Alignements Multiples

Anne Lopes

Analyse de Séquences Génomiques

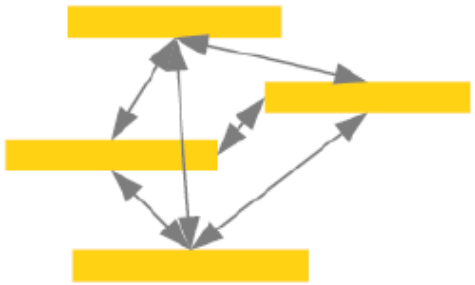
Mars 2019

Quelle est la similarité entre ces deux séquences?

Needleman & Wunsch (global)
Smith & Waterman (local

Quelles sont les séquences les plus similaires à la mienne?

FASTA, BLAST



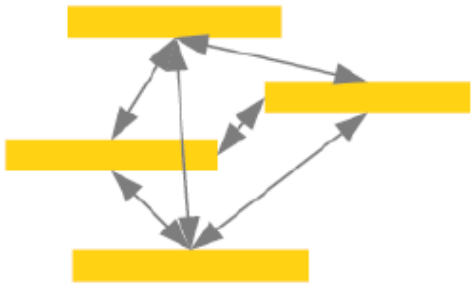
Quelles sont les points communs entre toutes les séquences?

Quelle est la similarité entre ces deux séquences?

Needleman & Wunsch (global)
Smith & Waterman (local

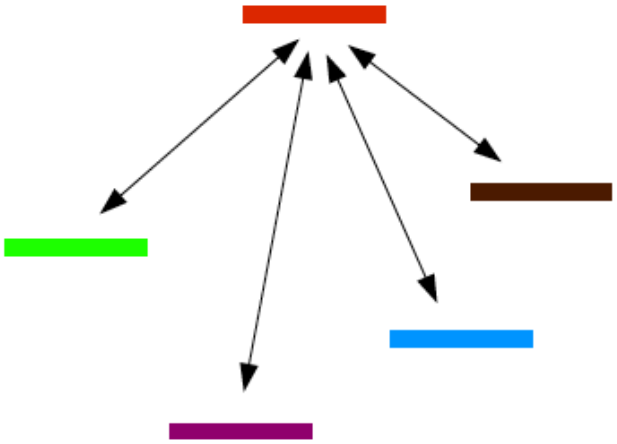
Quelles sont les séquences les plus similaires à la mienne?

FASTA, BLAST



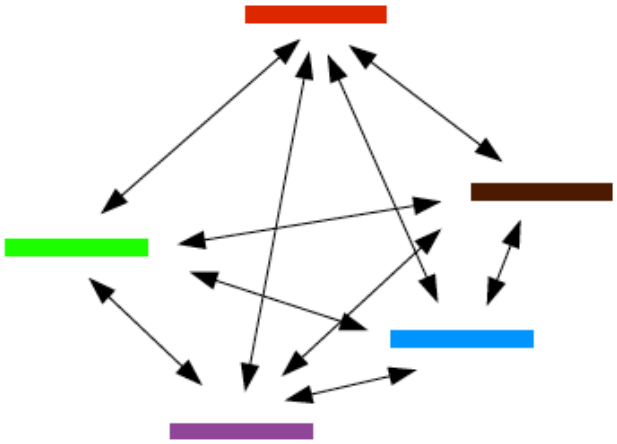
Quelles sont les points communs entre toutes les séquences?

Ce qu’on sait faire



alignement de paires

Ce qu’on veut faire



alignement simultané

intérêt ?

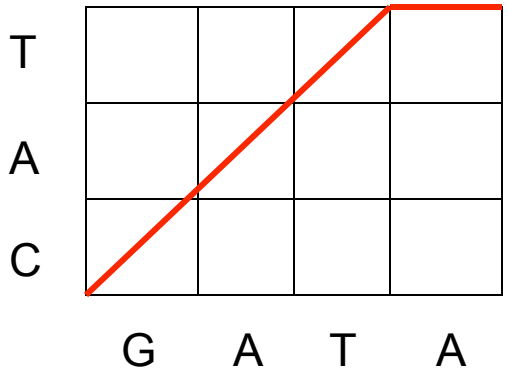
- identification des régions conservées et spécifiques :
  - rôle structural et/ou fonctionnel : (site actif, signature, sites covariants, certains aa ont été conservés
    pendant des millions d'années, pourquoi ?)

- prédiction de structure (protéine, ARN) :
  - zones conservées correspondent souvent aux structures secondaires
  - covariation => structure des ARN

- phylogénie :
  - retrouver l'histoire évolutive du jeu de séquences considérées

Complexité des alignements multiples

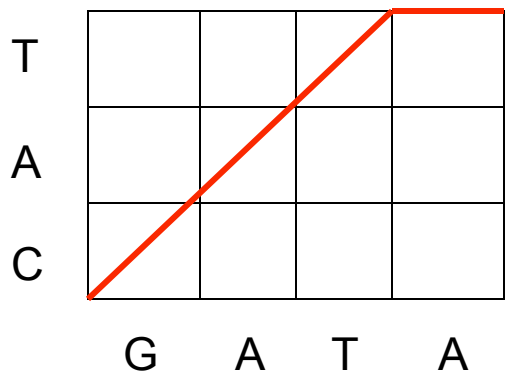
- alignement 2 à 2 => chemin dans une matrice de dimension 2



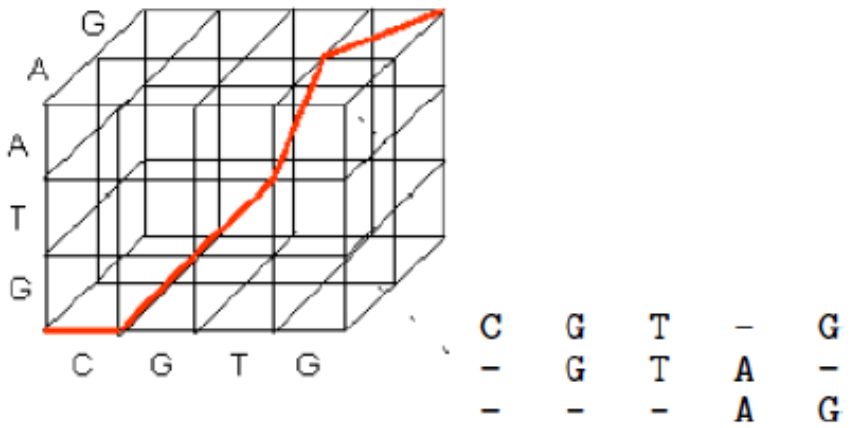
alignement de paire

Complexité des alignements multiples

- alignement 2 à 2 => chemin dans une matrice de dimension 2



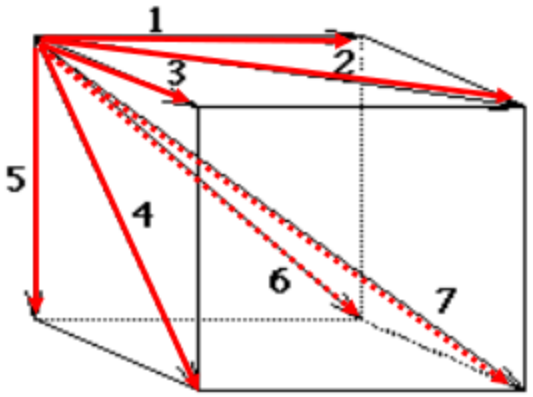
alignement de paire



alignement de 3 séquences

Complexité des alignements multiples

- alignement multiple inutilisable en pratique, combinatoire trop grande de n séquences => chemin dans une matrice de dimension n

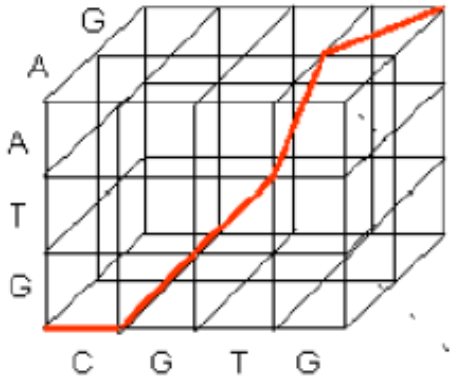


7 chemins

$$s_{i,j,k} = \max \begin{cases} s_{i-1,j-1,k-1} + \delta(v_i, w_j, u_k) \\ s_{i-1,j-1,k} + \delta(v_i, w_j, \_) \\ s_{i-1,j,k-1} + \delta(v_i, \_, u_k) \\ s_{i,j-1,k-1} + \delta(\_, w_j, u_k) \\ s_{i-1,j,k} + \delta(v_i, \_, \_) \\ s_{i,j-1,k} + \delta(\_, w_j, \_) \\ s_{i,j,k-1} + \delta(\_, \_, u_k) \end{cases}$$

La programmation dynamique peut s'étendre à l'alignement de k séquences de longueur n mais mais reste inefficace à cause du temps de calcul qui augmente de façon exponentielle ($7n^k$)

Complexité des alignements multiples




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

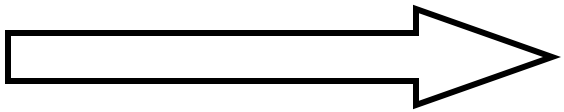
Utilisation d'algorithmes heuristiques :
Algorithmes approximatifs n'explorant pas tout l'espace des solutions mais permettant de donner des solutions approchant la solution optimale

On aligne donc les séquences de façon progressive

Complexité des alignements multiples

Nombreux programmes développés :

- clustal
- Muscle
- Mafft
- DiAlign
- Tcoffee



Autant de solutions différentes !

Score des alignements

Définition d'un score pour rendre compte de la qualité d'un alignement :

- indépendant du nombre de séquences
- indépendant de l'ordre des séquences
- reflétant la similarité

Score des alignements

Score d'entropie :

$$\Sigma_{\text{sur toutes les colonnes}} \; \Sigma_{X=A,T,G,C} \; p_X \log p_X$$

entropie d'alignement d'une colonne:

$$-(p_A \log p_A + p_C \log p_C + p_G \log p_G + p_T \log p_T)$$

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

- Colonne 1 = -[1*log(1) + 0*log0 + 0*log0 +0*log0]
  = 0
- Colonne 2 = -[(1/4)*log(1/4) + (3/4)*log(3/4) + 0*log0 + 0*log0]
  = -[(1/4)*(-2) + (3/4)*(-.415)] = +0.811
- Colonne 3 = -[(1/4)*log(1/4)+(1/4)*log(1/4)+(1/4)*log(1/4) +(1/4)*log(1/4)]
  = 4* -[(1/4)*(-2)] = +2.0
- Entropie d'alignement = 0 + 0.811 + 2.0 = +2.811

From A. Carbone

Score des alignements

Somme des paires

- les colonnes sont considérées comme indépendantes
- le score correspond à la somme des scores de tous les couples non ordonnés appartenant à une même colonne

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

```
          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
```

Approches heuristiques

- alignements progressifs : exemple CLUSTAL
  - calcul d'une matrice de similarité (programmation dynamique)
  - construction de l'arbre guide (Neighbor-Joining)
  - alignement progressif des nœuds de l'arbre par ordre décroissant de similarité

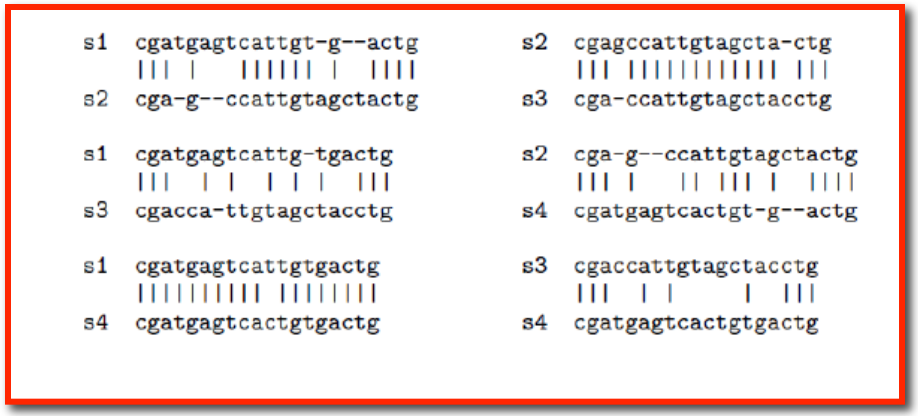
Exemple :

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

indel : -2, substitution : -1, identitée : 1

Score des alignements

- alignements progressifs : exemple CLUSTAL
  - calcul d'une matrice de similarité (programmation dynamique)
  - construction de l'arbre guide (Neighbor-Joining)
  - alignement progressif des nœuds de l'arbre par ordre décroissant de similarité



Score des alignements

- alignements progressifs : exemple CLUSTAL
  - calcul d'une matrice de similarité (programmation dynamique)
  - construction de l'arbre guide (Neighbor-Joining)
  - alignement progressif des nœuds de l'arbre par ordre décroissant de similarité

```
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
```


	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	

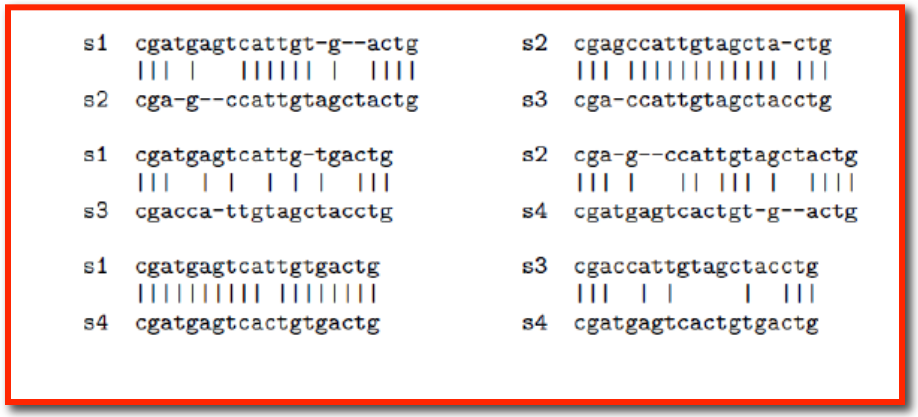
n séquences

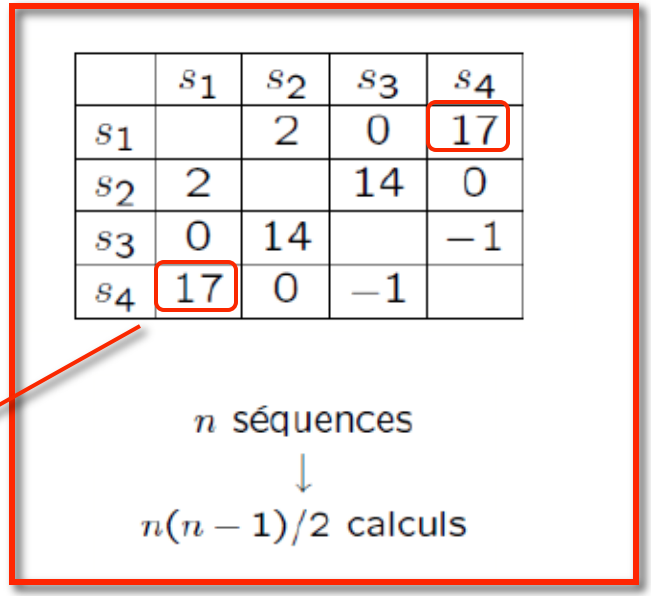
↓

$n(n-1)/2$ calculs

Score des alignements

- alignements progressifs : exemple CLUSTAL
  - calcul d'une matrice de similarité (programmation dynamique)
  - construction de l'arbre guide (Neighbor-Joining)
  - alignement progressif des nœuds de l'arbre par ordre décroissant de similarité





	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	

n séquences

↓

$n(n-1)/2$ calculs

$[(S_1,S_4),S2,S3]$

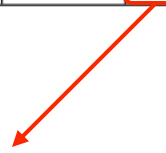
Score des alignements

- alignements progressifs : exemple CLUSTAL
  - calcul d'une matrice de similarité (programmation dynamique)
  - construction de l'arbre guide (Neighbor-Joining)
  - alignement progressif des nœuds de l'arbre par ordre décroissant de similarité

nouvelle matrice avec les nouveaux groupes : $[(S_1,S_4),S2,S3]$

	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	

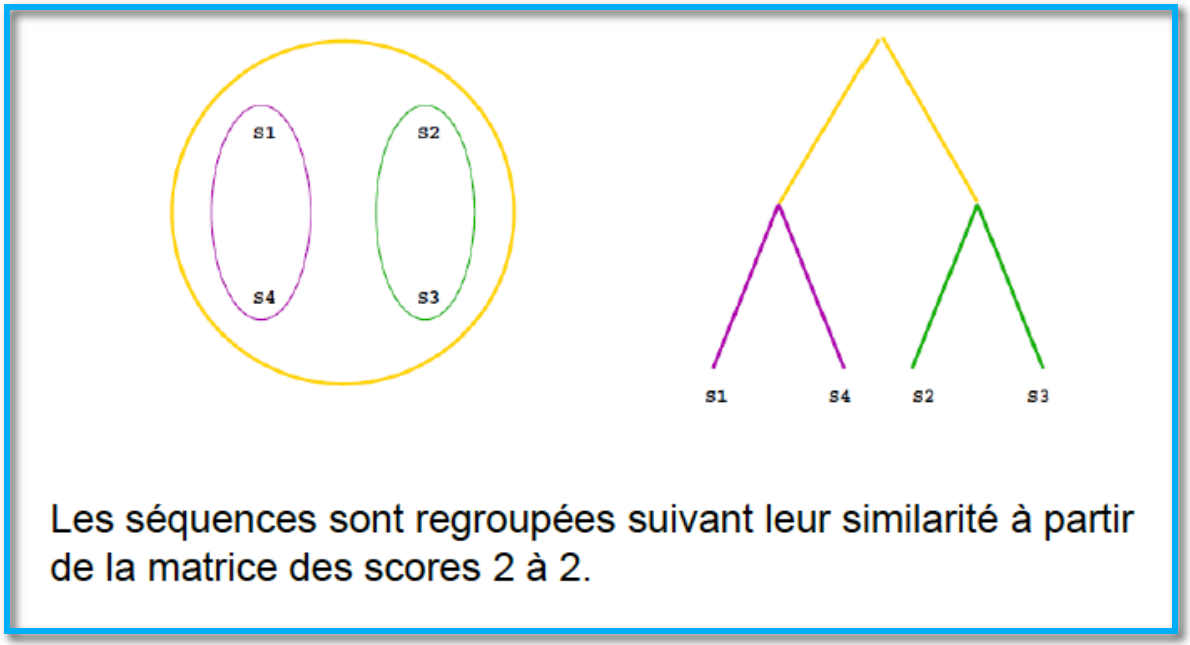
	S_{14}	S_2	S_3
S_{14}		1	-0,5
S_2	1		14
S_3	-0,5	14	



Nouveaux groupes : $[(S_1,S_4),(S2,S3)]$

Score des alignements

- alignements progressifs : exemple CLUSTAL
  - calcul d'une matrice de similarité (programmation dynamique)
  - construction de l'arbre guide (Neighbor-Joining)
  - alignement progressif des nœuds de l'arbre par ordre décroissant de similarité

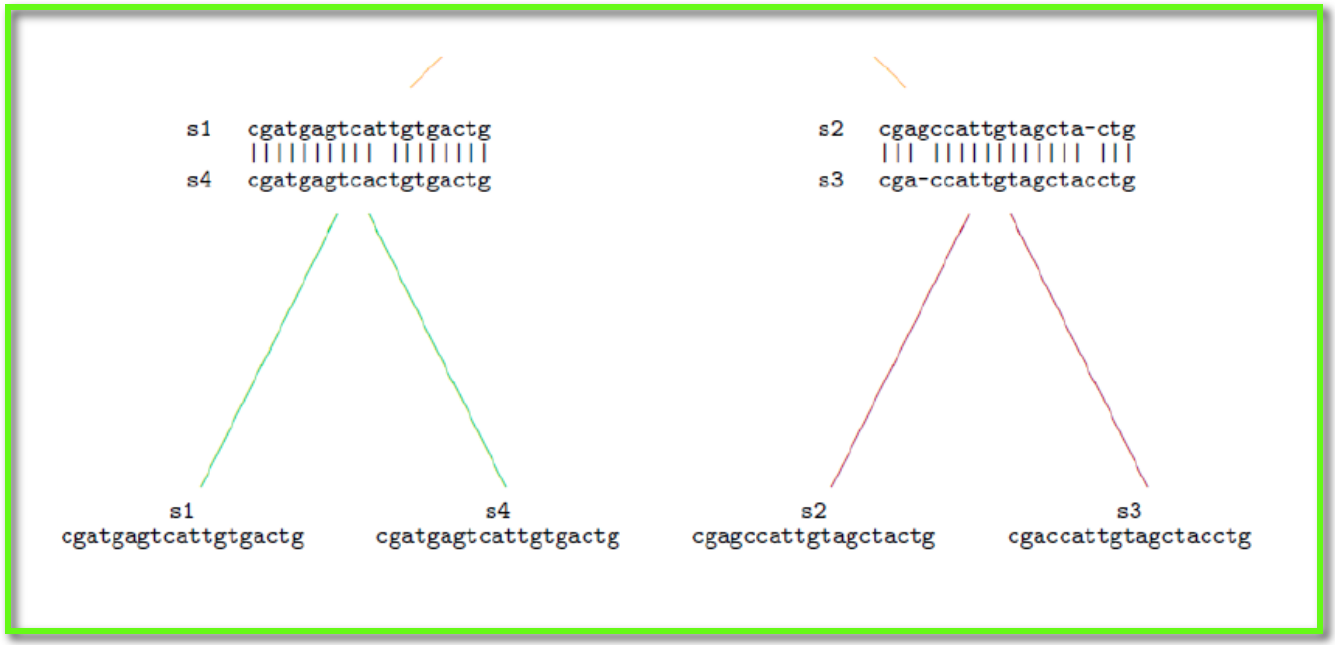


Les séquences sont regroupées suivant leur similarité à partir de la matrice des scores 2 à 2.

l'arbre « reflète » les relations évolutives entre les séquences

Score des alignements

- alignements progressifs : exemple CLUSTAL
  - calcul d'une matrice de similarité (programmation dynamique)
  - construction de l'arbre guide (Neighbor-Joining)
  - alignement progressif des nœuds de l'arbre par ordre décroissant de similarité



alignement de paires de séquences, séquence-profils, profils-profils en suivant l'arbre « guide »

Score des alignements

• alignements progressifs : exemple CLUSTAL

- calcul d'une matrice de similarité (programmation dynamique)
- construction de l'arbre guide (Neighbor-Joining)
- alignement progressif des nœuds de l'arbre par ordre décroissant de similarité

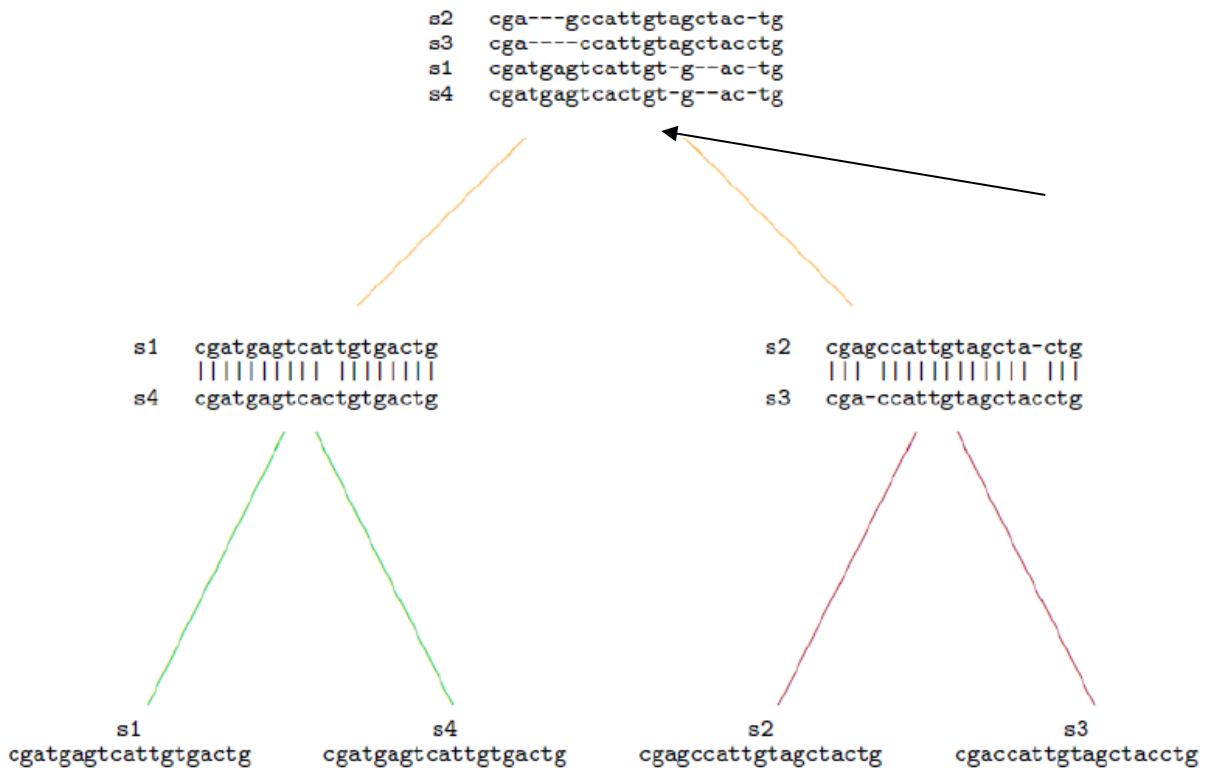
alignement de paires de séquences, séquence-profils, profils-profils

- à chaque étape : programmation dynamique
- matrice de substitution choisie en fonction de la divergence des groupes à aligner
- pénalité d'ouverture de « gap » dépend des séquences et positions à aligner
(pénalité/3 dans les régions hydrophiles, gly moins pénalisées que trp,
taille de la séquence…)

Score des alignements

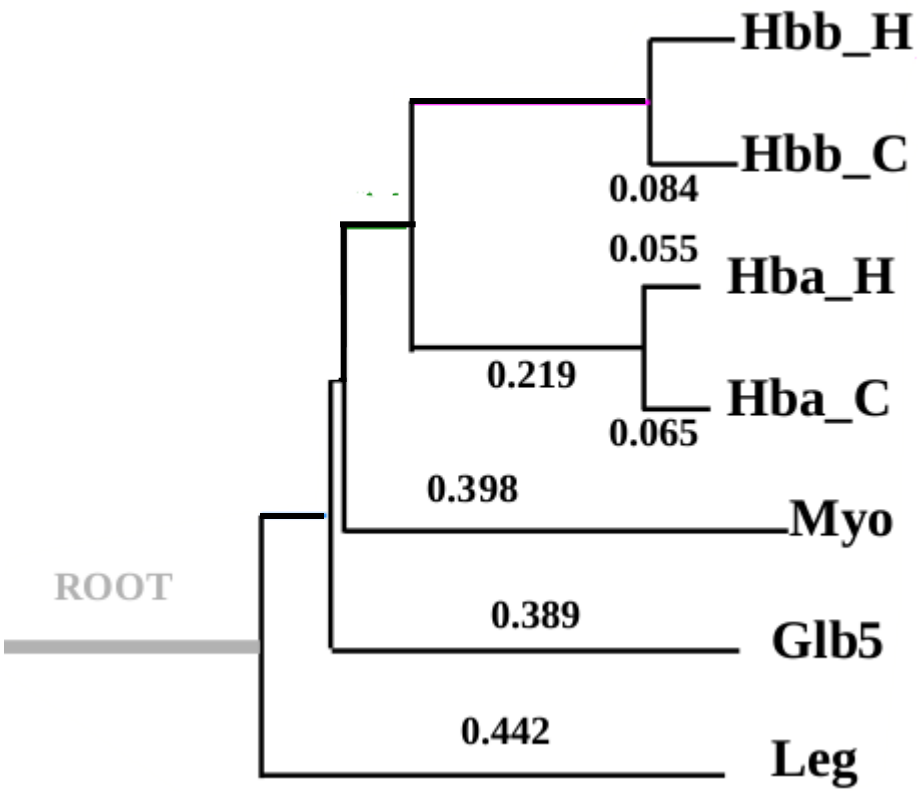
- alignements progressifs : exemple CLUSTAL

dernière étape : construction de l'alignement final



alignement des profils

Pondération des branches (ClustalW)



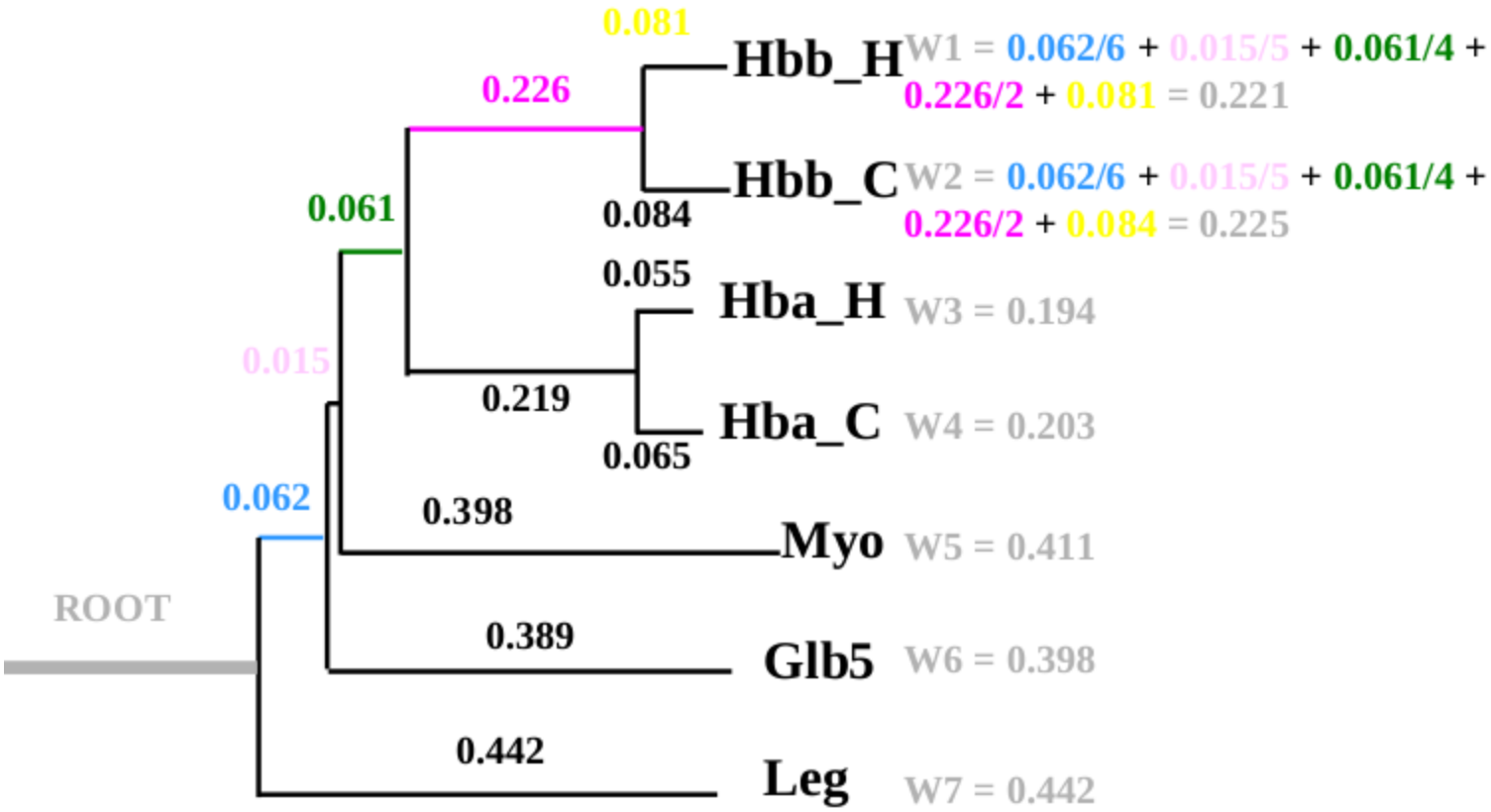
Pondération des branches (ClustalW)

Pondération des branches de l'arbre

- principe : attribuer un poids à chaque branche pour empêcher les groupes de séquences sur-représentées de dominer dans l'alignement
- dépend de la taille de la branche et du nombre de feuilles au bout de la branche (redondance de l'information)
- le poids d'une séquence correspond à la somme des longueurs de branches pondérées allant de la racine à la séquence

Pondération des branches (ClustalW)

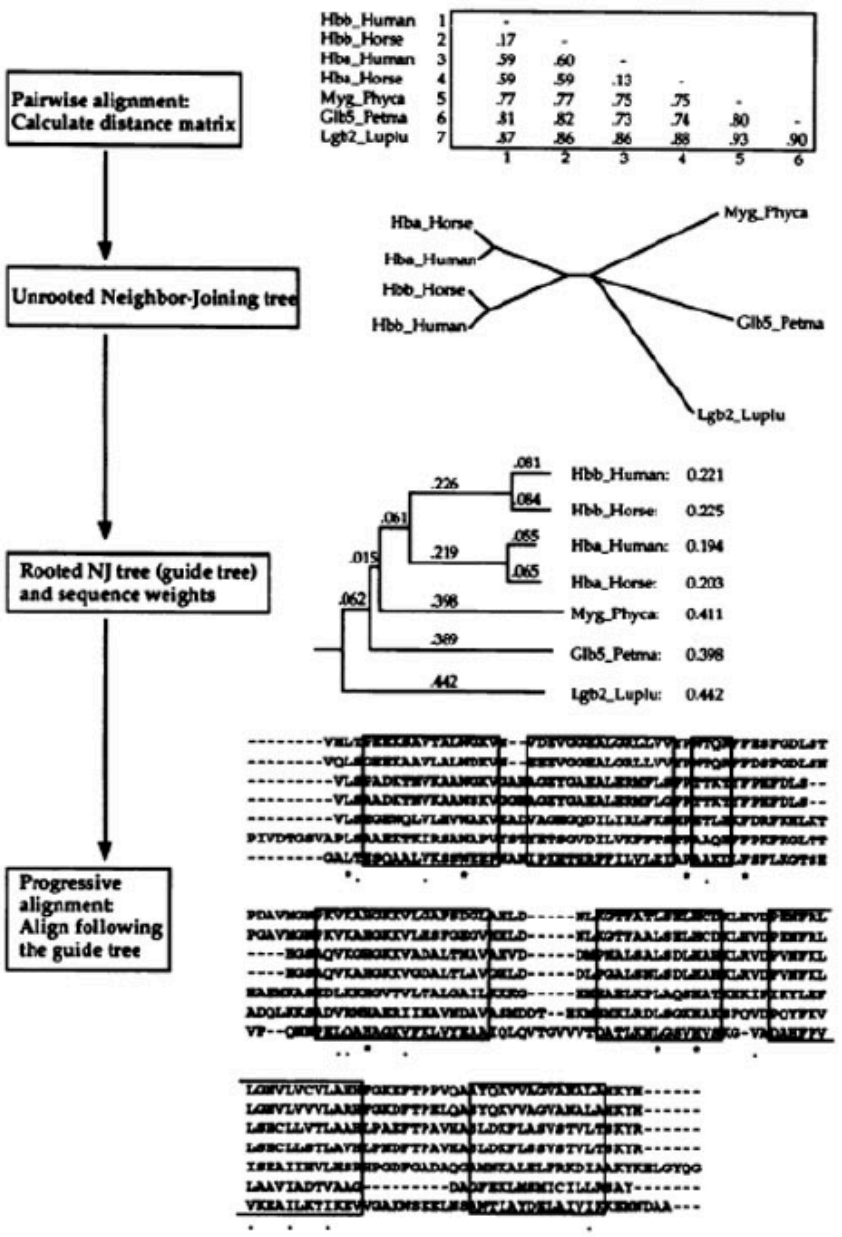
Pondération des branches de l'arbre



Score de l'alignement profil-profil

Score avec pondération des branches de l'arbre

ClustalW (W = weight)



Approches heuristiques - Conclusion

• **alignements progressifs : exemple CLUSTAL // PARAMETRES & LIMITES**

cas de clustalW (adapté aux protéines) :

- séquences pondérées en fonction de leur sur/sous représentation
- adaptation des matrices de similarité au fur et à mesure en fonction de la divergence des séquences à aligner (BLOSUM50, BLOSUM62, BLOSUM80)
- pénalité des gaps en fonction du type de résidu (gly plus sujette à être entourée de gaps que trp p. ex.)
- pénalité des gaps réduite dans les régions hydrophyles

paramètres principaux :

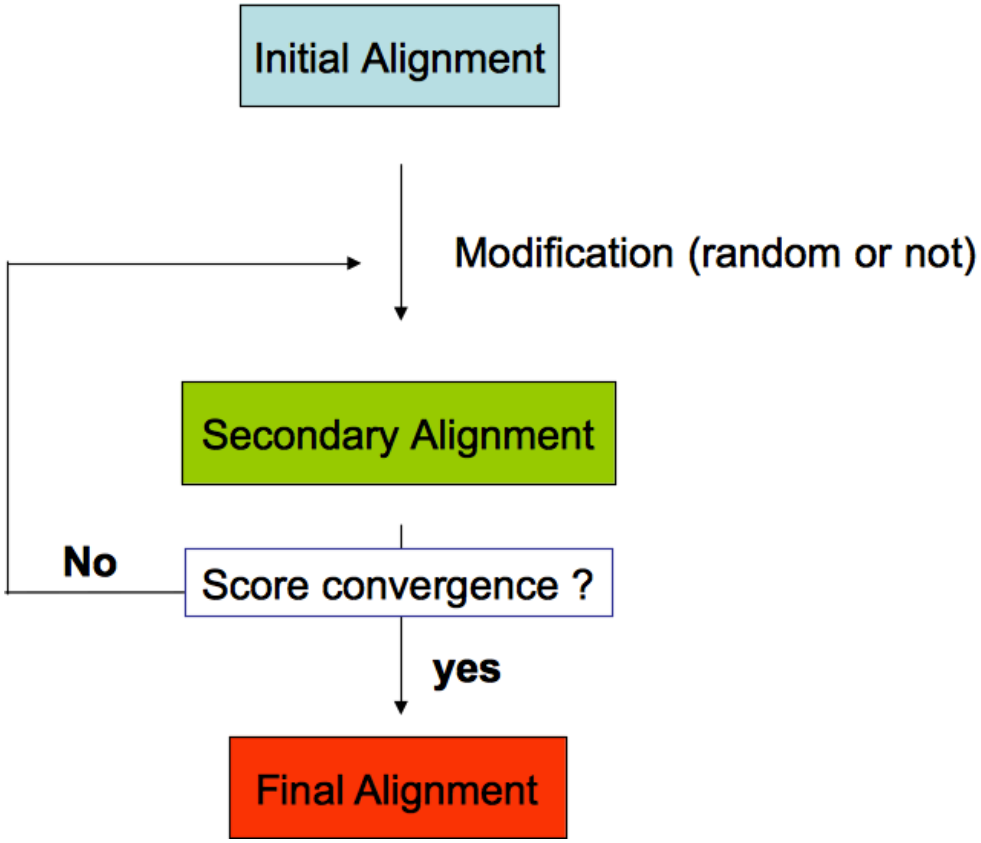
matrices (BLOSUM/PAM/Gonnet), pénalité ouverture/extension de gaps, qualité des alignements par pairs, qualité des alignements de paire

limites : algorithme qui ne revient pas en arrière et ajoute indel à chaque ajout de nouveau groupe (applicable aux familles proches mais limité pour les autres)

améliorations : méthodes améliorant la définition de l'arbre (A/R entre alignement multiple et l'arbre généré pour optimiser l'ensemble de façon itérative) (DiAlign, Muscle…)

Et après ? Autres solutions ?

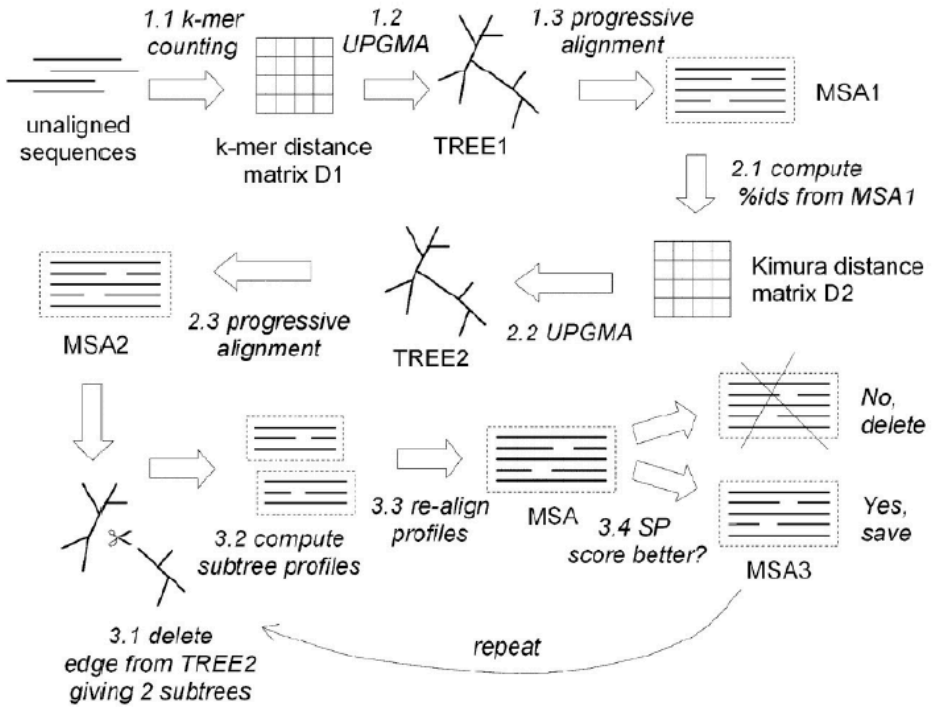
- alignements itératifs : exemple MUSCLE



Et après ? Autres solutions ?

- alignements itératifs : exemple MUSCLE

même stratégie que CLUSTAL au départ



Evaluation des programmes d'alignement

HOME
OVERVIEW
GALLERY
SOURCE
CREDITS

For any questions or suggestions please contact us

GoTo ...
ConSurf
ConSeq
Selecton
Pepitope
Epitopia
QuasiMotiFinder

The GUIDANCE server

guide-tree based alignment confidence

Select Algorithm GUIDANCE

Warning: GUIDANCE is not suitable for alignments of very few sequences. As a rule of thumb, use HoT for less than 8 sequences.

Type your sequences (FASTA format only)

OR

Upload your sequences file (FASTA format only) Parcourir...

Sequences Type: ○ Amino Acids ○ Nucleotides ○ Codons

Select the MSA algorithm MAFFT (default)

Warning: PRANK is significantly more time consuming. MAFFT is the fastest.

Please enter your email address (Optional)

Your email address will be used to update you the moment the results are ready.

Evaluation des programmes d'alignement

Table 3. Performance of aligners on the PREFAB protein reference alignment benchmark

Aligner	Overall (1927)	Time
DIALIGN	57.2	12 h, 25 min
CLUSTALW	58.9	2 h, 57 min
T-Coffee	63.6	144 h, 51 min
MUSCLE	64.8	3 h, 11 min
MAFFT	64.8	**2 h, 36 min**
ProbCons	66.9	19 h, 41 min
ProbCons-ext	**68.0**	37 h, 46 min

Evaluation des programmes d'alignement

Programme	Avantages	Précautions
ClustalW	Utilise moins de mémoire que les autres	Moins précis que tous les logiciels modernes, difficulté pour les grands jeux de séquences
DIALIGN	Tente de différencier régions alignables et non alignables	Moins précis que Clustal pour les alignements globaux
MAFFT **MUSCLE**	- Plus rapide et plus précis que ClustalW - Équilibre entre précision et vitesse de calcul	Pour les grands jeux de données (>1000 séquences), utiliser les options d'optimisation
T-Coffee	Bonne précision et capable de mélanger des données hétérogènes	Forte complexité en temps de calcul et utilisation de l'espace mémoire Inutilisable avec plus de 100 séquences

Evaluation des programmes d'alignement

Quand les séquences sont peu divergentes (%id > 35), toute les méthodes sont relativement fiables (alignement correct à 90%, SSR essentiellement)

Twilight zone (15-25%) : toutes les méthodes sont en difficulté !

Pas de méthode universelle ! En essayer plusieurs !

Dépend des attentes de l'utilisateur

Alternative : les HMM et profils HMM => voir Hhalign, HMMER

ajustement manuel : jalview, seaview, swissprot PDB viewer