Alignement multiple

Hélène Touzet

helene.touzet@univ-lille.fr

CNRS, Bonsai, CRIStAL

- Entrée : k séquences

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

- Les séquences doivent être « apparentées »
- Ne convient pas pour l'alignement de deux séquences

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

De nombreuses applications

- identification de sites conservés
- identification de variants
- modélisation de motifs
- conception d'amorces de PCR
- phylogénie
- bioinformatique structurale (ARN et protéine)

Comment construire un alignement multiple

- plus difficile qu'il n'y parait
- taille des données
  - longueur des séquences
  - nombre de séquences
- diversité des séquences
  - distance évolutive variable
  - pression de sélection non uniforme
  - combinaison de similitudes globales et locales

Score d'un alignement multiple

- Score SP - sums of pairs : somme des scores de ses colonnes
- Comment scorer une colonne ?
  - adaptable à un nombre quelconque de lignes
  - indépendant de l'ordre
  - reflète la similarité

$$scoreSP\begin{pmatrix} c_1 \\ \vdots \\ c_k \end{pmatrix} = \sum_{1 \leq i < j \leq k} score(c_i, c_j)$$

$c_1, \ldots, c_k \in \mathcal{A} \cup \{-\}$ et $score(-,-) = 0$

$\mathcal{A}$: alphabet ADN ou acides aminés

```
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
```

Identité : $+1$
Substitution : -2
Indel : -3

Définition alternative équivalente

- α : alignement multiple pour les séquences $s_1, \ldots, s_k$
- α_{ij} : projection de l'alignement pour s_i et s_j

$$\text{scoreSP}(\alpha) = \sum_{1 \leq i < j \leq k} \text{score}(\alpha_{ij})$$

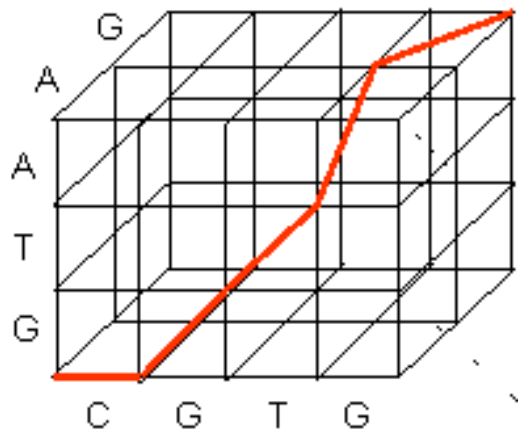
- retour à l'exemple

```
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
```

Identité : $+1$
Substitution : -2
Indel : -3

Algorithme exact pour l'alignement multiple

- trouver l'alignement multiple de score SP maximal
- changement de dimension
  - alignement deux à deux : chemin dans une matrice de dimension 2
  - alignement multiple de ℓ séquences : chemin dans une matrice de dimension ℓ
  - résolution par programmation dynamique




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

Exemple pour trois séquences, U, V et W

- matrice en dimension trois
- $\mathtt{Sim}(i,j,k)$: score optimal entre $U(1..i)$, $V(1..j)$ et $W(1..k)$.
- formule de récurrence :

$$
\begin{aligned}
\mathtt{Sim}(0,0,0) &= 0 \\
\mathtt{Sim}(0,0,k) &= \mathtt{Sim}(0,0,k-1) + SP(-,-,W(k)) \\
\mathtt{Sim}(0,j,0) &= \mathtt{Sim}(0,j-1,0) + SP(-,V(j),-) \\
\mathtt{Sim}(i,0,0) &= \mathtt{Sim}(i-1,0,0) + SP(U(i),-,-) \\
\mathtt{Sim}(0,j,k) &= \max \begin{cases} \mathtt{Sim}(0,j-1,k-1) + SP(-,V(j),W(k)) \\ \mathtt{Sim}(0,j-1,k) + SP(-,V(j),-) \\ \mathtt{Sim}(0,j,k-1) + SP(-,-,W(k)) \end{cases} \\
\mathtt{Sim}(i,0,k) &= \max \begin{cases} \mathtt{Sim}(i-1,0,k-1) + SP(U(i),-,W(k)) \\ \mathtt{Sim}(i-1,0,k) + SP(U(i),-,-) \\ \mathtt{Sim}(i,0,k-1) + SP(-,-,W(k)) \end{cases} \\
\mathtt{Sim}(i,j,0) &= \max \begin{cases} \mathtt{Sim}(i-1,j-1,k) + SP(U(i),V(j),-) \\ \mathtt{Sim}(i-1,j,k) + SP(U(i),-,-) \\ \mathtt{Sim}(i,j-1,k) + SP(-,V(j),-) \\ \mathtt{Sim}(i,j,k-1) + SP(-,-,W(k)) \end{cases}
\end{aligned}
$$

$$\mathtt{Sim}(i,j,k) \quad = \quad \max \begin{cases} \mathtt{Sim}(i-1,j-1,k-1) + SP(U(i),V(j),W(k)) \\ \mathtt{Sim}(i-1,j-1,k) + SP(U(i),V(j),-) \\ \mathtt{Sim}(i-1,j,k-1) + SP(U(i),-,W(k)) \\ \mathtt{Sim}(i-1,j,k) + SP(U(i),-,-) \\ \mathtt{Sim}(i,j-1,k-1) + SP(-,V(j),W(k)) \\ \mathtt{Sim}(i,j-1,k) + SP(-,V(j),-) \\ \mathtt{Sim}(i,j,k-1) + SP(-,-,W(k)) \end{cases}$$

Algorithme exact – complexité

- n : longueur des séquences
- 2 séquences : $O(n^2)$ en temps et en espace
- 3 séquences : $O(n^3)$ en temps et en espace
- ℓ séquences, $s_1, \ldots, s_\ell$
  - $\texttt{Sim}(i_1, \ldots, i_\ell)$: score optimal entre les ℓ préfixes $s_1(1..i_1), \ldots, s_\ell(1..i_\ell)$
  - table de taille n^ℓ
  - temps de calcul d'une case : dépend de $2^\ell - 1$ cases précédentes
  - temps de calcul de chaque scoreSP candidat : $\ell(\ell-1)/2$
  - temps de calcul exponentiel : $O(n^\ell 2^\ell \ell^2)$
- le problème de décision associé est NP-complet

Comment construire un alignement multiple en pratique

- recours à des heuristiques

  du grec `heuriskein`, trouver $\rightarrow$ qui sert à la découverte

- méthodes progressives
  - construction d'un arbre guide généré à partir des comparaisons deux à deux des séquences
  - incorporation des séquences suivant cet arbre pour former l'alignement multiple
  - exemples : heuristique en étoile, `CLUSTAL W`, `CLUSTAL Omega`, `kalign`

- méthodes itératives
  - ajoût d'étapes de correction de l'alignement
  - exemples : `T-coffee`, `MUSCLE`, `MAFFT`

- Ce qui change entre les différents logiciels
  - la méthode de comparaison des séquences deux à deux globale/locale, exacte/approchée
  - la méthode de construction de l'arbre guide
  - les étapes de correction
- Autant d'alignements que de programmes

Alignement progressif

```
> s1
cgatgagtcattgtgactg
> s2
cgagccattgtagctactg
> s3
cgaccattgtagctacctg
> s4
cgatgagtcactgtgactg
```

`indel:-2, substitution:-1, identité:1`

- comparaison des séquences deux par deux
- combinaison des alignements

Heuristique en étoile

- sélection d'une séquence centrale
- construction de l'alignement multiple, en partant de la séquence centrale, puis en incorporant une à une les autres séquences

- Étape 1 : Alignements globaux de toutes les séquences deux par deux

```
s1 cgatgagtcattgt-g--actg   s2 cgagccattgtagcta-ctg
   ||| |   ||||||  |  ||||      ||| ||||||||||||| |||
s2 cga-g--ccattgtagctactg   s3 cga-ccattgtagctacctg

s1 cgatgagtcattg-tgactg     s2 cga-g--ccattgtagctactg
   |||  | |  | | |  |||         ||| |   || ||| |  ||||
s3 cgacca-ttgtagctacctg     s4 cgatgagtcactgt-g--actg

s1 cgatgagtcattgtgactg      s3 cgaccattgtagctacctg
   |||||||||| ||||||||         |||  | | |   |  |||
s4 cgatgagtcactgtgactg       s4 cgatgagtcactgtgactg
```

Tableau des scores

	s_1	s_2	s_3	s_4
s_1		2	0	17
s_2	2		14	0
s_3	0	14		-1
s_4	17	0	-1	

ℓ séquences

$\downarrow$

$\ell(\ell-1)/2$ alignements

- Étape 2 : sélection de la séquence centrale à partir du tableau des scores : séquence qui maximise la somme des similarités avec l'ensemble des autres séquences

	s_1	s_2	s_3	s_4	
s_1		2	0	17	19
s_2	2		14	0	16
s_3	0	14		-1	13
s_4	17	0	-1		16

- Étape 3 : construction de l'alignement multiple par juxtaposition des alignements deux à deux avec la séquence centrale

```
s1 cgatgagtcattgt-g--actg  s1 cgatgagtcattg-tgactg
   ||| |   |||||| |  ||||     |||  | |  | | |  |||
s2 cga-g--ccattgtagctactg  s3 cgacca-ttgtagctacctg
```

```
s1 cgatgagtcattgtgactg
   |||||||||| ||||||||
s4 cgatgagtcactgtgactg
```

Alignement multiple

```
s1 cgatgagtcattg-t-g--actg
s2 cga-g--ccattg-tagctactg
s3 cgacca-ttgtagct-ac--ctg
s4 cgatgagtcactg-t-g--actg
```

L'intégration d'une nouvelle séquence se fait en prenant la séquence centrale comme guide. C'est toujours possible en étirant les gaps de l'alignement multiple courant.
Once a gap, always a gap

Approches avec un arbre guide

```
s1  cgatgagtcattgt-g--actg
    ||| |   |||||| |  ||||
s2  cga-g--ccattgtagctactg
```

```
s2  cgagccattgtagcta-ctg
    ||| ||||||||||||| |||
s3  cga-ccattgtagctacctg
```

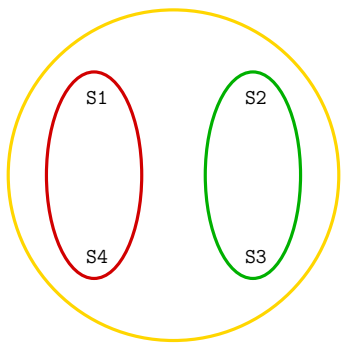
```
s1  cgatgagtcattg-tgactg
    |||  | |  | | |  |||
s3  cgacca-ttgtagctacctg
```

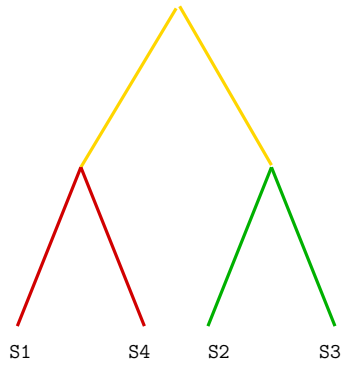
```
s2  cga-g--ccattgtagctactg
    ||| |   || ||| |  ||||
s4  cgatgagtcactgt-g--actg
```

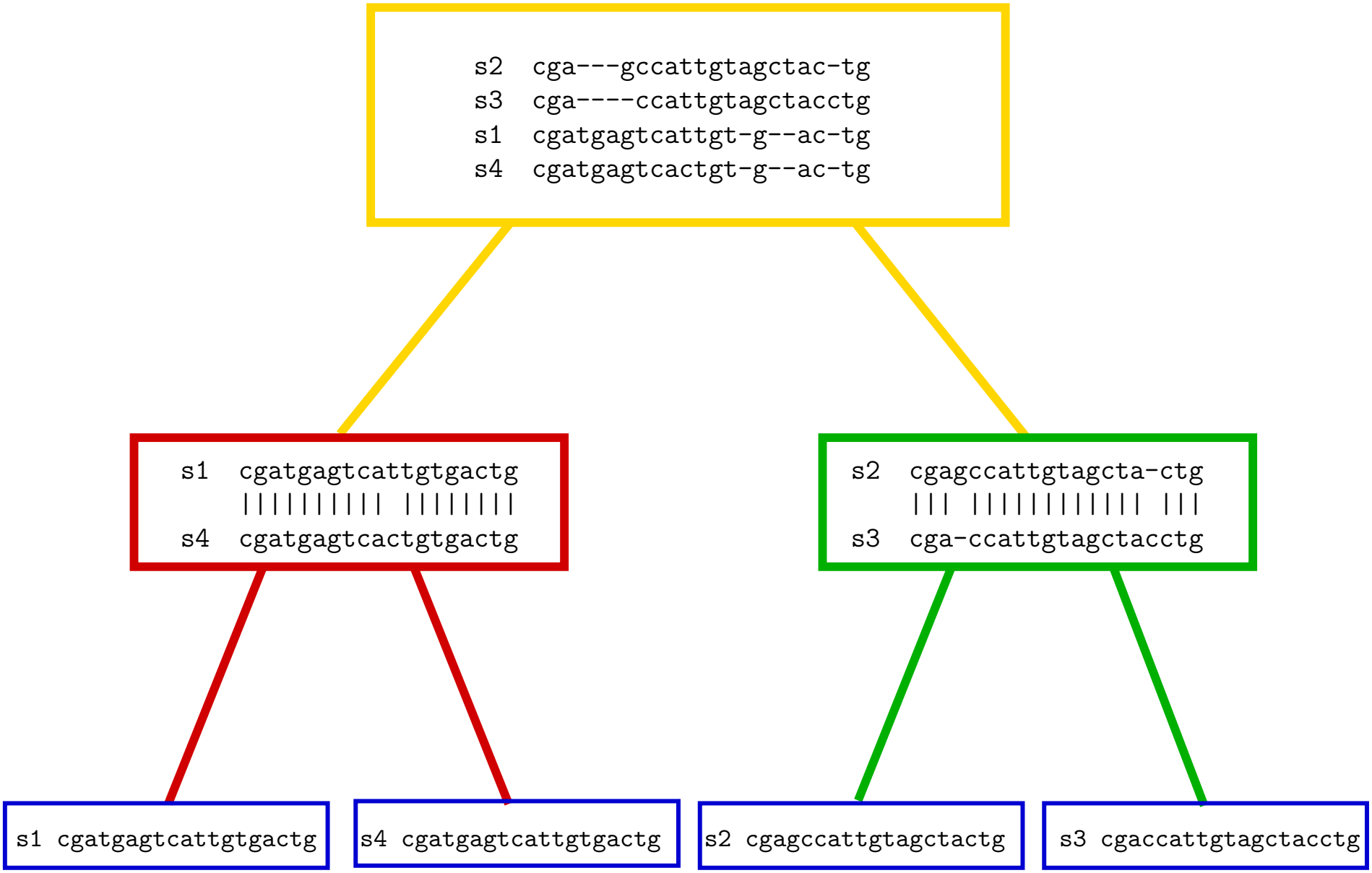
```
s1  cgatgagtcattgtgactg
    |||||||||| ||||||||
s4  cgatgagtcactgtgactg
```

```
s3  cgaccattgtagctacctg
    |||  | |     |  |||
s4  cgatgagtcactgtgactg
```

	s1	s2	s3	s4
s1		2	0	17
s2	2		14	0
s3	0	14		−1
s4	17	0	−1	

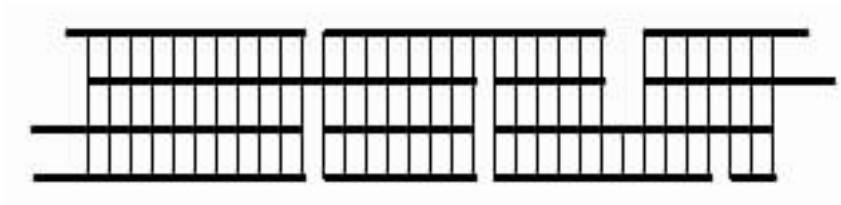





s2 cga---gccattgtagctac-tg
s3 cga----ccattgtagctacctg
s1 cgatgagtcattgt-g--ac-tg
s4 cgatgagtcactgt-g--ac-tg
s1 cgatgagtcattgtgactg
s4 cgatgagtcactgtgactg
s2 cgagccattgtagcta-ctg
s3 cga-ccattgtagctacctg
s1 cgatgagtcattgtgactg
s4 cgatgagtcattgtgactg
s2 cgagccattgtagctactg
s3 cgaccattgtagctacctg

Clustal W

- Cluster + Alignment
- alignement global des séquences deux à deux



- construction de l'arbre guide avec l'algorithme de Neighbor-Joining
- série d'optimisations pour l'alignement de séquences protéiques

  Adaptation des matrices de similarité au fil de l'algorithme en fonction de la divergence des séquences à aligner, pénalités de gaps spécifiques à chaque résidu, réduites dans les régions hydrophiles

Clustal Omega

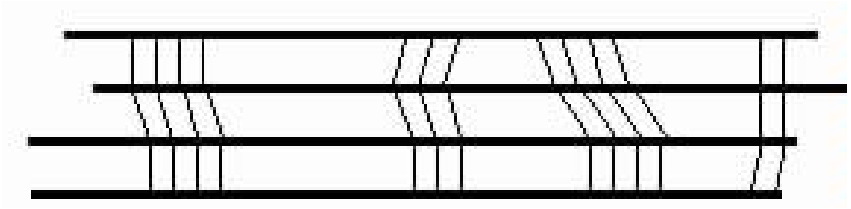
- successeur de Clustal W
- nouvelle méthode pour construire l'arbre guide
  $\rightarrow$ évite de considérer les alignements entre toutes les paires de séquences
  $\rightarrow$ plus rapide
- utilisation de Modèles de Markov cachés (HMM) pour fusionner les alignements à partir de l'arbre guide
  $\rightarrow$ meilleure qualité

Comparaison des séquences à aligner

- X : ensemble initial de séquences
- R : sous-échantillon aléatoire de X
  - on tire au sort $\log_2(|X|)$ séquences de X, de longueurs représentative (tri sur la longueur puis pas fixe)
  - pour chaque séquence r de R, on calcule l'alignement avec toutes les séquences restantes de X, et on choisit une séquence ℓ particulièrement éloignée, que l'on ajoute à R
- pour chaque séquence s appartenant à X
  - on calcule $F(s) = (d(r_1, s), \ldots, d(r_{|R|}, s))$, $r_1, \ldots r_{|R|} \in R$
  - les différents vecteurs F sont comparés avec la distance euclidienne

kalign

- alignement local des séquences deux à deux à partir de motifs communs



- arbre guide construit avec la même heuristique que Clustal Omega

- rapide et peu gourmand en mémoire
  $\rightarrow$ adapté aux jeux de données volumineux

```
>S1
GARFIELD THE LAST FAT CAT
>S2
GARFIELD THE FAST CAT
>S3
GARFIELD THE VERY FAST CAT
>S4
THE FAT CAT
```

```
CLUSTAL multiple sequence alignment by Clustal Omega

S4        --------THEFAT-----CAT
S3        GARFIELDTHEVERYFASTCAT
S1        GARFIELDTHELASTFAT-CAT
S2        GARFIELDTHEFASTCAT----
                    ***.
```

Alignements 2 à 2 pour KALIGN

```
S1  GARFIELDTHELASTFA-TCAT
    ||||||||||||    || ||||
S2  GARFIELDTHE----FASTCAT
```

```
S2  GARFIELDTHE----FASTCAT
    ||||||||||||    ||||||
S3  GARFIELDTHEVERYFASTCAT
```

```
S1  GARFIELDTHELASTFA-TCAT
    ||||||||||||    || |||
S3  GARFIELDTHEVERYFASTCAT
```

```
S2  GARFIELDTHEFASTCAT
              ||||| ||||
S4  --------THEFA-TCAT
```

```
S1 GARFIELDTHELASTFATCAT
             |||    ||||||
S4 --------THE----FATCAT
```

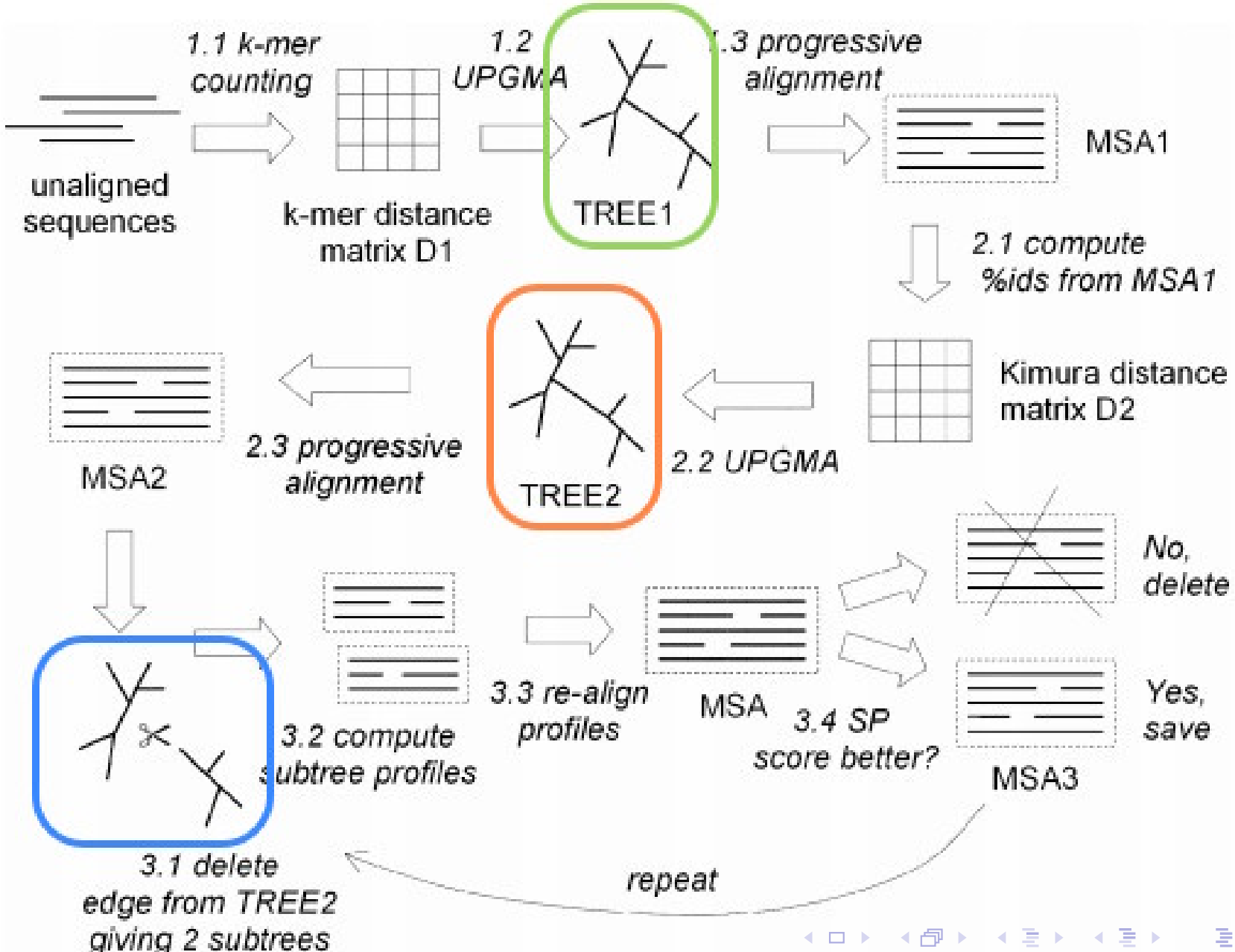
```
S3  GARFIELDTHEVERYFASTCAT
              |||    || ||||
S4  --------THE----FA-TCAT
```

```
CLUSTAL multiple sequence alignment by KALIGN

S1      GARFIELDTHELASTFA-TCAT
S2      GARFIELDTHE----FASTCAT
S3      GARFIELDTHEVERYFASTCAT
S4      --------THE----FA-TCAT
                  ***    ** ****
```

MUSCLE

MUltiple Sequence Comparison by Log-Expectation



MUSCLE

MUltiple Sequence Comparison by Log-Expectation

- TREE1 : construit très rapidement, sur la base de k-mers communs (sans alignement)
- TREE2 : vrai arbre, avec des distances obtenues à partir de TREE1
- TREE3 : détection de sous-groupes de séquences

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


S3              GARFIELDTHEVERYFASTCAT
S4              --------THE----FA-TCAT
S1              GARFIELDTHELASTFA-TCAT
S2              GARFIELDTHE----FASTCAT
                         ***    ** ****
```

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.

https://www.ebi.ac.uk/Tools/msa/

Comparaison des performances

BaliBASE

base de données d'alignements multiples de protéines basés sur la structure secondaire (+ 150 familles)

	BB1.1	BB1.2	BB2	BB3	BB4	BB5	temps
Clustal Ω	0.358	0.789	0.450	0.575	0.579	0.533	539.91
T-Coffee	0.410	0.848	0.402	0.491	0.545	0.587	81041.50
Kalign	0.365	0.790	0.360	0.476	0.504	0.435	21.88
MUSCLE	0.318	0.804	0.350	0.409	0.450	0.460	789.57
MAFFT	0.258	0.749	0.316	0.425	0.480	0.496	68.24

BB1.1 BB1.2 : séquences équidistantes avec différents niveaux de conservation
BB2 : protéines homologues + 1 séquence orpheline
BB3 : sous-groupes avec moins de 25% d'identité entre les groupes
BB4 : extensions N/C-terminales
BB5 : insertions internes

Prefab - Protein Reference Alignment Benchmark

%identité	< 20	$20 - 40$	$40 - 70$	> 70	temps (sec)
MAFFT	0.569	0.876	0.961	0.979	4544
T-Coffee	0.558	0.865	0.950	0.972	175789
Clustal Ω	0.535	0.866	0.967	0.980	1698
MUSCLE	0.507	0.850	0.946	0.976	2068
Kalign	0.474	0.817	0.957	0.979	80

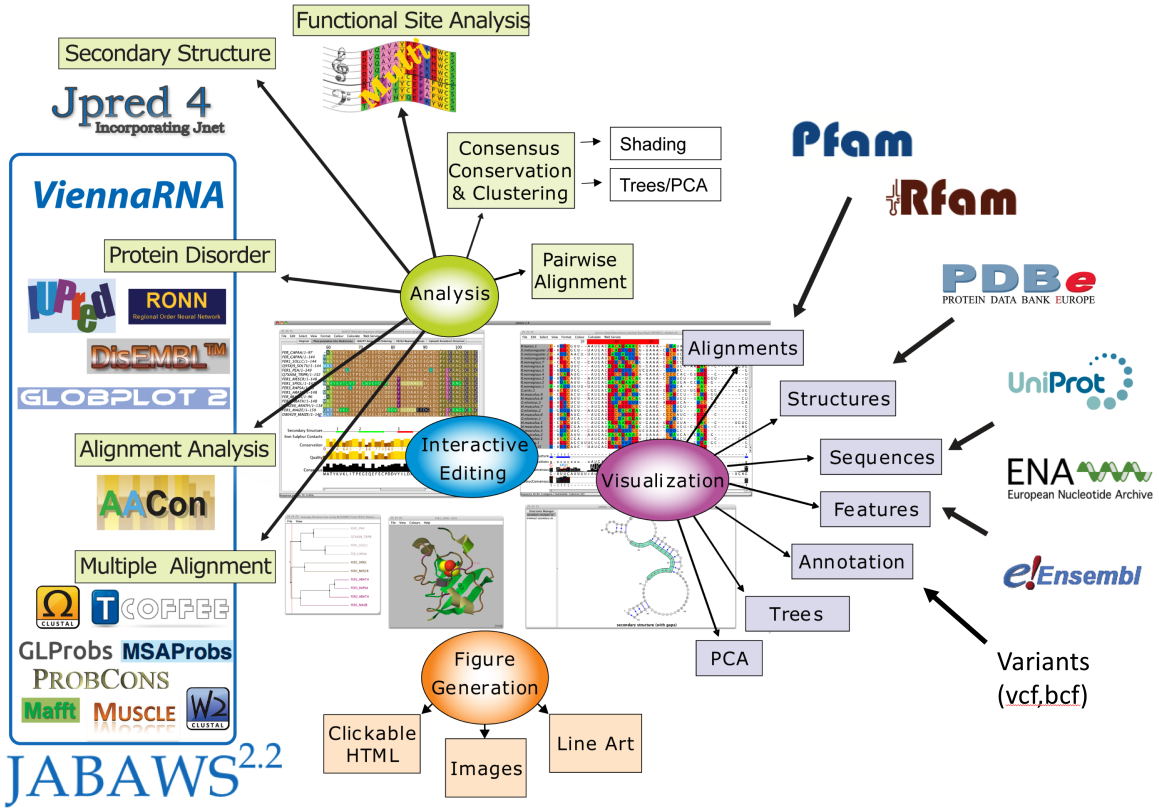
Paramétrages différents

Usage avancé

- Plus de fonctionnalités sur les sites d'origine que sur l'EBI
  - Clustal W, clustal Ω : `http://www.clustal.org`
  - T-coffee : `http://tcoffee.crg.ca`
  - MAFFT : `https://mafft.cbrc.jp/alignment/software`
  - Muscle : `https://www.drive5.com/muscle`
  - kalign : `https://msa.sbc.su.se/cgi-bin/msa.cgi`

- Exploration des données avec un éditeur d'alignement multiple

jalview



- Installation locale, JAVA
- Très complet : edition, structure, annotation, variants, phylogénie
- Un peu complexe à prendre en main
- `https://www.jalview.org`

- Mview : colorisation, minimal, en ligne à l'EBI
  `https://www.ebi.ac.uk/Tools/msa/mview`

- seaview : orienté phylogénie
  `http://doua.prabi.fr/software/seaview`

- bioedit
  uniquement sous Windows

Séquence consensus

```
sequence 1          C C T A T G G G C T A C A A G C C A
sequence 2          C A T C C T G T C C C T A T G G A A
sequence 3          T C - - A A G G C C G C A T G - A A
sequence 4          T C - - A A G G C A G C A T G G A A

consensus 100%      N N . . N N G N C N N N A N G N N A
consensus  70%      N C . . N N G G C N N C A T G G A A
majoritaire         N C . . A A G G C C G C A T G G A A
```

Code IUPAC pour l'ADN

- International Union of Pure and Applied Chemistry
- alphabet à 15 lettres qui décrit toutes les combinaisons de nucléotides possibles

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

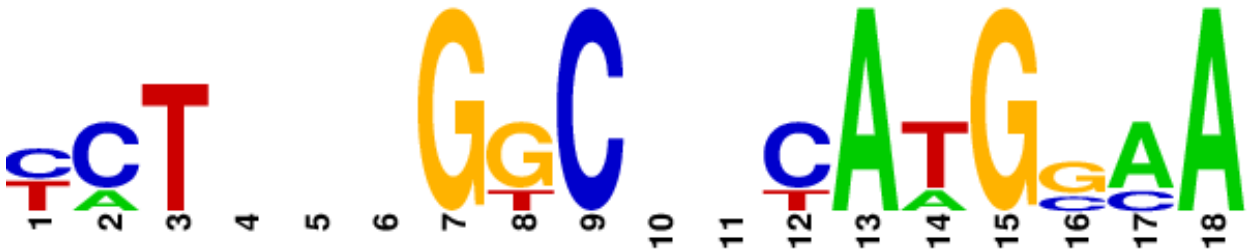
```
sequence 1      C C T A T G G G C T A C A A G C C A
sequence 2      C A T C C T G T C C C T A T G G A A
sequence 3      T C - - A A G G C C G C A T G - A A
sequence 4      T C - - A A G G C A G C A T G G A A

IUPAC           Y M - - H D G K C H V Y A W G - M A
```

Sequence logo

```
sequence 1      C C T A T G G G C T A C A A G C C A
sequence 2      C A T C C T G T C C C T A T G G A A
sequence 3      T C - - A A G G C C G C A T G - A A
sequence 4      T C - - A A G G C A G C A T G G A A

IUPAC           Y M - - H D G K C H V Y A W G - M A
```



https://weblogo.berkeley.edu

Formule pour le sequence logo

```
sequence 1      C C T A T G G G C T A C A A G C C A
sequence 2      C A T C C T G T C C C T A T G G A A
sequence 3      T C - - A A G G C C G C A T G - A A
sequence 4      T C - - A A G G C A G C A T G G A A
```

Pour une colonne donnée :

- f_b : proportion de la base b dans la colonne
- H : hauteur d'une colonne

$$H = \log_2(4) + \sum_{b=1}^{4} f_b \log_2(f_b)$$

- proportion de chaque base dans la colonne : f_b
- exemples

Messages de conclusion

- les logiciels d'alignement multiple ont chacun leurs spécificités
- critères de choix
  - taille des données
  - nature des données
- usages avancés en fonction de l'application visée

Quelques exercices d'application

https://helene-touzet.cnrs.fr/Teaching/October2020/TP_msa.html