

L'Alignement multiple et ses applications

Hélène TOUZET

`touzet@lifl.fr`

Définition de l'alignement multiple

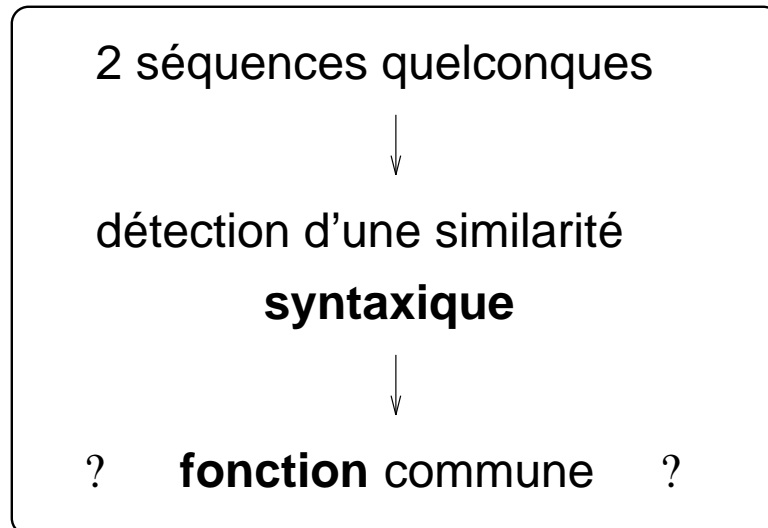
▷ **Entrée** : k séquences

```
* * * * * * * * * * * * * *
* * * * * * * * * *
* * * * * * * * * * * * * *
* * * * * * * * * * *
```

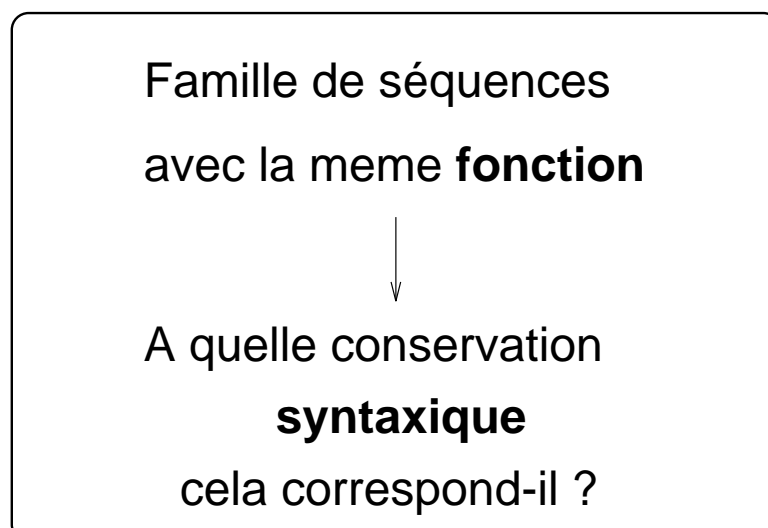
▷ **Sortie** : un tableau contenant les k séquences, avec des indels

```
* * * * * * * * * - * * * *
* * * - - - * * * - * * * *
* * * - * * * * * * * * * *
* * * - - * * - - * * * * *
```

Alignement 2 à 2



Alignement multiple

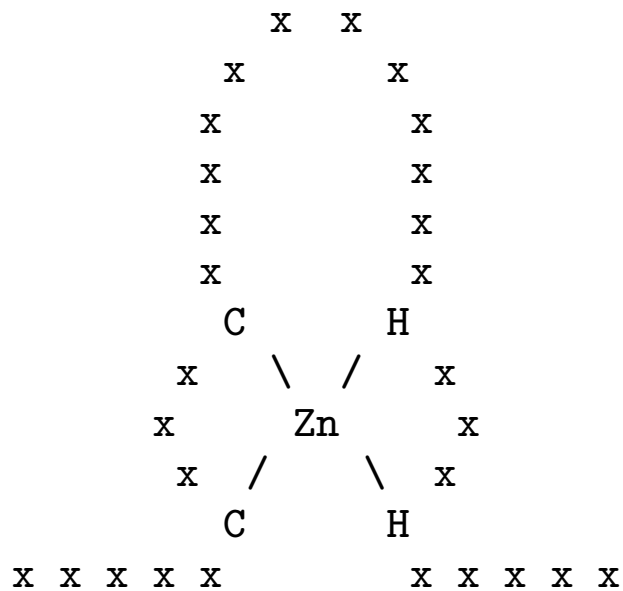
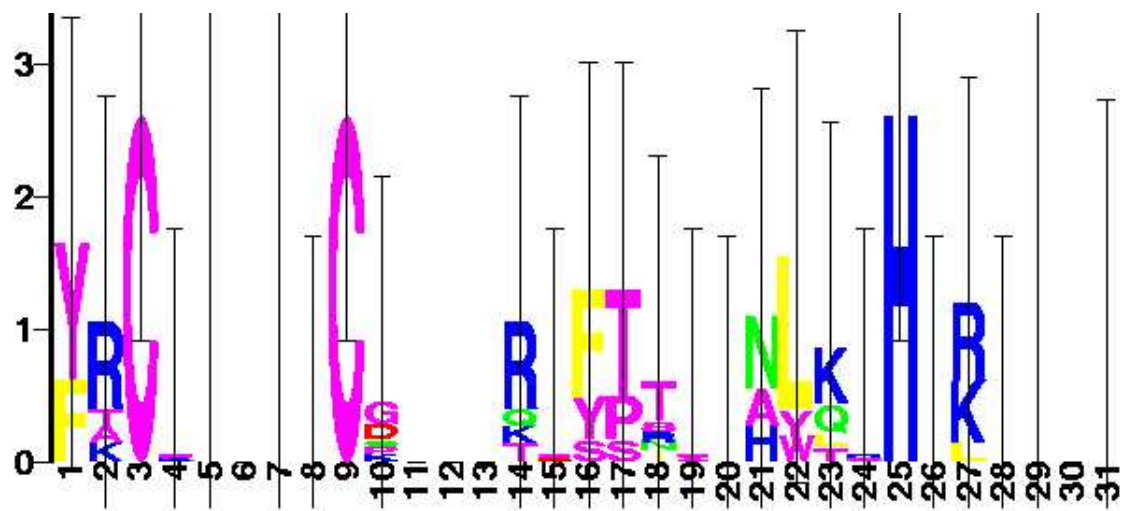


Application 1 : modélisation de motifs

Exemple : Doigt de zinc

TTY1_HUMAN/383-407	YVCPF-DGCN---KKFAQSTNLKSHILT---H
YKQ8_CAEEL/78-102	YKCT---VCR---KDISSESRLRTHMFKQ-HH
BASO_HUMAN/719-742	FQCD---ICK---KTFKNACSVKIHKHN--MH
ZG29_XENLA/62-84	FVCT---VCG---KTYKYKHGLNTHLHS---H
P43_XENBO/106-130	LKCSV-PGCK---RSFRKKRALRIHVSE---H
IKAR_MOUSE/488-512	FECN---MCG---YHSQDRYEFSSHITRG-EH
Q92610/1043-1069	YTCG---YCTEDSPSFPRPSLLESHISL--MH
TRA1_CAEEL/306-331	YKCEF-ADCE---KAFSNASDRAKHQNR--TH
ZN10_HUMAN/383-405	YKCN---QCG---IIFSQNSPFIVHQIA---H
GLI1_XENLA/283-310	FVCHW-QDCSRELRPFKAQYMLVVHMRR---H
XFIN_XENLA/276-298	FRCS---ECS---RSFTHNSDLTAHMRK---H
TF3A_BUFAM/72-97	CKCET-ENCN---LAFTTASNMLHFKR--AH
ZG58_XENLA/174-196	FVCT---ECN---LSFAGLANLRSHQHL---H
P43_XENBO/163-187	YRCSY-EDCQ---TVSPTWTALQTHLKK---H
TSH_DROME/354-378	FRCV---WCK---QSFPTLEALTTHMKDS-KH
ZN76_HUMAN/165-189	FRCGY-KGCG---RLYTTAHHLKVHERA---H
TF3A_BUFAM/219-244	YRCPR-ENCD---RTYTTKFNLKSHILT--FH
SUHW_DROAN/349-373	YACK---ICG---KDFTRSYHLKRHQKYS-SC
ZN76_HUMAN/285-309	YTCPE-PHCG---RGFTSATNYKNHVRI---H
SRYC_DROME/469-492	FKCN---YCP---RDFTNFPNWLKHTRR--RH
EVI1_HUMAN/761-784	YRCK---YCD---RSFSISSNLQRHVRN--IH
...	

Extrait de Pfam, entrée zf-C2H2



Modélisation

Prosite : motif

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

Exemple : Site de fixation de la cellulose

```

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

Extrait de Prosite, entrée PS00562

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

```

+-----+
|               +-----|-----+
|               |       |       |
xxxxxxCxxxxxxxxxxCxxxxxCxxxxxxxxxxCx
*****
  
```

Les 4 cystéines sont impliquées dans des liaisons di-sulfures.

Exemple Site de fixation du facteur de transcription SP1

ctccgcccga
ccccgcccaca
ccccgccccca
ccccgccccca
ccccgcccccg
ccacgccccca

	<i>A</i>	<i>C</i>	<i>G</i>	<i>T</i>
1	0	6	0	0
2	0	5	0	1
3	1	5	0	0
4	0	6	0	0
5	0	0	6	0
6	0	6	0	0
7	0	6	0	0
8	0	6	0	0
9	1	4	1	0

Application 2 : Structure d'ARN

```
G A G C C C A G U U C
  A G G A C U C U U C
A A U C A C C C G A U
```

Changement de base compensatoire:

Quand une base impliquée dans un appariement mute, la base complémentaire mute également, pour préserver la paire, et donc, la structure secondaire.

Méthode des covariations

Étape 1 : *construction d'un alignement multiple*

```
G  A  G  C  —  C  C  A  G  U  U  C
—  A  G  G  A  C  —  U  C  U  U  C
A  A  U  C  A  C  C  C  G  A  U  —
—  A  G  G  A  C  —  U  C  U  U  C
```

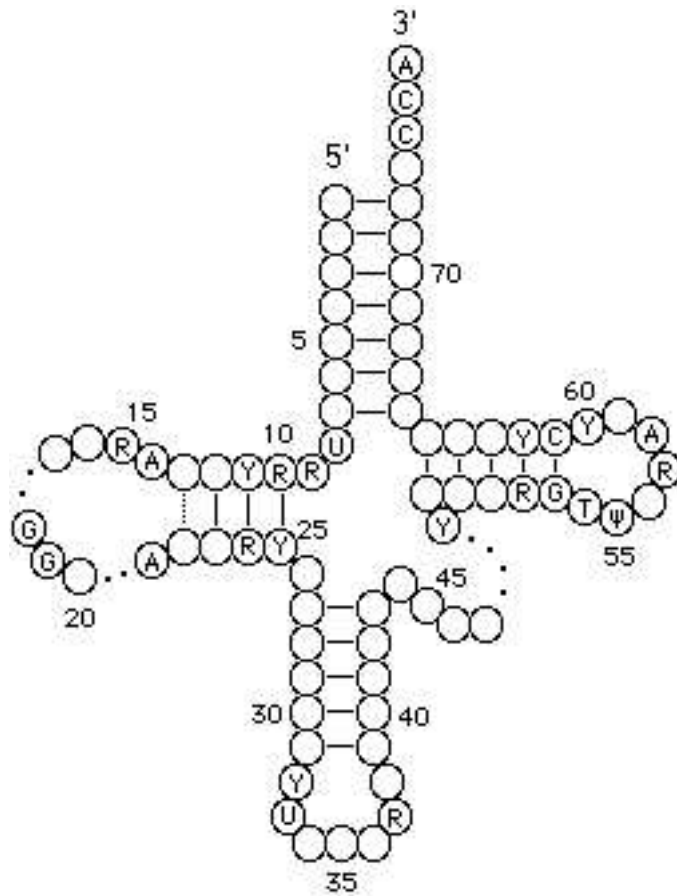
Étape 2 : *détection des positions corrélées*

présomption d'appariement

DA1650	TGC	LEUCONOSTOC	LACTIS	GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTTCGATCCCGCTATTCTCCA---
DA1660	TGC	E. COLI		GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA
DA1670	TGC	LEUCONOSTOC	MESEN.	GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTTCGATCCCGCTATTCTCCA---
DA1710	TGC	MYCOBACT.	LEPRAE	GGGGCCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGAAGGCAGATGTCAGGGGTTTCGATCCCGCTAGGCTCCA---
DA1730	TGC	TRICHODESMIUM	SP.	GGGGGTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGAAGGCAGAAGTCAGCGGTTTCGA.TCCGCTTACCCCA---
DA1780	TGC	AEROMONAS	HYDROPH.	GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA
DA1790	TGC	PREVOTELLA	RUMINI.	GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA
DA1810	TGC	PSEUDOMONAS	CEPAC.	GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTTCGGTTTCGATCCCGTCTGCCTCCACCA
DA1820	TGC	PSEUDOMONAS	AER.	GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTTCAGGAGTTTCGATCCTCCTTGGCTCCACCA
DA1830	TGC	PSEUDOMONAS	GLAD.	GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTTCGGTTTCGATCCCGTCTGCCTCCACCA
DA1840	TGC	PSEUDOMONAS	FLUOR.	GGGGCCATAGCTCAGCTGGGGAGAGCGCCTGCCTTTGCACGCAGGAGGTCAACGGTTTCGATCCCGTTTTGGCTCCA---
DA1850	TGC	PSEUDOMONAS	MALLEI	GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTTCGGTTTCGATCCCGTCTGCCTCCACCA
DA1860	TGC	CAMPYLOBAC.	JEJUNI	GGGGCATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTTCAGCGGTTTCGATCCCGCTATTCTCCACCA
DA1890	TGC	CAULOBACTER	CRES.	GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTTCGGTTTCGAATCCGTCTGGCTCCACCA
DA1900	TGC	BRUCELLA	SUIS	GGGGCCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGGAGGTTTCGATCCCGTCCGGCTCCACCA
DA1910	TGC	BRUCELLA	MELITENS.	GGGGCCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCGTTCGGTTTCGATCCCGTCCGGCTCCACCA
DA1920	TGC	BRUCELLA	ABORTUS	GGGGCCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGTTCGGTTTCGATCCCGTCCGGCTCCACCA

Alignement d'ARN de transfert (extrait de la compilation de Bayreuth, Germany)

Vérification . . .



Structure secondaire de l'ARNt

Application 3 : phylogénie moléculaire

- ▷ Retracer l'historique des espèces à partir des mutations observées
- ▷ Données : gènes communs aux familles étudiées, pas trop divergents
- ▷ Résultat : classification sous forme d'arbre phylogénétique

Méthodes de parcimonie

Rasoir d'Occam

*"Pluralitas non est
ponenda sine neccesitate"*



- ▷ Privilégier l'arbre qui minimise le nombre de mutations
- ▷ Le nombre global de mutations est obtenu en faisant la somme des mutations le long de chaque branche

Exemple

1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G

3 arbres non enracinés possibles

Inconvénient : long

Méthodes de distance

- ▷ Point de départ : alignement multiple
- ▷ Matrice de toutes les distances deux à deux
- ▷ Classification hiérarchique

On construit l'arbre à partir des feuilles en regroupant progressivement les noeuds 2 à 2 pour former des **clusters**.

Exemple : UPGMA

Unweight Pair Group Method with Arithmetic mean

- ▷ À chaque étape, on regroupe les deux clusters les plus proches.
- ▷ La distance est calculée en faisant la moyenne arithmétique

Reprise de l'exemple précédent

Matrice initiale

	1	2	3	4
1	0	4	5	6
2		0	5	4
3			0	2
4				0

Après une itération

	1	2	3 + 4
1	0	4	5,5
2		0	4,5
3 + 4			0

Comment construire un alignement multiple ?

Clustal

Thompson *et al.* - 1994

CLUSTAL = cluster + alignement

Étape 1 : alignements globaux 2 à 2

Étape 2 : arbre phylogénétique (clusters)

Étape 3 : alignement multiple obtenu par combinaisons des alignements 2 à 2

Exemple :

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

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

Étape 1 : *calcul des scores de similarité de tous les alignements globaux 2 à 2*

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

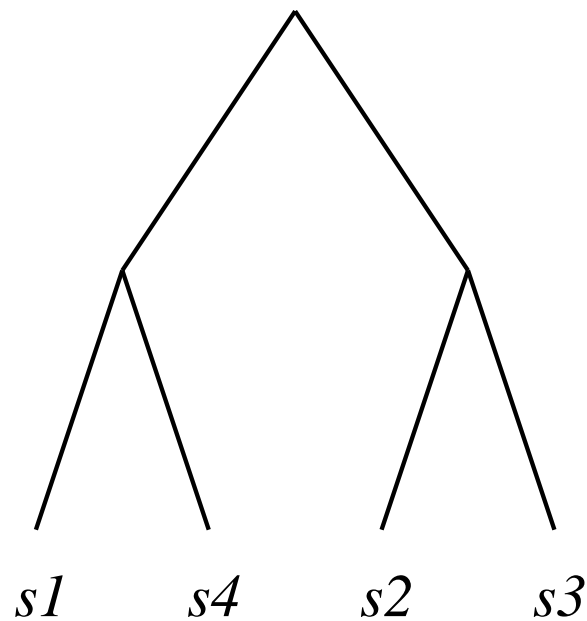
Tableau des scores d'alignement:

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

n séquences
 ↓
 $n(n - 1)/2$ calculs

Étape 2 : *construction de l'arbre guide*

Arbre obtenu avec l'algorithme de Neighbor-Joining



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

les séquences les plus proches sont alignées, et ainsi de suite.

Étape 3 : *construction de l'alignement multiple final*

Alignement progressif des séquences, en les incorporant dans l'ordre de l'arbre guide.

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

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



```
s2  CGA---GCCATTGTAGCTAC-TG
s3  CGA----CCATTGTAGCTACCTG
s1  CGATGAGTCATTGT-G--AC-TG
s4  CGATGAGTCACTGT-G--AC-TG
```

Once a gap, always a gap