possibles et établissement d'une matrice de distances basée sur les scores des alignements

⬇

Construction d'un arbre guide à l'aide des distances d'alignements calculées précédemment

⬇

Alignement des séquences ou des groupes de séquences en suivant l'ordre déterminé par l'arbre guide

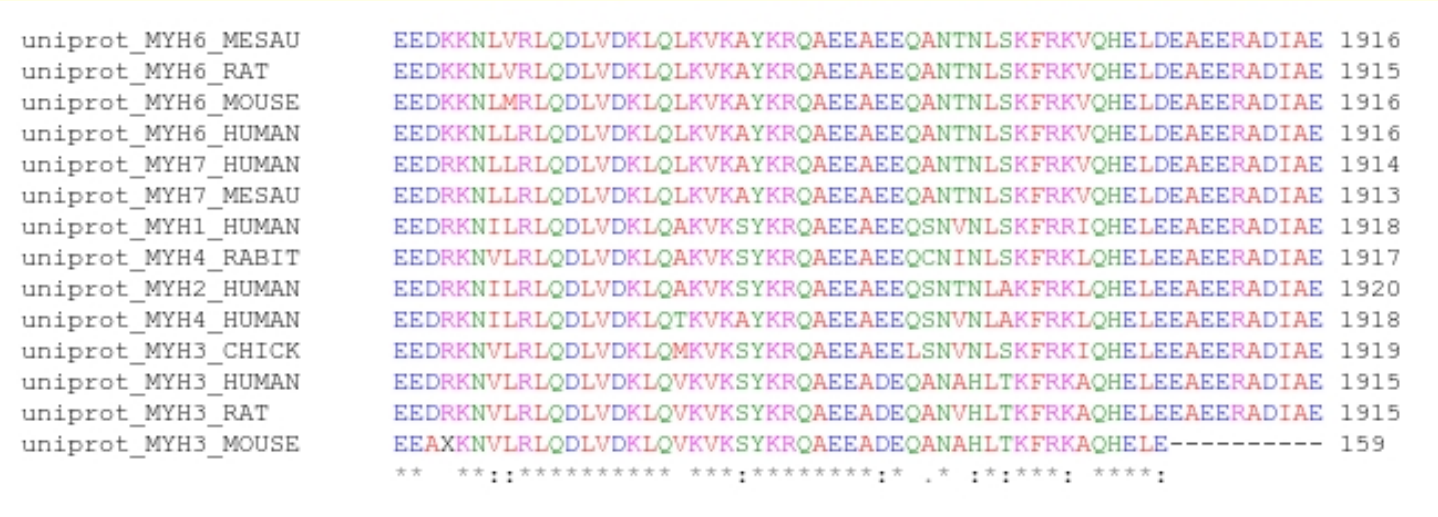


- Pondération des séquences :
  Une séquence similaire à d'autres dans le groupe a un poids faible, alors qu'une séquence moins proche des autres a un poids plus fort

  $\Rightarrow$ On diminue ainsi le poids des groupes de séquences similaires et privilégie les changements dans l'arbre évolutif

- Pénalité de gap :
  CLUSTALW pénalise les gaps de manière à les placer entre les domaines conservés (utilise une matrice spéciale et $\neq$ pénalités suivant les régions)

Méthode de score de CLUSTALW $\neq$ SP-score



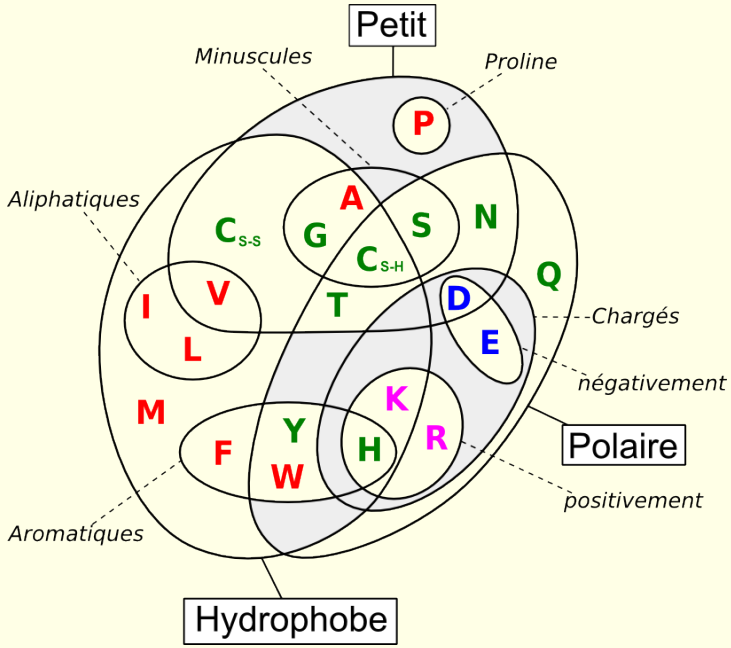
- CLUSTALW affiche par défaut les symboles suivants pour indiquer le degrés de conservation dans chaque colonne :

  → * caractère identique dans toute la colonne

  → : substitutions conservatives (suivant la table des couleurs)

  → . substitutions semi-conservatives

Code couleur de CLUSTALW



A.A.	Couleur	Description
AVFPMILW	Rouge	Petits (petits + hydrophobes (incl. aromatiques -Y))
DE	Bleu	Acides
RK	Magenta	Basiques - H
STYHCNGQ	Vert	Hydroxyl + sulfhydryl + amine + G
Autres	Gris	Unusual amino/imino acids . . .

(http://fr.wikipedia.org/wiki/Acide_aminé)

- Un autre type d'alignement progressif : la méthode de l'étoile centrale

- Principe : alignement 2 à 2 entre une séquence fixée (le centre de l'étoile) et toutes les autres séquences

- Méthode :
  1. Alignements globaux de toutes les séquences 2 par 2
  2. Choisir la séquence centre s_c en fonction des alignements par paire (la plus proche de toutes les autres)
  3. Construction de l'alignement multiple par juxtaposition des alignements 2 à 2 entre toutes les s_i pour $i \neq c$ et s_c (s_c est la séquence guide)

- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
  - → Alignement exact (ex : MSA, DCA)
  - → Alignement progressif (ex : CLUSTALW)
  - → Amélioration de l'approche progressive (ex : T-COFFEE, MAFFT, DIALIGN,. . .)
  - → Autres méthodes d'alignement multiple
  - → Evaluation de MSA
- La suite ...

- Inconvénients de l'approche progressive :
  → toute création d'un trou est définitive ⇒ « *once a gap, ever a gap* »
  → comparaison simultanée de 2 séquences uniquement
  → résultat dépendant des méthodes choisies (score, ordre, . . .)

- Différentes méthodes d'amélioration de l'alignement progressif
  → principe de la consistance
  → raffinement itératif
  → utilisation de modèles probabilistes : profils HMM
  → . . .

- Rmq : fonction de score ("objective/scoring function") = paramètre le + critique d'une méthode d'aln

Fonction de score basée sur la consistance

- Les méthodes de score utilisées par les algorithmes d'alignements par paires sont un élément trés influent dans l'algorithme progressif :
  - → basées sur les matrices (ClustalW, Muscle, Kalign, . . .)
  - → basées sur la consistance (T-Cofee, MAFFT, ProbCons, . . .)

- Contraintes de consistance: prises en compte de combinaisons consistantes de séquences

```
SéqA  GARFIELD THE LAST FAT CAT
SéqB  GARFIELD THE FAST CAT
SéqC  GARFIELD THE VERY FAST CAT
SéqD  THE FAT CAT
```

```
SéqA  GARFIELD THE LAST FAT CAT
SéqB  GARFIELD THE FAST CAT ---

SéqA  GARFIELD THE LAST FA-T CAT
SéqC  GARFIELD THE VERY FAST CAT

SéqA  GARFIELD THE LAST FAT CAT
SéqD  -------- THE ---- FAT CAT
```

```
SéqB  GARFIELD THE ---- FAST CAT
SéqC  GARFIELD THE VERY FAST CAT

SéqB  GARFIELD THE FAST CAT
SéqD  -------- THE FA-T CAT

SéqC  GARFIELD THE VERY FAST CAT
SéqD  -------- THE ---- FA-T CAT
```

Alignement A-B retenu:

```
SéqA  GARFIELD THE LAST FAT CAT
SéqB  GARFIELD THE FAST --- CAT
```

=> Utilisation de { données internes (alignements par paire)
                  { données externes (structures 3D)

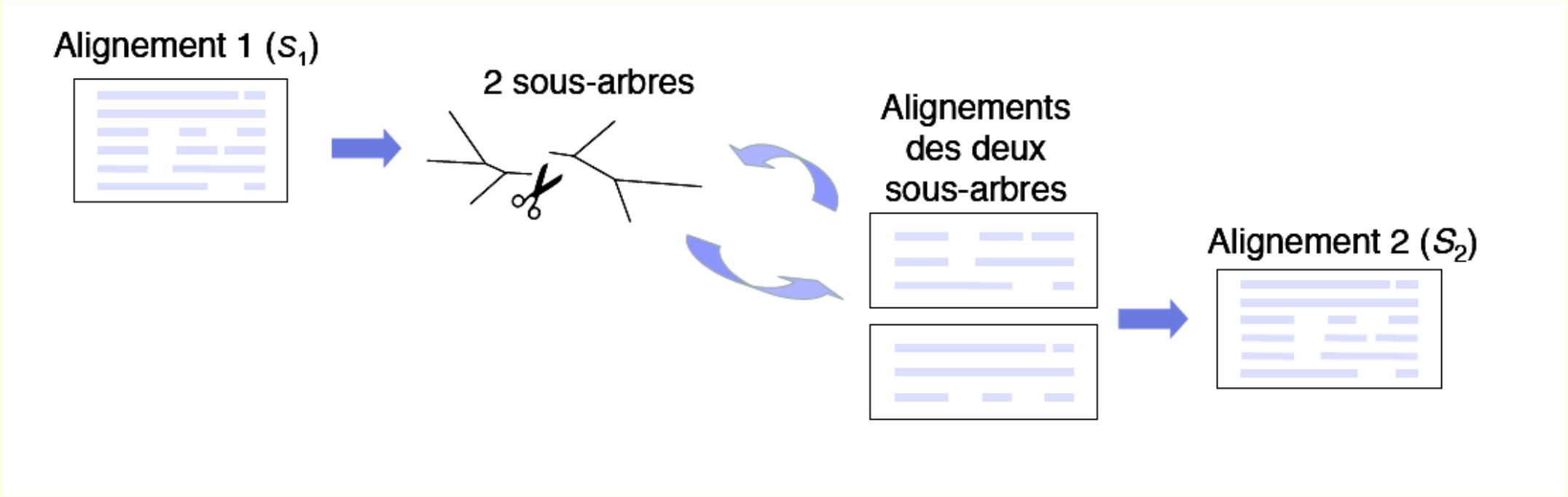
- Le pb majeur avec l'alignement progressif est qu'une erreur faite au début de l'alignement ne peut être corrigée par la suite

  $\Rightarrow$ une étape de rafinement est ajoutée (Muscle, MAFFT, . . .)

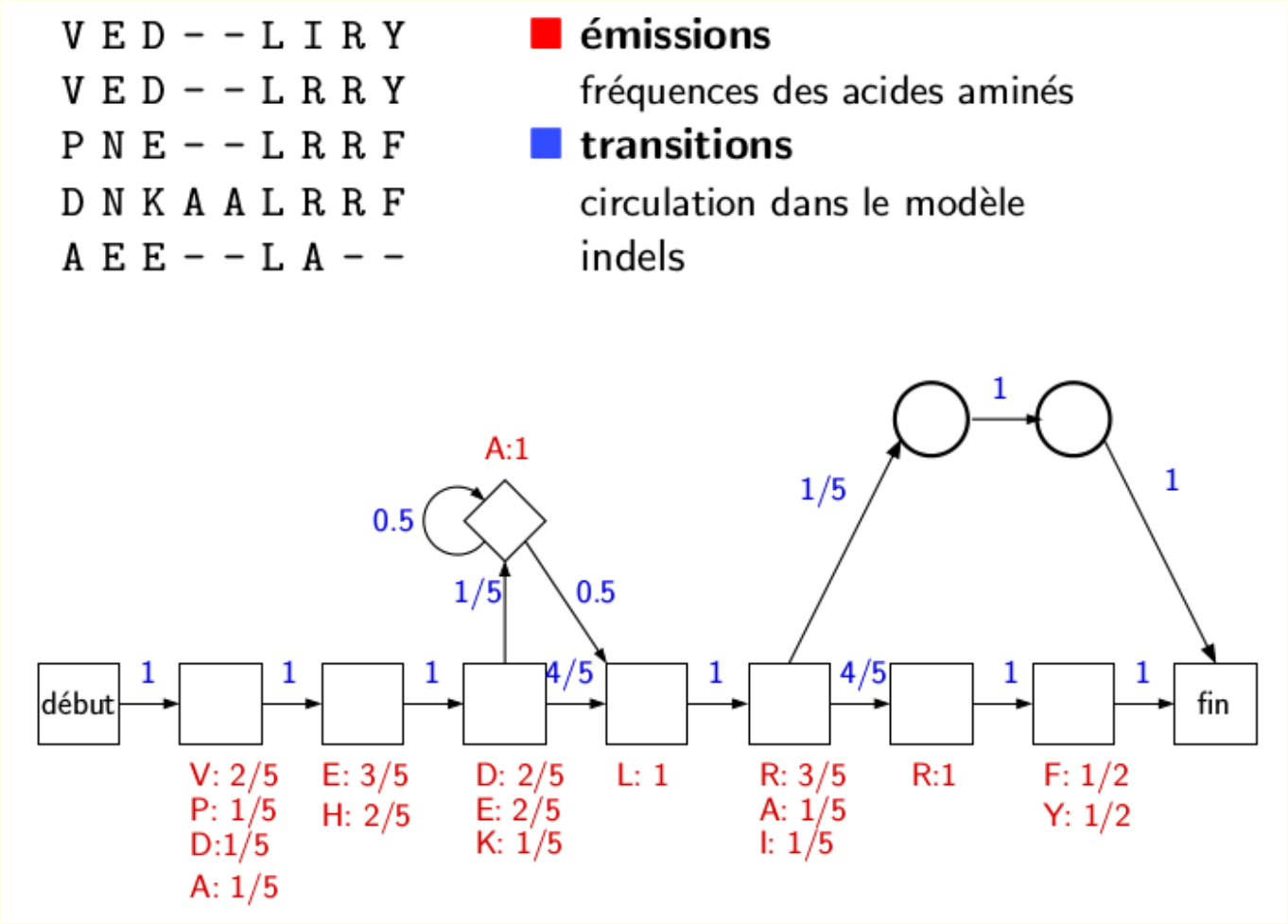
- Objectif = améliorer le score d'alignement global en réalignant des sous-groupes de séquences de manière répétée puis en alignant ces sous-groupes dans l'alignement global de toutes les séquences

- La sélection de ces sous-groupes peut se faire sur la base de l'arbre guide, ou la séparation d'une à deux séquences du reste, ou de manière aléatoire

Exemple de rafinement d'alignement progressif



- Représentation des alignements par des profils HMM (Clustal Omega, ProbCons, . . .)
- Modèles de Markov cachés (HMM)



- Basés sur l'approche progressive
- Utilisant différentes méthodes pour améliorer l'aln

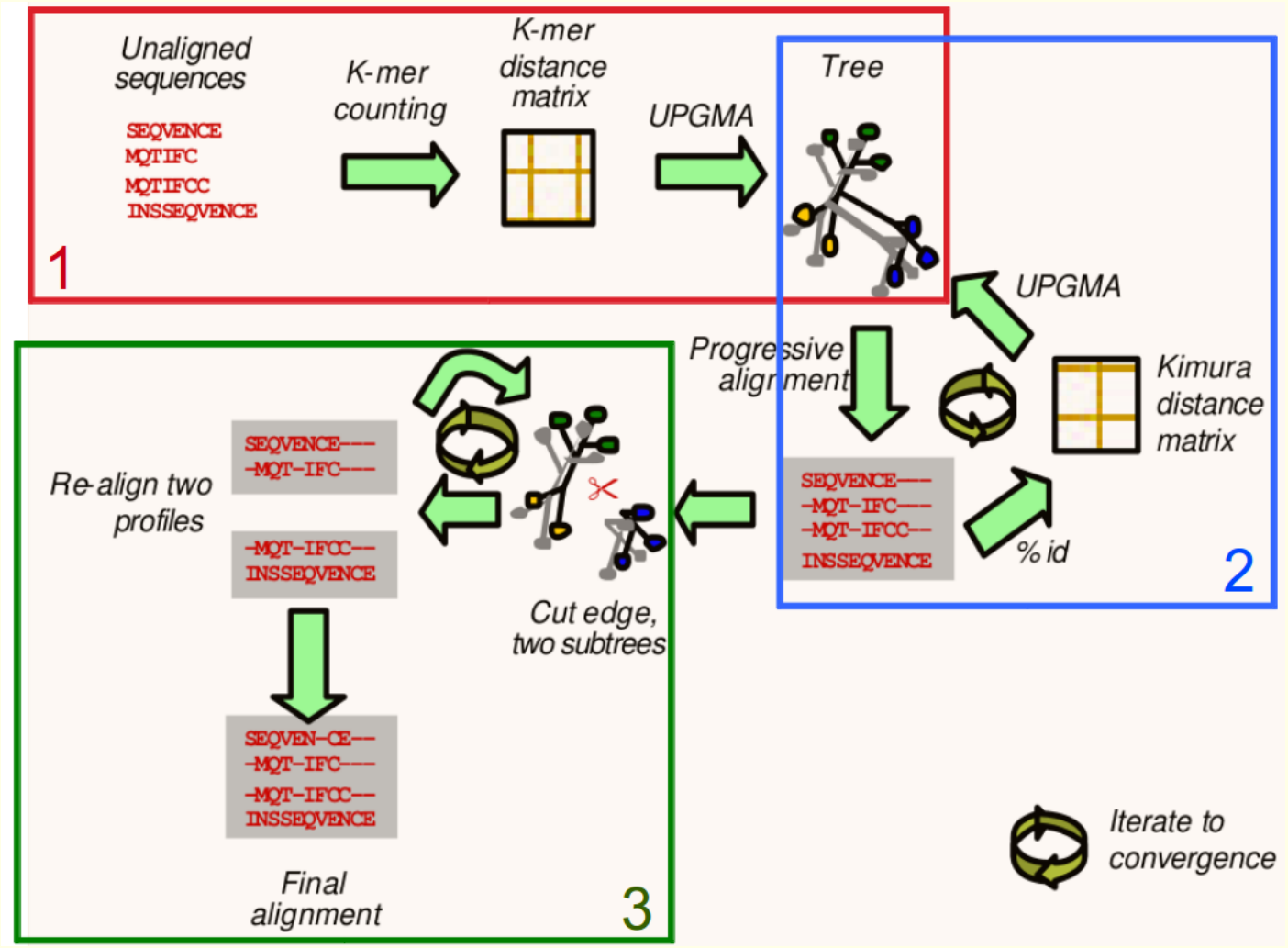
→ consistence : T-COFFEE, M-COFFEE
→ rafinement itératif : MUSCLE
→ consistence + rafinement itératif : MAFFT
→ HMM : Clustal Omega
→ consistence + HMM : ProbCons
→ similarités locales : KALIGN2, DIALIGN, POA

- T-COFFEE V5.05 (2007), [Notredame et al, 2000]
- Méthode :
  1. Construction d'une bibliothèque d'alignement 2 à 2, globaux et locaux, pour chaque paire de séquences du jeu d'entrée
  2. Phase de "scorage" pour donner un poids aux alignements puis aux caractères alignés (« *library extension* »)
  3. Cette bibliothèque est ensuite utilisée pour guider une phase d'alignement progressif pour trouver un MSA préservant la consistance des alignements 2 à 2
- Alignements de qualité mais lent pour de gros jeux de séquences
- Plusieurs formats de sortie : .aln (CLUSTALW), FASTA, PHYLIP et T-COFFEE + 2 formats d'arbres
- Originalité : permet de combiner différents résultats (exemple : CLUSTALW, DIALIGN et alignement structural)

T-COFFEE : versions et résultats

- Différentes versions spécifiques et permettant l'utilisation de données externes (autres programmes, structures, ...)
- M-COFFEE [Wallace et al., 2006] : méta-méthode qui combine les résultats de différentes méthodes (Muscle-Mafft-Kalign-Clustal Omega) en utilisant le principe de consistance pour produire un alignement consensus

- MUSCLE V3.8 (2010) [Edgar, R.C., 2004]



1. k-mer clustering
⇒ alignement progressif rapide

2.progressive alignment + tree refinement
⇒amélioration de l'alignement progressif

3.tree dependent refinement
⇒ raffinement itératif

- 3 stratégies d'utilisation (versions) de MUSCLE : compromis entre qualité et rapidité (étapes 1, 2, 3)
  I. MUSCLE, options par défaut (étapes 1 à 3) ⇒ qualité maximum
  II. MUSCLE-fast (étape 1) ⇒ rapidité maximum
  III. MUSCLE-prog (étapes 1 et 2) ⇒ compromis entre qualité et rapidité
- Options : ajout de séquences à un alignement existant, alignements de profils, ...
- Plusieurs formats de sortie : .CLUSTALW, FASTA, PHYLIP, HTML (ci-dessus), GCG MSF
- + les arbres guides (aprés la 1ere ou 2eme itération)

- MAFFT V7 (2013) [Katoh K and Standley D M, 2013]

- Même principe que MUSCLE
  → Étape 1 : alignement progressif rapide
  → Étape 2 : amélioration de l'alignement progressif
  → Étape 3 : rafinement itératif

- Alignement par paire (étape 1) :
  → Séquence réécrit dans un système où les acides aminés sont décrits par leur polarité et leur volume
  → Segments de similarité entre chaque paire de séquence repérés par une analyse par Transformée de Fourrier rapide (FFT)
  → Alignement par paire sur la base de ses segments (algorithme restreint de prog. dyn. global)

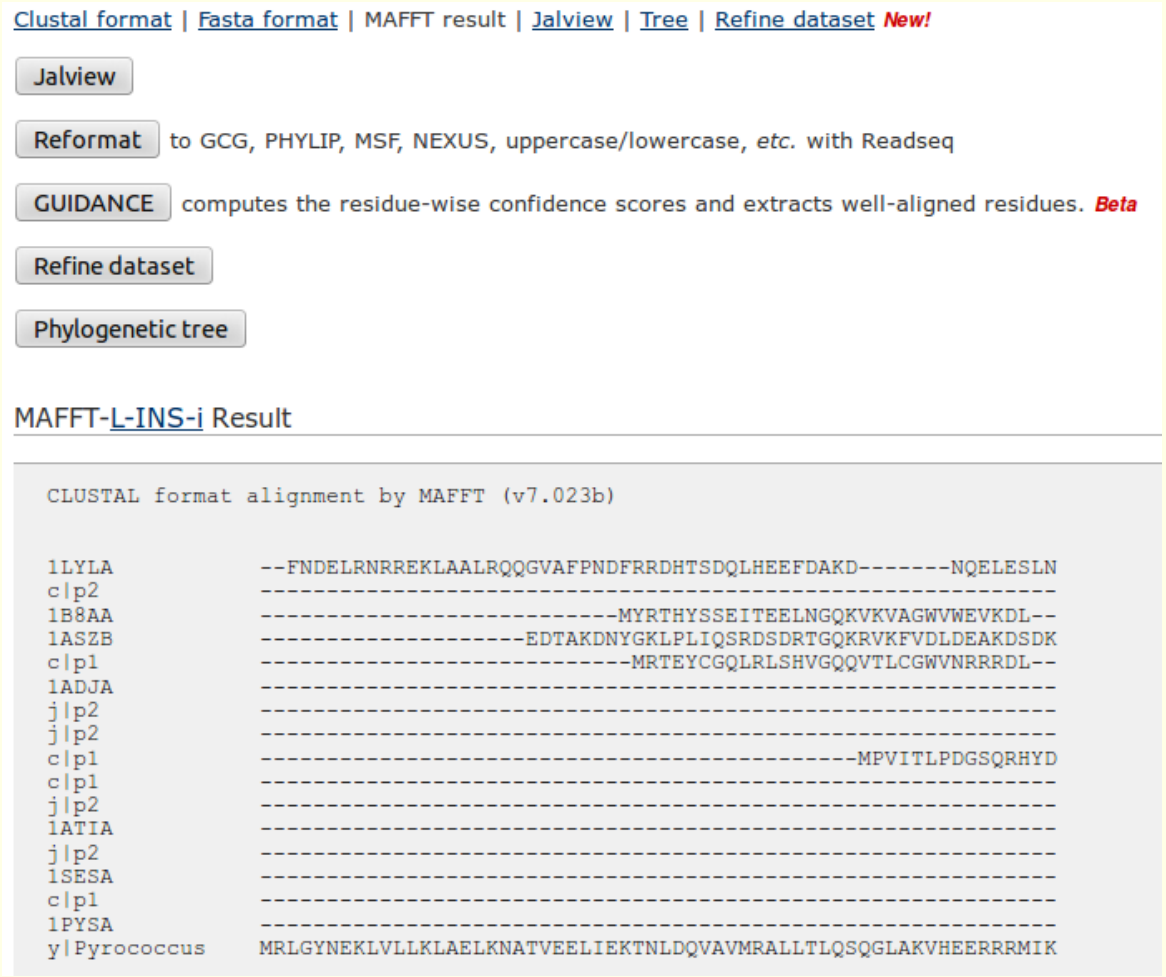
MAFFT - Utilisation

- 3 stratégies d'utilisation (versions) de MAFFT
  - I. FFT-NS-1/2 : méthode progressive (étape 1 / étape 1 et 2)
  - II. FFT/NW-NS-i : méthode de raffinements itératif utilisant le score WSP (étape 1 à 3)
  - III. L/E/G-INS-i : méthode de raffinements itératif utilisant le score WSP (weighted sum-of-pairs) et des scores basés sur la consistance (étape 1 à 3)

- Méthodes pour construire les alignements par paire
  - → Algorithme FFT (I et II)
  - → Algorithme d'alignement global de Needleman et Wunsch (III G-INS-i)
  - → Algorithme d'alignement local de Smith et Waterman (III L/E/-INS-i)

- Autres options possibles :
  - → Aligner un grand nombre de séquences ($\geq$10 000, PartTree)
  - → Ajout de séquences à un aln existant, aln de profils

MAFFT - Résultat



Clustal format | Fasta format | MAFFT result | Jalview | Tree | Refine dataset *New!*

Jalview

Reformat to GCG, PHYLIP, MSF, NEXUS, uppercase/lowercase, *etc.* with Readseq

GUIDANCE computes the residue-wise confidence scores and extracts well-aligned residues. *Beta*

Refine dataset

Phylogenetic tree

MAFFT-L-INS-i Result

```
CLUSTAL format alignment by MAFFT (v7.023b)


1LYLA           --FNDELRNRREKLAALRQQGVAFPNDFRRDHTSDQLHEEFDAKD-------NQELESLN
c|p2            ------------------------------------------------------------
1B8AA           ---------------------------MYRTHYSSEITEELNGQKVKVAGWVWEVKDL--
1ASZB           --------------------EDTAKDNYGKLPLIQSRDSDRTGQKRVKFVDLDEAKDSDK
c|p1            ----------------------------MRTEYCGQLRLSHVGQQVTLCGWVNRRRDL--
1ADJA           ------------------------------------------------------------
j|p2            ------------------------------------------------------------
j|p2            ------------------------------------------------------------
c|p1            ---------------------------------------------MPVITLPDGSQRHYD
c|p1            ------------------------------------------------------------
j|p2            ------------------------------------------------------------
1ATIA           ------------------------------------------------------------
j|p2            ------------------------------------------------------------
1SESA           ------------------------------------------------------------
c|p1            ------------------------------------------------------------
1PYSA           ------------------------------------------------------------
y|Pyrococcus    MRLGYNEKLVLLKLAELKNATVEELIEKTNLDQVAVMRALLTLQSQGLAKVHEERRRMIK
```

- Plusieurs formats de sortie possibles pour l'alignement multiple
- Visualisation de l'alignement et de l'arbre phylogénétique
- Rafinement du jeux de données

- Clustal Ω [Sievers et al, 11]
- Même principe que Clustal W mais :
  - → Construction de l'arbre guide = version optimisée de mBed ([Blackshields et al. 2010], séquences remplacées par des vecteurs de distance)
  - → Alignement progressif = alignement de profils HMM (Hhalign, [Söding, 2005])
- Options possibles :
  - → Alignement de profil externe (EPA) : utilisation d'un profil HMM existant (ou construit à partir d'un alignement existant) pour guider l'alignement du jeux de données
  - → Utiliser le principe de l'EPA dans un schéma itératif pour améliorer l'arbre guide et/ou l'alignement progressif
  - → Aligner un grand nombre de séquences ($\geq$10 000)
  - → Ajout de séquences à un aln existant, aln de profils

- Méthodes se basant sur des similarités locales et adaptées pour aligner des séquences très divergentes ou de longueurs différentes

- KALIGN2 [Lassmann et al., 2009]
  → Alignement progressif utilisant, pour le calcul des distances entre séquences, l'algorithme Wu-Manber de reconnaissance de chaîne (précision et vitesse améliorées)
  → Option : programmation dynamique pour aligner les profils

- POA [Lee et al., 2002]
  → Représentation en graphe d'un alignement multiple de séquences (PO-MSA)
  → Alignement par programme dynamique par paire de ces graphes

- DIALIGN [Morgenstern, 2004]

- DIALIGN 2.2.1 (oct. 2007) [Morgenstern, 2004]

→ Version améliorée : DIALIGN-TX [Subramanian et al., 2008]

- Spécialité : se base sur des similarités locales pour aligner des séquences très divergentes ou de longueur différentes

- Méthode :
  1. Repérer les régions alignées sans gaps dans les alignements 2 à 2 ($\sim$ diagonales continues dans un dotplot)
  2. Cherche un ensemble compatibles de diagonales pondérées pouvant produire un ali. et maximisant la somme des poids
  3. DIALIGN construit un alignement progressivement à partir de ces diagonales

- Schéma de score dans DIALIGN : score de l'aln = somme des scores des diagonales qui le composent $\Rightarrow$ pas de pénalités de gap

- Différentes versions de Dialign

→ DIALIGN-TX
  ▷ Dernière version de DIALIGN

→ Anchored DIALIGN
  ▷ Contraintes définies par l'utilisateur (points d'ancrage des régions alignées)

→ DIALIGN-PFAM
  ▷ Utilisation des informations de PFAM

→ CHAOS-DIALIGN
  ▷ Aln multiple ET aln par paire de séquences génomiques
  ▷ Utilisation d'un algorithme (CHAOS) qui identifie rapidement les régions de fortes similarités (= point d'ancrage des régions alignées par DIALIGN)

```
dog_il4       20565   AGAGCCTGGT CTGGAGCAAA GTTGATGTCT ACCTGTGCTT TCTTTAGCAG
hum_il4       21730   AGAGCCTGGT CTGGGGCAAA GTTGATGTCT ACCTGTGCTT TCTTTAGCAG
mus_il4       31390   ---------- ---GGGCCGA GCTGATGTCT ACCTGTGCTT TCTTTAGCAG

                      1111111111 1112222222 2222222222 2222222222 2222222222


dog_il4       20615   ATCAGATAta gGAG---TAC ACCAGTCGGG CATGAGCCTC TCCAGCTCTA
hum_il4       21780   ATCAGATAta gGAG---CAC ACCAGCCGGG CATGAGCCTC TCCAGCTCTA
mus_il4       31427   ATCAGATAta gccgcagCAC AGCAGTCGGG CATGAGCCTC TCCAACTCTA

                      2222222000 0222000333 3334444444 4444444444 4444444444


dog_il4       20662   AGGTGATGAT GACCAAGGCC AGTGTGGAGC CCTTGAacTG CAGCAGCTGG
hum_il4       21827   AGGTGATGAT GACCAAGGCC AGTGTGGAGC CCTTGAACTG CAGCAGCTGG
mus_il4       31477   AGGTGATGAT GACTAAAGCA AATGTGGAGC CCTTGAACTG CAGCAGCTGG

                      4444444444 4433666666 6666666666 6666662299 9999999999
```

- 2 formats de sortie possibles : FASTA et DIALIGN ci-dessus pour l'alignement multiple
- la liste des segments (alignements locaux sans gap par paire) PHYLIP
- la liste des points d'ancrage créés par CHAOS

- Les MSA peuvent servir à faire des recherches plus sensibles dans les banques de données qu'en utilisant 1 seule séq. requête

- Rappel du fonctionnement de PSI-BLAST
  1. On cherche des séquences similaires à la requête
  2. Construction d'un profil ou PSSM
  3. (en boucle) On cherche avec ce profil et on y intègre les nouvelles séquences trouvées
  4. Fin qd l'ensemble de séq. est stable ou le nb max d'itérations est dépassée

- PSI-BLAST utilise des alignement multiples différents de ceux obtenus avec les outils classiques

→ Normalement les MSA sont plus long que les séquences qu'ils contiennent mais PSI-BLAST supprime les régions qui nécessiteraient d'introduire des gap dans la séq. requête

- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
  - → Alignement exact (ex : MSA, DCA)
  - → Alignement progressif (ex : CLUSTALW)
  - → Amélioration de l'approche progressive (ex : T-COFFEE, MAFFT, DIALIGN,. . .)
  - → Autres méthodes d'alignement multiple
  - → Evaluation de MSA
- La suite ...

Approche utilisant des informations externes

- Approche récente $\Rightarrow$ nouvelle catégorie d'algorithmes

- Objectif = améliorer la signification biologique d'un aln en intégrant/utilisant des informations aditionnelles externes
  - $\rightarrow$ informations fonctionnelles (homologie) ou structurelles (stucture secondaire/tertiaire) extraites de bd biologiques
  - $\rightarrow$ contraintes définies par l'utilisateur

- Utilisation d'une des approches d'alignement multiple améliorée par différentes stratégies

- Une des stratégies : l'extension du principe de consistance
  - $\rightarrow$ utilisation d'un jeux de données comme point de départ pour explorer et retrouver toutes les informations connexes contenues dans les bd publics afin de constituer la librairie utilisée pour la construction de l'aln multiple global

- Exemple de programmes : Expresso, PROMALS3D, 3D-Coffee, ...

Approche orientée phylogénétiquement (1/3)

- Methodes classiques d'alignement (progressif)

  → essaient de minimiser le nombre d'insertion / délétion

  → produisent donc des alignements compacts
  ⇒ problèmes si régions non-homologues ou gaps informatifs pour l'analyse phylogénétique

- Développement d'approches guidées par la phylogénie ("Phylogeny-aware")

  → Objetif : obtenir un alignement "évolutivement" correct

  → Meilleure gestion des insertions / délétions

  → Alignements pour la phylogénie

  → Approche co-estimation de l'alignement et de la phylogénie

Approche orientée phylogénétiquement (2/3)

- SATé-II [Liu K et al., 2011]
  - → Alignement multiple avec raffinement itératif dérivé de MAFFT
  - → Estimation de l'alignement multiple qui supporte l'arbre le plus vraisemblable
  - → PASTA [Mirarab et al., 2014] : basé sur SATé-II pour aligner un grand volume de données (¿10 000 seq)

- PRANK [Löytynoja and Goldman, 2005 ; 2008 ; 2010]
  - → Programme d'alignement multiple probabiliste (HMM)
  - → Prend en compte les informations phylogénétiques contenues dans les indels
    - ⇒ indels = événements évolutifs distincts
    - ⇒ meilleur gestion des insertions
    - ⇒ évite la sur-estimation du nb d'évènement de délétion
  - → Améliore les alignements si insertions présentes mais lent

Approche orientée phylogénétiquement (3/3)

- PAGAN [Löytynoja et al., 2012] et ProGraphMSA [Szalkowski, 2012]

  → Améliorent le principe de PRANK (placement des insertions / délétion en prenant en compte les informations phylogénétiques)

  → Utilisent l'algorithmique des graphes pour améliorer la reconnaissance des régions non-homologues

  → Plus rapides que PRANK

  → PAGAN : pour l'ajout de (fragements) séquences à un alignement de référence (utilisation en métagénomique, NGS)

  → ProGraphMSA : pour l'alignement multiple global

- Lorsqu'on dispose de structures 3D, on peut comparer des protéines en les superposant
- Mais il faut connaître quels sont les a.a. qui se correspondent
- C'est le but des méthodes d'alignement multiple structurel
  → Alignement structurel des séquences
  → Superposition 3D de leur structure
- Logiciels, pour 2 ou plusieurs protéines : SSAP, DALI, STAMP, ...
- Méthode de STAMP

  1. Alignement structurel de 2 séquences
  2. Les structures sont superposées selon cet alignement
  3. Calcul d'une matrice de scores de similarité structurelle
  4. Prog. dyn. sur cette matrice pour trouver meilleur score & ali.
  5. Étant donné l'alignement on recommence jusqu'à convergence

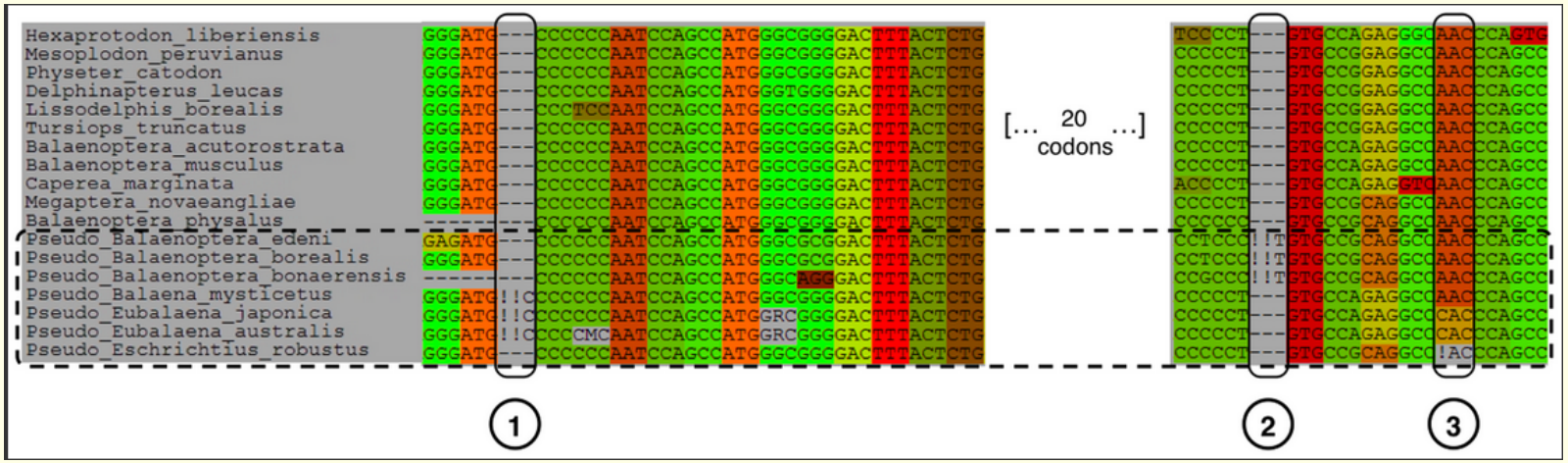
  Si plusieurs séquences, toutes les paires de structures sont comparées et un arbre est calculé, et on le suit comme pour un alignement progressif

- MACSE [Ranwez V et al, 11]

  → Même principe que Muscle

  → Prise en compte des décalages de phase et des codons Stop

    → Aln de jeux de données contenant des seq non-fonctionnelles sans perturber la structure du codon sous-jacent.



    → Détection de décalages de phase dans des seq de BD publiques

    → Aln de reads/contigues NGS contre une seq codante de référence

- Les composants algorithmiques principaux des programmes d'alignement multiple les plus utilisés

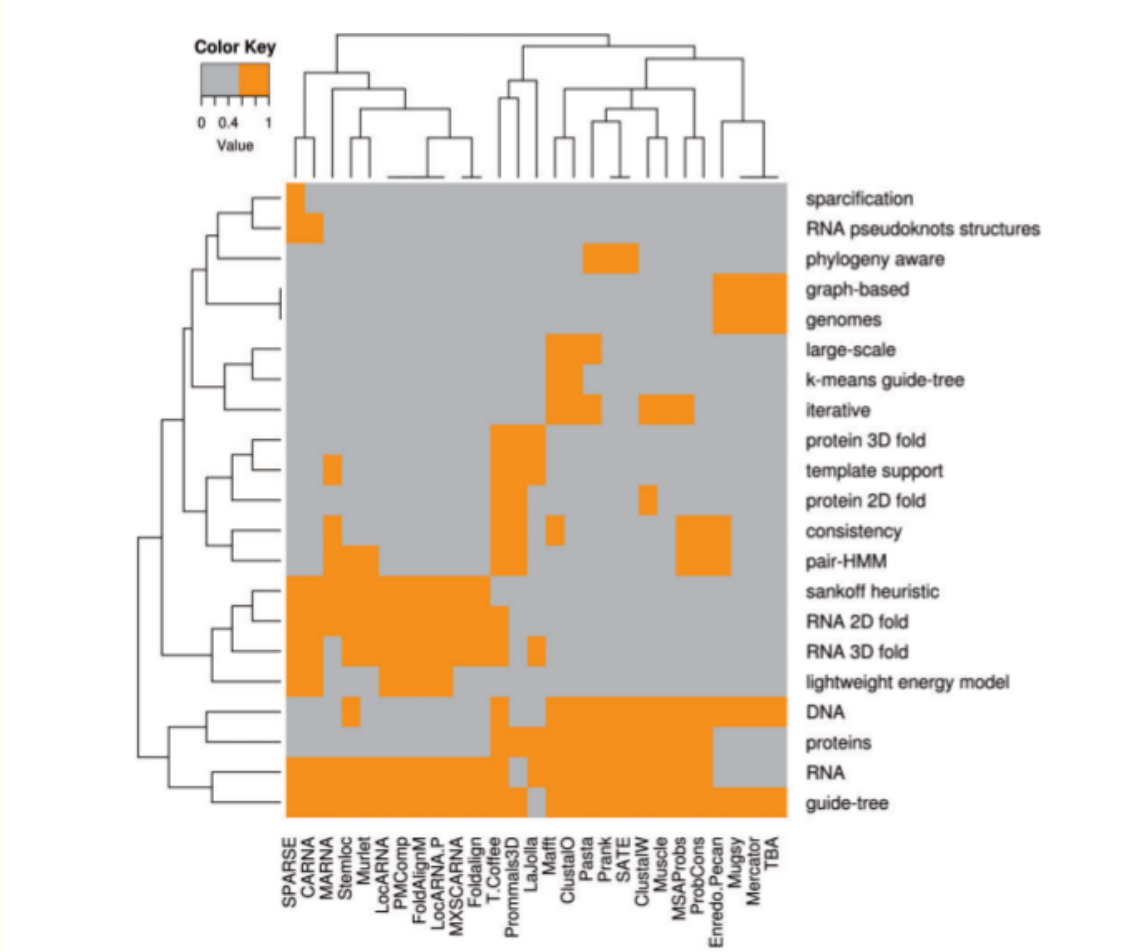


Figure 1. Main algorithmic components of the most widely used multiple aligners. On the heatmap, orange entries indicate a feature implemented in the considered method. Both the aligners and the components were clustered by similarity using the R-package.

[Chatzou *et al*, 2015]

- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
  - → Alignement exact (ex : MSA, DCA)
  - → Alignement progressif (ex : CLUSTALW)
  - → Amélioration de l'approche progressive (ex : T-COFFEE, MAFFT, DIALIGN,. . .)
  - → Autres méthodes d'alignement multiple
  - → Evaluation de MSA
- La suite ...

Sensibilité et sélectivité

- Notions utiles pour comparer des méthodes
- Exemple :

	Séq. membres famille	Séq. non membres
Séq. au-dessus du seuil	Vrais positifs	Faux positifs
Séq. en dessous du seuil	Faux négatifs	Vrai négatifs

- Sensibilité $= \frac{\text{Vrais positifs}}{\text{Vrais positifs}+\text{Faux négatifs}}$

  $\Rightarrow$ capacité à détecter les vraies instances de l'objet recherché (VP) = récupérer le max de séquences de la famille étudiée au risque d'avoir beaucoup d'intrus

  $\Rightarrow$ Favoriser la sensibilité = minimiser les faux négatifs

- Sélectivité $= \frac{\text{Vrais positifs}}{\text{Vrais positifs}+\text{Faux positifs}}$

  $\Rightarrow$ capacité à rejeter les fausses instances (FP) = récupérer le min d'intrus au risque de ne pas retenir certaines séquences de la famille étudiée

  $\Rightarrow$ Favoriser la sélectivité = minimiser les faux positifs

Benchmark

- Bases de données d'alignements multiples de référence : BaliBASE, PREFAB, OXBENCH, SABmark, SSSD, HomFam, . . .
- Utilisé pour attester de la qualité des logiciels d'alignement multiple
- 4 catégories : basé sur la simulation, la consistence, la structure, la phylogenie.
- Vue d'ensemble des benchmark principaux d'alignements multiples et leurs propriétés les plus pertinantes

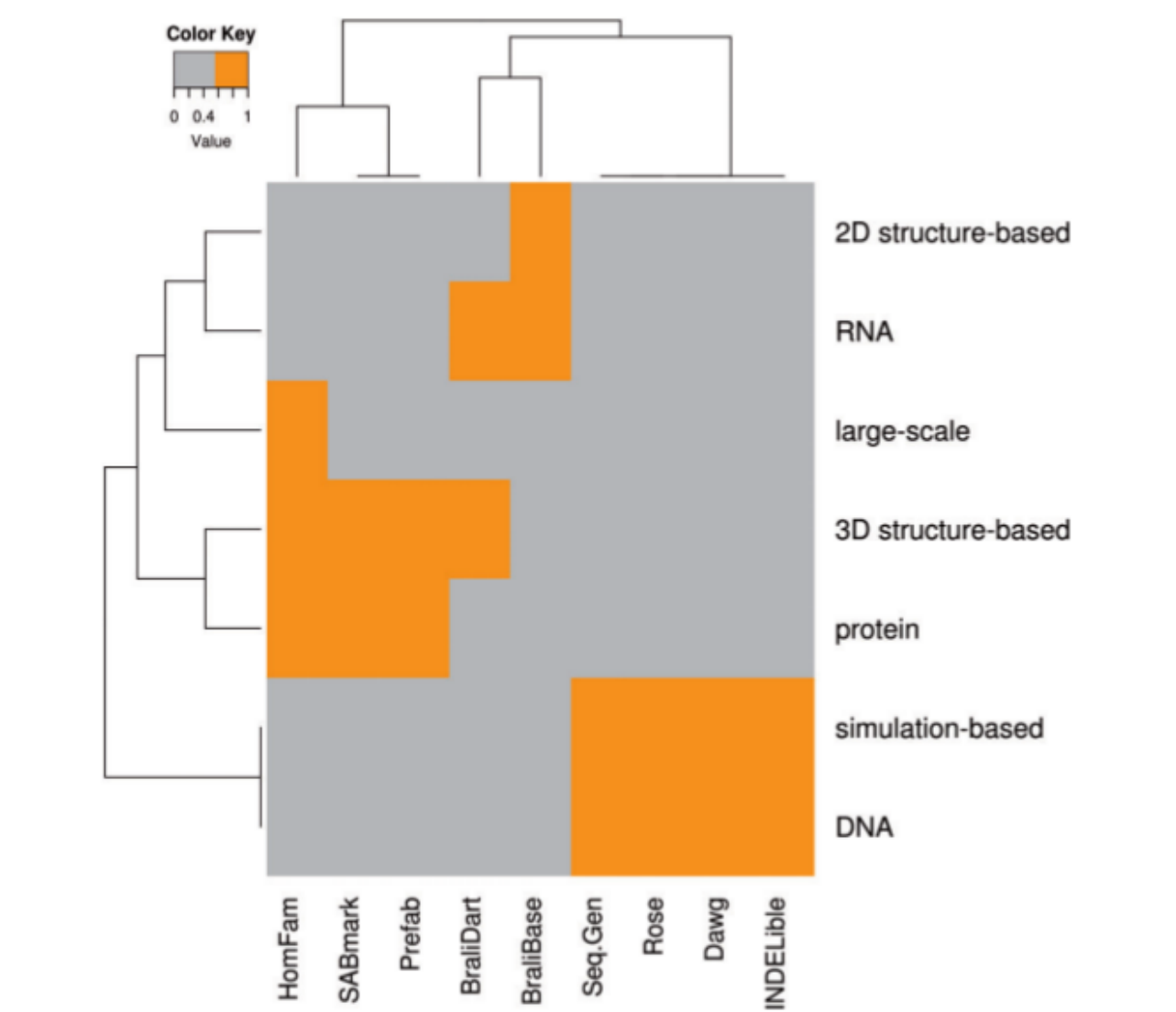


Figure 2. Main benchmark methods and their most relevant properties. On the heatmap, orange entries indicate a property describing a given method. Both properties and benchmarks were clustered by similarity.

[Chatzou *et al*, 2015]

- Bases de données d'alignements multiples de référence : BaliBASE, PREFAB, OXBENCH, SABmark, SSSD, HomFam, ...

- Utilisé pour attester de la qualité des logiciels d'alignement multiple

- BaliBASE 3.0 [Thompson et al., 2005]
  - → Plus de 200 familles de protéines
  - → Alignements de qualité (structure secondaire et vérification manuelle)
  - → Plusieurs ensembles de référence (pb d'aligenment spécifique)
    - ▷ Référence 1 : séquences équidistantes avec $\neq$ niveaux de conservation
    - ▷ Référence 2 : protéines homologues + 1 séquence orpheline
    - ▷ Référence 3 : sous-groupes avec - de 25% d'identité entre les groupes
    - ▷ Référence 4 : extensions N/C-terminales
    - ▷ Référence 5 : insertions internes
    - ▷ Référence 6-8 : régions transmembranaires / répétées, inversion de domaines
    - ▷ Référence 9 : motifs linéaires

- Pour évaluer la qualité d'un alignement
- Pour identifier les sites / les portions les plus informatifs
- 3 categories de méthodes

  → celles qui utilisent des informations structurelles pour évaluer la précision

  → celles qui dépendent d'un indice de conservation pour identifier les positions les plus susceptibles d'être correctes

  → celles dépendant d'une certaine forme d'instabilité numérique locale pour identifier les portions les plus stables d'un alignement multiple

- Qualité des MSA ⇒ qualité/précision des analyses (phylogénie, . . .)
- Mais différents problèmes exsitent dans les MSA
  - → Grandes zones de gap
  - → Séquences de longueurs différentes
  - → Régions aves un alignement ambigu
  - → Variabilité importante des caractères (régions trés divergentes)
  - → . . .

Nettoyage du MSA pour avoir un aln de meilleur qualité
Suppression des régions non informatives, mal conservées

- Logiciels : BMGE, Trimal, Gblocks, Noisy, TCS, . . .

- Gblocks [Castresana, 2000] : logiciel trés utilisé pour nettoyer un alignement (mais qui date un peu . . .)

- Classement des sites en 3 catégories en fonction de la variabilité observée à chaque site : Conservé, Non conservé, Très conservé

- Détermination d'un ensemble de blocs de site bien alignés en éliminant :
  → Les sites non conservés
  → Leur voisins jusqu'au 1er site très conservé
  → Les blocs de sites restant trop court

- Différents paramètres :
  → Moduler le niveau de stringence avec lequel les sites les plus variables sont éliminés
  → Paramètres par défaut : effet conservateur (élimine + de sites que les sites non homologues)

- trimAl [Capella-Gutiérrez et al., 2009]
- Adapté à de larges jeux de données
- Analyse site par site (ou pour un ensemble de sites fixés)
- Différents critères/paramètres
  - → Proportion de séquences ayant un gap (gap score)
  - → Niveau de similarité des acides aminés (similarity score)
  - → Niveau de consistance parmi plusieurs alignements (consistency score)
- Seuil de conservation = pourcentage minimum de sites de l'alignement original souhaité par l'utilisateur dans l'alignement nettoyé
  - → Si seuil non atteint avec les contraintes de score, alors contraintes de score diminuées
- Sélection automatique des paramètres possible (3 modes - différentes utilisations des scores de gap et de similarité)

- BMGE [Criscuolo and Gribaldo, 2010] = Block Mapping and Gathering with Entropy

- Identification d'un ensemble de colonnes variables dans l'alignement :
  $\rightarrow$ calcul d'une mesure d'entropie sur une fenêtre glissante
  $\rightarrow$ suppression de colonnes se situant au-dessus d'un certain seuil

- Mesure de l'entropie : prend en compte la similarité des acides nucléiques / aminés correspondant à un niveau de divergence défini par l'utilisateur

- Analyse site par site : chaque caractère est une observation indépendante

- Noisy [Dress et al., 2008]

- Objectif = déduire les colonnes qui sont phylogénétiquement non informatives en évaluant le degré des sites homoplastiques par rapport aux colonnes aléatoires.

- Méthodes

  → Evaluation de la distribution des caractères des colonnes suivant un ordonnancement circulaire des taxons

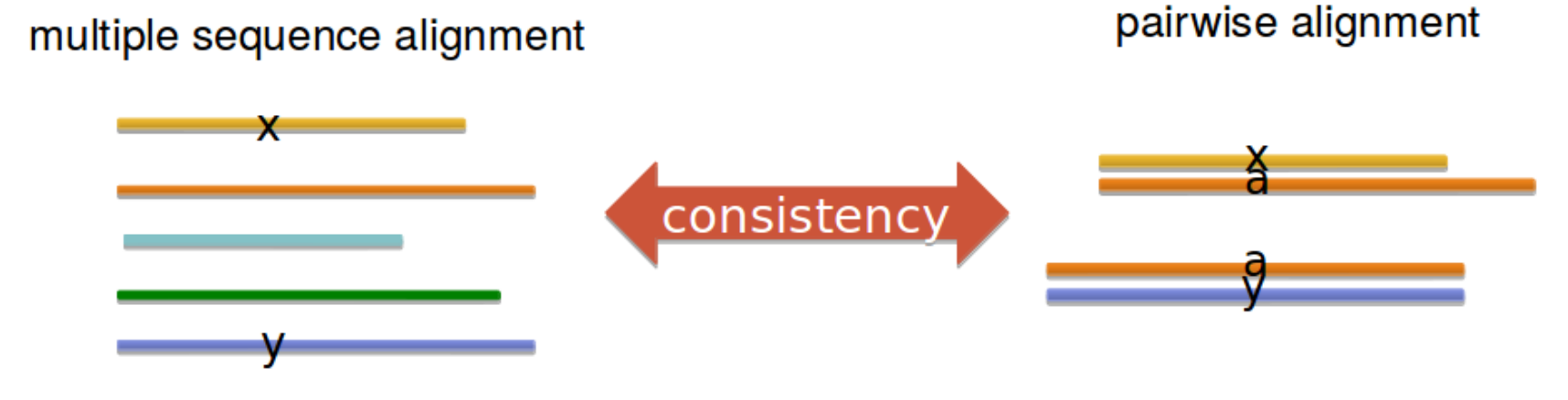
  → Identification des sites homoplastiques phylogénétiquement non informatifs

- Besoin d'un alignement d'au moins 15 séquences pour bien fonctionner

- TCS [Chang et al., 2014]
- Transitive Consistency Score
- Version étendue de la fonction de score de T-Coffee

  → Basée sur la consitence et le principe de transitivité

  → Utilisation de différentes librairies d'alignements par paire (ClustalW / Lalign, ProbCons pair-HMM, MAFFT / MUSCLE / Kalign, ...)



Indexes de qualité des alignements multiples et leurs caractéristiques

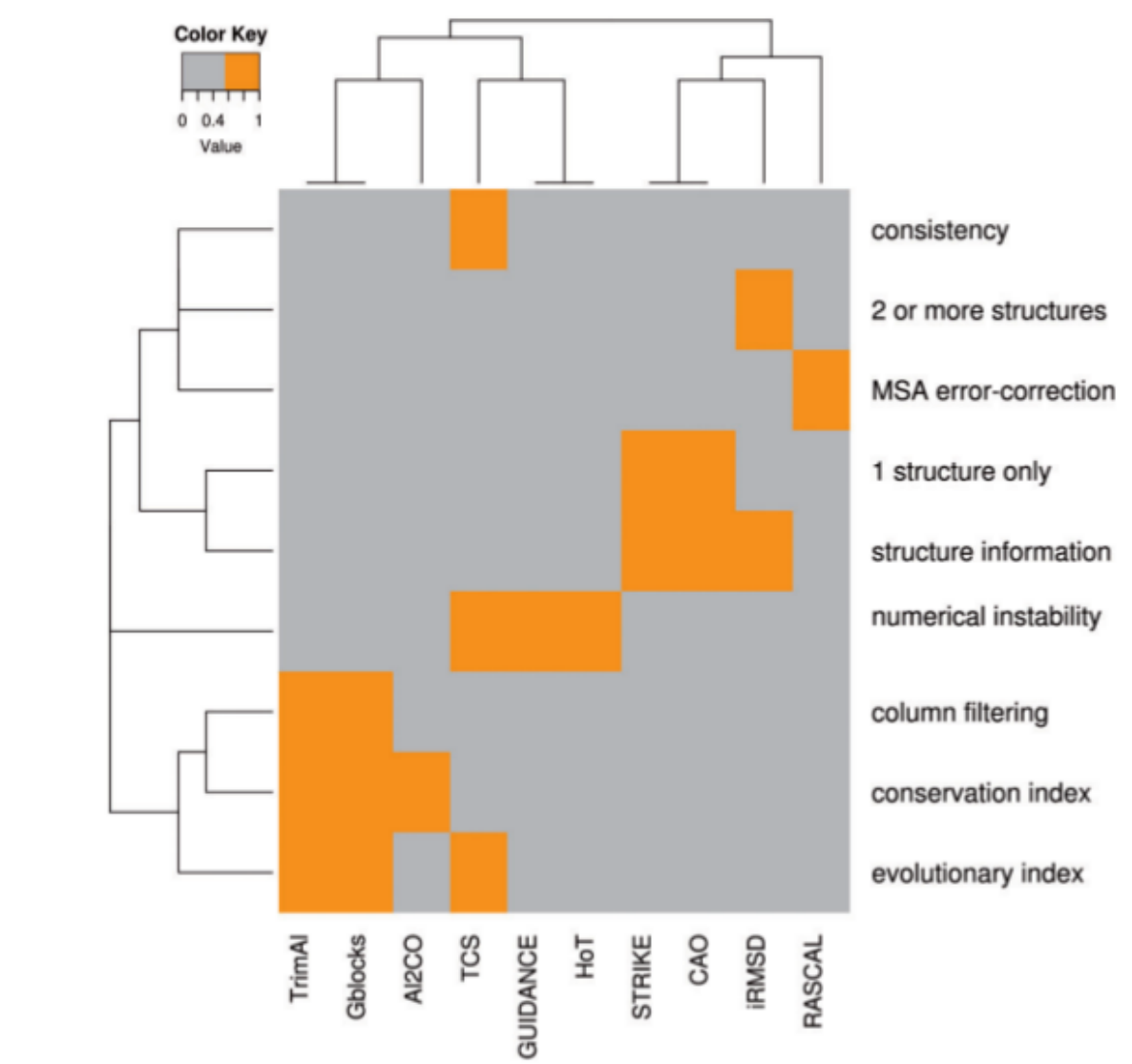


Figure 3. MSA quality indexes and their features. Features with zero are not used by the specific quality index.

[Chatzou *et al*, 2015]

Mais récemment, ...

Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. Tan G., Muffato M., Ledergerber C., Herrero J., Goldman N., Gil M., Dessimoz C. Syst Biol. 2015 Sep ;64(5) :778-91

≪ Although our results suggest that light filtering (up to 20% of alignment positions) has little impact on tree accuracy and may save some computation time, contrary to widespread practice, we do not generally recommend the use of current alignment filtering methods for phylogenetic inference. ≫

- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
- La suite ...