Alignement multiple

Équipe Bonsai

http://www.lifl.fr/bonsai

année 2012

Définition de l'alignement multiple

Définition de l'alignement multiple

- entrée : k séquences

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

Définition de l'alignement multiple

- entrée : k séquences

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

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

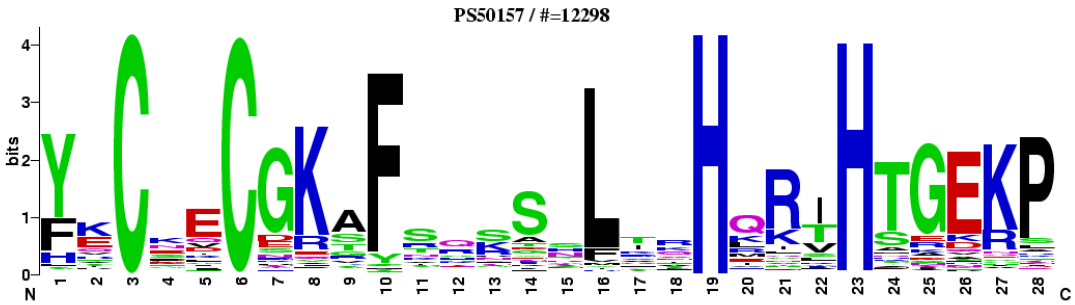
```
C A T G C G A G T A - G T A G
C A T G - - - G T A - G T A G
C C T G - G A G T A C G T A G
C A T G - - A G - - C G T A G
```

Motif doigt de zinc (*C2H2-type*)

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

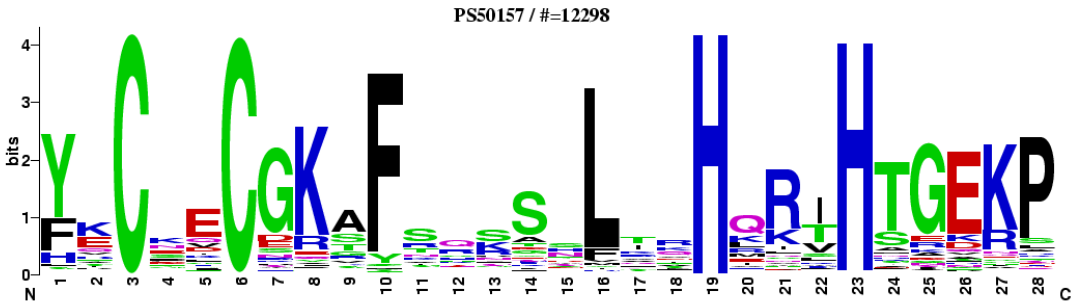
Motif doigt de zinc (*C2H2-type*)

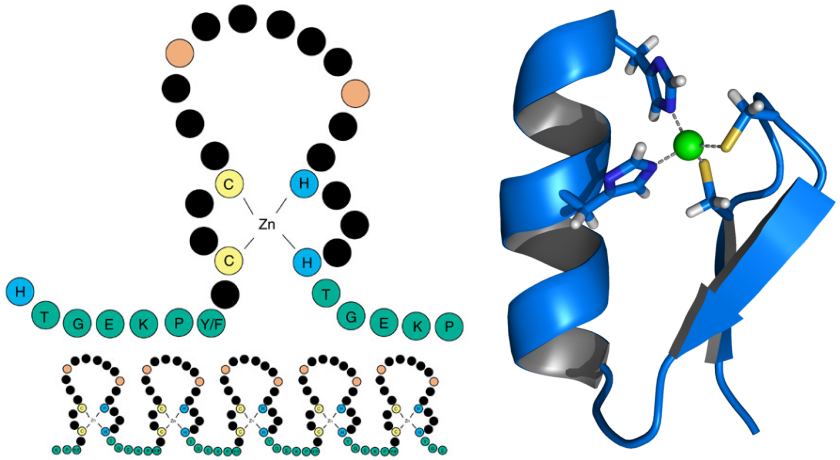
```
TYY1_HUMAN YVCPFDGCNKKFAQSTNLKSHILT--H
YKQ8_CAEEL YKCT--VCRKDISSSESLRTHMFKQHH
BASO_HUMAN FQCD--ICKKTFKNACSVKIHHKN-MH
ZG2-9_XENL FVCT--VCGKTYKYKHGLNTHLHS--H
P43_XENBO  LKCSVPGCKRSFRKKRALRIHVSE--H
IKAR_MOUSE FECN--MCGYHSQDRYEFSSHITRGEH
TRA1_CAEEL YKCEFADCEKAFSNASDRAKHQNR-TH
ZN10_HUMAN YKCN--QCGIIFSQNSPFIVHQIA--H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



Motif doigt de zinc (*C2H2-type*)

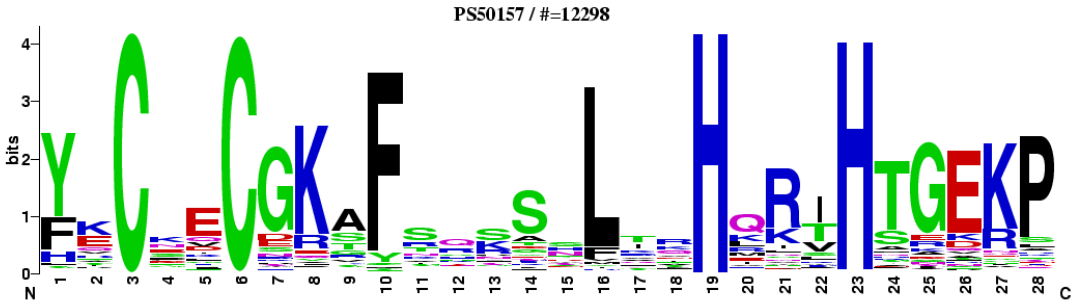
```
TYY1_HUMAN YVCPFDGCNKKFAQSTNLKSHILT--H
YKQ8_CAEEL YKCT--VCRKDISSSESLRTHMFKQHH
BASO_HUMAN FQCD--ICKKTFKNACSVKIHHKN-MH
ZG2-9_XENL FVCT--VCGKTYKYKHGLNTHLHS--H
P43_XENBO  LKCSVPGCKRSFRKKRALRIHVSE--H
IKAR_MOUSE FECN--MCGYHSQDRYEFSSHITRGEH
TRA1_CAEEL YKCEFADCEKAFSNASDRAKHQNR-TH
ZN10_HUMAN YKCN--QCGIIFSQNSPFIVHQIA--H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

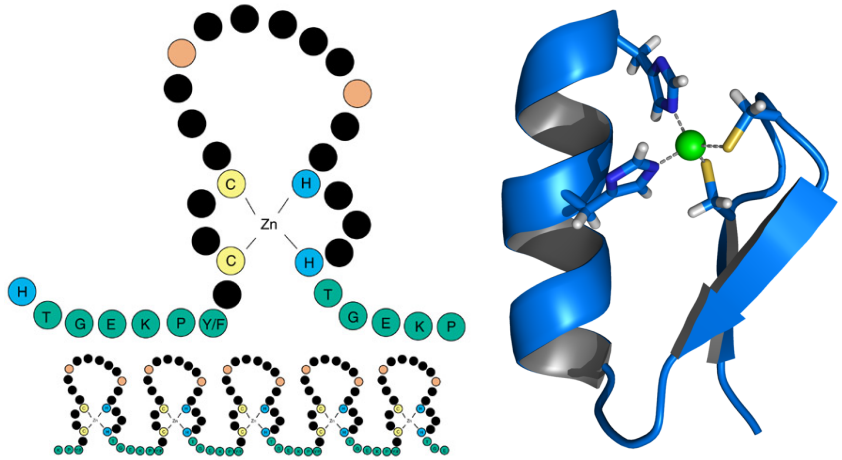




Motif doigt de zinc (*C2H2-type*)

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





modélisation : motif Prosite

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

Site de fixation de la cellulose

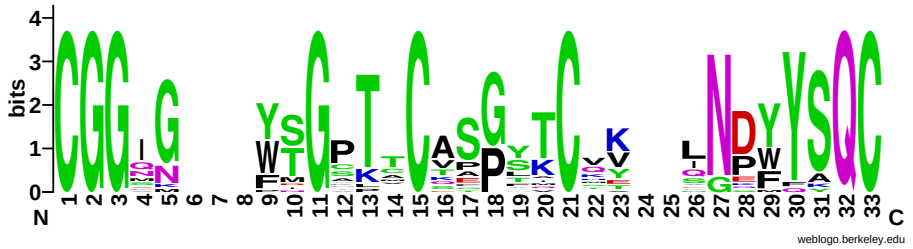
```
HWGQCGGI---GYSGCKTCTSGTTCQYSNDYYSQCL
HYGQCGGI---GYSGPTVCASGTTCQVLNPYYSQCL
QWGQCGGI---GYTGSTTCASPYTCHVLNPYYSQCY
VWGQCGGQ---NWSGPTCCASGSTCVYSNDYYSQCL
LYGQCGGA---GWTGPTTCQAPGTCKVQNQWYSQCL
IWGQCGGN---GWTGATTCASGLKCEKINDWYYQCV
VWGQCGGN---GWTGPTTCASGSTCVKQNDFYSQCL
DWAQCGGN---GWTGPTTCVSPYTCTKQNDWYSQCL
QWGQCGGQ---NYSGPTTCKSPFTCKKINDFYSQCQ
RWQQCGGI---GFTGPTQCEEPYICTKLNDWYSQCL
HWAQCGGI---GFSGPTTCPEPYTCAKDHDIYSQCV
LYEQCGGI---GFDGVTCCSEGLMCMKMGPYYSQCR
VWAQCGGQ---NWSGTPCCTSGNKCVKLNDFYSQCQ
PYGQCGGM---NYSGKTMCSPGFKCVELNEFFSQCD
AYYQCGGSKSAYPNGNLACATGSKCVKQNEYYSQCV
EYAACGGE---MFMGAKCCKFGLVCYETSGKWSQCR
```

extrait de Prosite, entrée PS00562

Site de fixation de la cellulose

```
HWGQCGGI---GYSGCKTCTSGTTCQYSNDYYSQCL
HYGQCGGI---GYSGPTVCASGTTCQVLNPYYSQCL
QWGQCGGI---GYTGSTTCASPYTCHVLNPYYSQCY
VWGQCGGQ---NWSGPTCCASGSTCVYSNDYYSQCL
LYGQCGGA---GWTGPTTCQAPGTCKVQNQWYSQCL
IWGQCGGN---GWTGATTCASGLKCEKINDWYYQCV
VWGQCGGN---GWTGPTTCASGSTCVKQNDFYSQCL
DWAQCGGN---GWTGPTTCVSPYTCTKQNDWYSQCL
QWGQCGGQ---NYSGPTTCKSPFTCKKINDFYSQCQ
RWQQCGGI---GFTGPTQCEEPYICTKLNDWYSQCL
HWAQCGGI---GFSGPTTCPEPYTCAKDHDIYSQCV
LYEQCGGI---GFDGVTCCSEGLMCMKMGPYYSQCR
VWAQCGGQ---NWSGTPCCTSGNKCVKLNDFYSQCQ
PYGQCGGM---NYSGKTMCSPGFKCVELNEFFSQCD
AYYQCGGSKSAYPNGNLACATGSKCVKQNEYYSQCV
EYAACGGE---MFMGAKCCKFGLVCYETSGKWSQCR
```

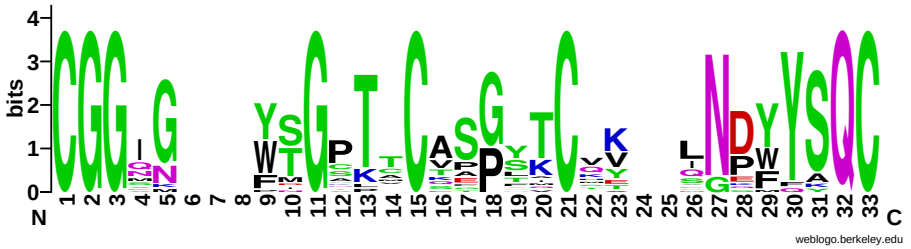
extrait de Prosite, entrée PS00562

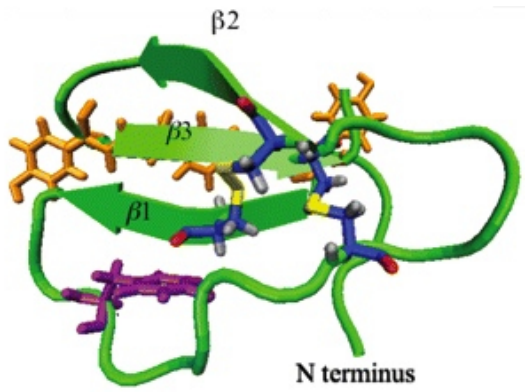


Site de fixation de la cellulose

```
HWGQCGGI---GYSGCKTCTSGTTCQYSNDYYSQCL
HYGQCGGI---GYSGPTVCASGTTCQVLNPYYSQCL
QWGQCGGI---GYTGSTTCASPYTCHVLNPYYSQCY
VWGQCGGQ---NWSGPTCCASGSTCVYSNDYYSQCL
LYGQCGGA---GWTGPTTCQAPGTCKVQNQWYSQCL
IWGQCGGN---GWTGATTCASGLKCEKINDWYYQCV
VWGQCGGN---GWTGPTTCASGSTCVKQNDFYSQCL
DWAQCGGN---GWTGPTTCVSPYTCTKQNDWYSQCL
QWGQCGGQ---NYSGPTTCKSPFTCKKINDFYSQCQ
RWQQCGGI---GFTGPTQCEEPYICTKLNDWYSQCL
HWAQCGGI---GFSGPTTCPEPYTCAKDHDIYSQCV
LYEQCGGI---GFDGVTCCSEGLMCMKMGPYYSQCR
VWAQCGGQ---NWSGTPCCTSGNKCVKLNDFYSQCQ
PYGQCGGM---NYSGKTMCSPGFKCVELNEFFSQCD
AYYQCGGSKSAYPNGNLACATGSKCVKQNEYYSQCV
EYAACGGE---MFMGAKCCKFGLVCYETSGKWSQCR
```

extrait de Prosite, entrée PS00562

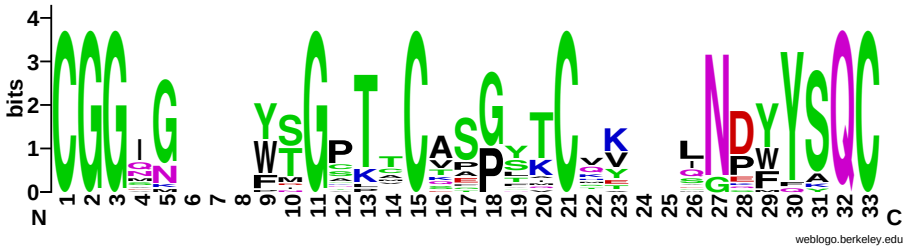


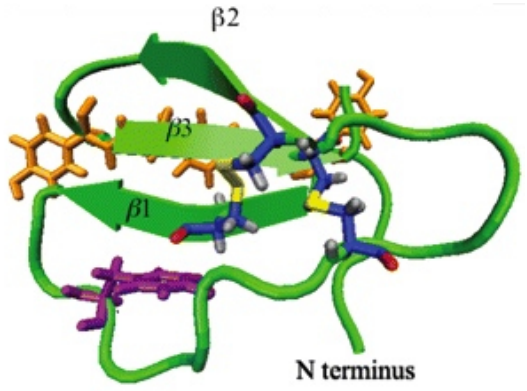


Site de fixation de la cellulose

```
HWGQCGGI---GYSGCKTCTSGTTCQYSNDYYSQCL
HYGQCGGI---GYSGPTVCASGTTCQVLNPYYSQCL
QWGQCGGI---GYTGSTTCASPYTCHVLNPYYSQCY
VWGQCGGQ---NWSGPTCCASGSTCVYSNDYYSQCL
LYGQCGGA---GWTGPTTCQAPGTCKVQNQWYSQCL
IWGQCGGN---GWTGATTCASGLKCEKINDWYYQCV
VWGQCGGN---GWTGPTTCASGSTCVKQNDFYSQCL
DWAQCGGN---GWTGPTTCVSPYTCTKQNDWYSQCL
QWGQCGGQ---NYSGPTTCKSPFTCKKINDFYSQCQ
RWQQCGGI---GFTGPTQCEEPYICTKLNDWYSQCL
HWAQCGGI---GFSGPTTCPEPYTCAKDHDIYSQCV
LYEQCGGI---GFDGVTCCSEGLMCMKMGPYYSQCR
VWAQCGGQ---NWSGTPCCTSGNKCVKLNDFYSQCQ
PYGQCGGM---NYSGKTMCSPGFKCVELNEFFSQCD
AYYQCGGSKSAYPNGNLACATGSKCVKQNEYYSQCV
EYAACGGE---MFMGAKCCKFGLVCYETSGKWSQCR
```

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

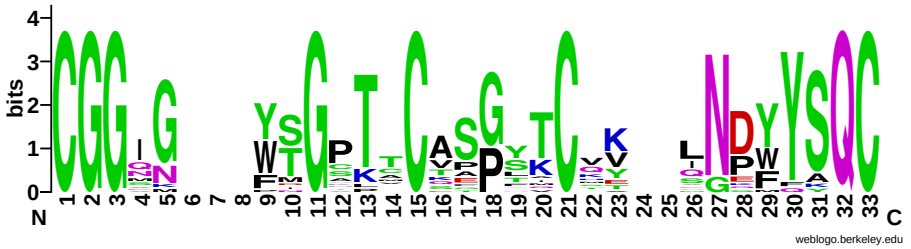
```
      +----------------+
      |          +-----|---------+
      |          |     |         |
xxxxxxCxxxxxxxxxxCxxxxxCxxxxxxxxxCx
      ****************************
```

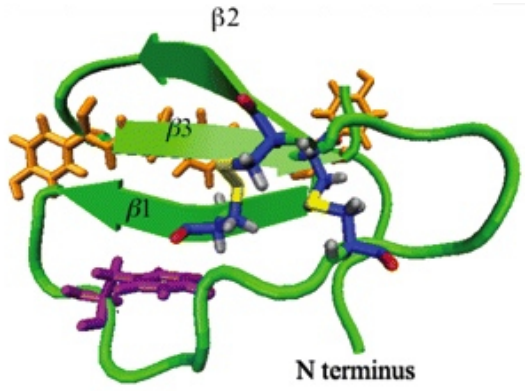
les 4 cystéines sont impliquées dans des liaisons di-sulfures

Site de fixation de la cellulose

```
HWGQCGGI---GYSGCKTCTSGTTCQYSNDYYSQCL
HYGQCGGI---GYSGPTVCASGTTCQVLNPYYSQCL
QWGQCGGI---GYTGSTTCASPYTCHVLNPYYSQCY
VWGQCGGQ---NWSGPTCCASGSTCVYSNDYYSQCL
LYGQCGGA---GWTGPTTCQAPGTCKVQNQWYSQCL
IWGQCGGN---GWTGATTCASGLKCEKINDWYYQCV
VWGQCGGN---GWTGPTTCASGSTCVKQNDFYSQCL
DWAQCGGN---GWTGPTTCVSPYTCTKQNDWYSQCL
QWGQCGGQ---NYSGPTTCKSPFTCKKINDFYSQCQ
RWQQCGGI---GFTGPTQCEEPYICTKLNDWYSQCL
HWAQCGGI---GFSGPTTCPEPYTCAKDHDIYSQCV
LYEQCGGI---GFDGVTCCSEGLMCMKMGPYYSQCR
VWAQCGGQ---NWSGTPCCTSGNKCVKLNDFYSQCQ
PYGQCGGM---NYSGKTMCSPGFKCVELNEFFSQCD
AYYQCGGSKSAYPNGNLACATGSKCVKQNEYYSQCV
EYAACGGE---MFMGAKCCKFGLVCYETSGKWSQCR
```

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

```
      +----------------+
      |          +-----|---------+
      |          |     |         |
xxxxxxCxxxxxxxxxxCxxxxxCxxxxxxxxxCx
      ****************************
```

les 4 cystéines sont impliquées dans des liaisons di-sulfures

Structure d'ARN

Structure d'ARN

- on dispose d'une famille d'ARN possédant la même structure

```
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
```

Structure d'ARN

- on dispose d'une famille d'ARN possédant la même structure
- pour un appariement de la structure donné :

  si une base mute dans la structure d'ARN, la base qui s'y apparie doit muter aussi . . .

```
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
```

Structure d'ARN - méthode comparative

1. contruction de l'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  -
```

2. détection de positions corrélées

Structure d'ARN - méthode comparative

1. contruction de l'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	-

2. détection de positions corrélées

Structure d'ARN - méthode comparative (ARNt)

```
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTCAGCGGTTCGATCCCGCTTAGCTCCACCA
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTCGATTCCCCTAGGCTCCA---
GGGGGTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAGTCAGCGGTTCGATTCCGCTTACCCCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGGAGTTCGATCCTCCTTGGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCTGGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAACGGTTCGATCCCGTTTGGCTCCA---
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGCGGTTCGATCCCGCTATTCTCCACCA
GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTCGGTTCGAATCCGTCTGGCTCCACCA
GGGGCCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGGAGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
```

Structure d'ARN - méthode comparative (ARNt)

```
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTCAGCGGTTCGATCCCGCTTAGCTCCACCA
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTCGATTCCCCTAGGCTCCA---
GGGGGTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAGTCAGCGGTTCGATTCCGCTTACCCCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGGAGTTCGATCCTCCTTGGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCTGGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAACGGTTCGATCCCGTTTGGCTCCA---
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGCGGTTCGATCCCGCTATTCTCCACCA
GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTCGGTTCGAATCCGTCTGGCTCCACCA
GGGGCCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGGAGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
```

Structure d'ARN - méthode comparative (ARNt)

```
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTCAGCGGTTCGATCCCGCTTAGCTCCACCA
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTCGATTCCCCTAGGCTCCA---
GGGGGTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAGTCAGCGGTTCGATTCCGCTTACCCCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGGAGTTCGATCCTCCTTGGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCTGGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAACGGTTCGATCCCGTTTGGCTCCA---
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGCGGTTCGATCCCGCTATTCTCCACCA
GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTCGGTTCGAATCCGTCTGGCTCCACCA
GGGGCCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGGAGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
```

Structure d'ARN - méthode comparative (ARNt)

```
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTCAGCGGTTCGATCCCGCTTAGCTCCACCA
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTCGATTCCCCTAGGCTCCA---
GGGGGTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAGTCAGCGGTTCGATTCCGCTTACCCCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGGAGTTCGATCCTCCTTGGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCTGGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAACGGTTCGATCCCGTTTGGCTCCA---
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGCGGTTCGATCCCGCTATTCTCCACCA
GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTCGGTTCGAATCCGTCTGGCTCCACCA
GGGGCCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGGAGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
```

Structure d'ARN - méthode comparative (ARNt)

```
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTCAGCGGTTCGATCCCGCTTAGCTCCACCA
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTCGATTCCCCTAGGCTCCA---
GGGGGTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAGTCAGCGGTTCGATTCCGCTTACCCCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGGAGTTCGATCCTCCTTGGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCTGGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAACGGTTCGATCCCGTTTGGCTCCA---
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGCGGTTCGATCCCGCTATTCTCCACCA
GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTCGGTTCGAATCCGTCTGGCTCCACCA
GGGGCCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGGAGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
```

Structure d'ARN - méthode comparative (ARNt)

```
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTCAGCGGTTCGATCCCGCTTAGCTCCACCA
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTCGATTCCCCTAGGCTCCA---
GGGGGTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAGTCAGCGGTTCGATTCCGCTTACCCCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGGAGTTCGATCCTCCTTGGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCTGGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAACGGTTCGATCCCGTTTGGCTCCA---
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGCGGTTCGATCCCGCTATTCTCCACCA
GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTCGGTTCGAATCCGTCTGGCTCCACCA
GGGGCCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGGAGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
```

Structure d'ARN - méthode comparative (ARNt)

```
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTCAGCGGTTCGATCCCGCTTAGCTCCACCA
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTCGATTCCCCTAGGCTCCA---
GGGGGTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAGTCAGCGGTTCGATTCCGCTTACCCCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGGAGTTCGATCCTCCTTGGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCTGGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAACGGTTCGATCCCGTTTGGCTCCA---
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGCGGTTCGATCCCGCTATTCTCCACCA
GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTCGGTTCGAATCCGTCTGGCTCCACCA
GGGGCCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGGAGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
```

Structure d'ARN - méthode comparative (ARNt)

```
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTCAGCGGTTCGATCCCGCTTAGCTCCACCA
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTCGATTCCCCTAGGCTCCA---
GGGGGTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAGTCAGCGGTTCGATTCCGCTTACCCCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGGAGTTCGATCCTCCTTGGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCTGGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAACGGTTCGATCCCGTTTGGCTCCA---
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGCGGTTCGATCCCGCTATTCTCCACCA
GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTCGGTTCGAATCCGTCTGGCTCCACCA
GGGGCCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGGAGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
```

Structure d'ARN - méthode comparative (ARNt)

```
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTCAGCGGTTCGATCCCGCTTAGCTCCACCA
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTCGATTCCCCTAGGCTCCA---
GGGGGTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAGTCAGCGGTTCGATTCCGCTTACCCCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGGAGTTCGATCCTCCTTGGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCTGGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAACGGTTCGATCCCGTTTGGCTCCA---
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGCGGTTCGATCCCGCTATTCTCCACCA
GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTCGGTTCGAATCCGTCTGGCTCCACCA
GGGGCCCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGGAGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
```

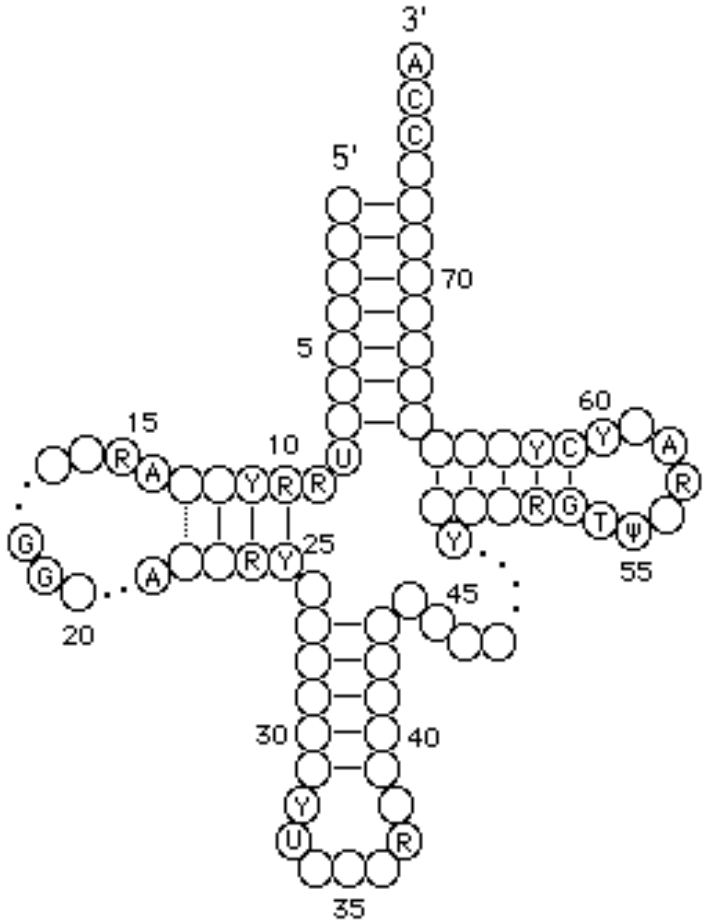
Structure d'ARN - méthode comparative (ARNt)

```
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTCAGCGGTTCGATCCCGCTTAGCTCCACCA
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTCGATTCCCCTAGGCTCCA---
GGGGGTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAGTCAGCGGTTCGATTCCGCTTACCCCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGGAGTTCGATCCTCCTTGGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCTGGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAACGGTTCGATCCCGTTTGGCTCCA---
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGCGGTTCGATCCCGCTATTCTCCACCA
GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTCGGTTCGAATCCGTCTGGCTCCACCA
GGGGCCCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGGAGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
```

Structure d'ARN - méthode comparative (ARNt)

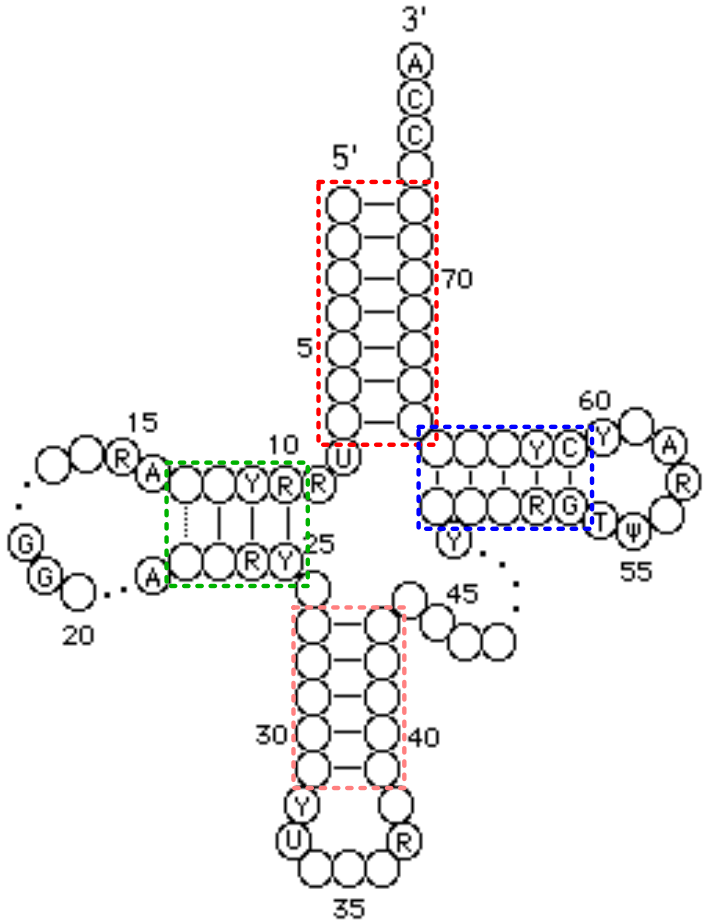
```
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTCAGCGGTTCGATCCCGCTTAGCTCCACCA
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTCGATCCCGCTATTCTCCA---
GGGGCCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTCGATTCCCCTAGGCTCCA---
GGGGGTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAGTCAGCGGTTCGATTCCGCTTACCCCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGGAGTTCGATCCTCCTTGGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCTGGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAACGGTTCGATCCCGTTTGGCTCCA---
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA
GGGGCATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAGCGGTTCGATCCCGCTATTCTCCACCA
GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTCGGTTCGAATCCGTCTGGCTCCACCA
GGGGCCCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGGAGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
GGGGCCCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTCGATCCCGTCCGGCTCCACCA
```

Structure d'ARN - méthode comparative (ARNt)



Structure secondaire de l'ARNt

Structure d'ARN - méthode comparative (ARNt)



Structure secondaire de l'ARNt

Démarche

Alignement 2 à 2

Deux séquences quelconques

↓

Détecter une similarité **syntaxique**

↓

Y a-t-il une **fonction** commune ?

Alignement multiple

Famille de séquences avec la même **fonction**

↓

À quelle conservation **syntaxique** cela correspond-il ?

Score d'un alignement multiple

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

$\Rightarrow$ la somme des scores associés à chaque colonne

somme des paires (Sum of Pairs)

$$SP(m_i) = \sum_{1 \leq j < k \leq n} s(m_i^j, m_i^k)$$

m_i = la i-ème colonne de l'alignement

m_i^j = j-ème aa dans la colonne i

Exemple

jeu de scores :

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

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

Somme des paires

Définition alternative (équivalente)

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

$$SP(\alpha) = \sum_{1 \leq i < j \leq n} score(\alpha_{ij})$$

Exemple

jeu de scores :

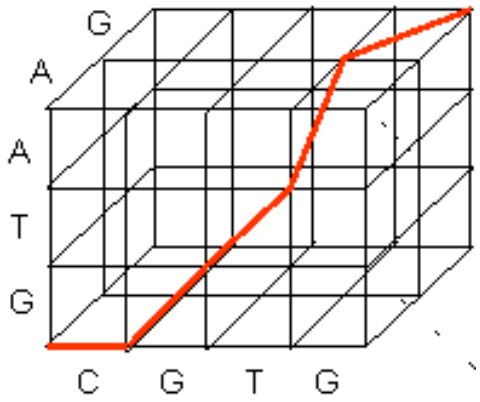
$$s(x,x)=1, \quad s(x,y)=-1, \quad s(x,-)=s(-,x)=-2, \quad s(-,-)=0$$

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

Algorithme exact

$\Rightarrow$ par programmation dynamique

- alignement deux à deux $\Rightarrow$ chemin dans une matrice de dimension 2
- alignement multiple de n séquences $\Rightarrow$ chemin dans une matrice de dimension n

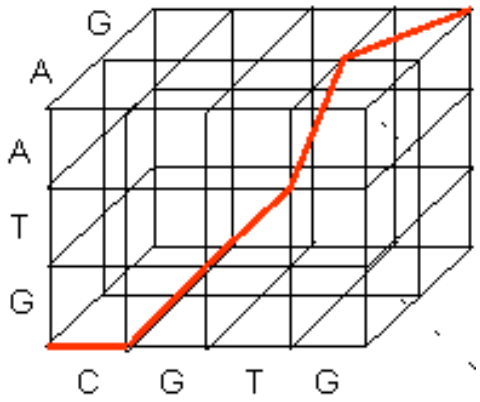



```
C   G   T   -   G
-   G   T   A   -
-   -   -   A   G
```

Algorithme exact

$\Rightarrow$ par programmation dynamique

- alignement deux à deux $\Rightarrow$ chemin dans une matrice de dimension 2
- alignement multiple de n séquences $\Rightarrow$ chemin dans une matrice de dimension n




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

$\Rightarrow$ impossible de l'utiliser en pratique.

$\approx$ 100 ans pour 8 séquences de 100bp.

Utilisation d'heuristiques

Definition (Heuristique)

Algorithme utilisant des règles simples pour diminuer l'espace de recherche des solutions (mais ne donnant pas forcément la meilleure solution)

- Clustal (*le plus populaire : Clustal, ClustalW, Clustal-Omega*)
- Dialign2 (*complémentaire à Clustal*)
- T-coffee, Muscle, Pima, Multalin, MA-FFT...

$\Rightarrow$ autant de programmes qui produisent des alignements différents !

Alignement basé sur un arbre guide

- idée : reconstruire l'alignement multiple à partir d'un arbre guide (clusters)
  - feuilles = séquences
  - noeuds = alignements
- partir des feuilles puis remonter dans l'arbre
  - utilisation de la technique de profile alignment $\Rightarrow$ produire un seul alignement multiple avec deux (par prog. dyn.)

MultAlin

CLUSTer + ALignement $\Rightarrow$ CLUSTAL

MultAlin

F. Corpet, 1988

principe :

1. calcule une matrice de similarité des paires
2. construit un arbre de clustering hierarchique (UPGMA)
3. construit l'alignement multiple en suivant l'arbre
4. reconstruit une arbre de clustering hierarchique avec les nouveaux alignements paire à paire issus de l'alignement trouvé
5. réitère le processus jusqu'à stabilisation de l'arbre de clustering

MultAlin - fonctionnement de UPGMA

- agglomère les 2 séquences de score maximal (~ distance minimale dans UPGMA)
- calcule les nouveaux scores entre ce cluster et les autres en faisant la moyenne

$$s(C_1, C_2) = \frac{1}{\mathrm{Card}(C_1) \times \mathrm{Card}(C_2)} \sum_{c \in C_1, c' \in C_2} s(c, c')$$

MultAlin - exemple 1

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

MultAlin - exemple 1

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

1. calcul des meilleurs alignements 2 à 2 :

scores (Mach = 1, Mismatch =-1, Indel = -1)

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

MultAlin - exemple 1

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

1. calcul des meilleurs alignements 2 à 2 :

   scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

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

MultAlin - exemple 1

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

1. calcul des meilleurs alignements 2 à 2 :

scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

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

MultAlin - exemple 1

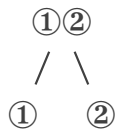
① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

1. calcul des meilleurs alignements 2 à 2 :

scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

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



MultAlin - exemple 1

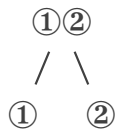
① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

1. calcul des meilleurs alignements 2 à 2 :

scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

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



① TACCATGA
② TACCAT-A

MultAlin - exemple 1

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

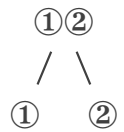
1. calcul des meilleurs alignements 2 à 2 :

scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

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

	①②	③	④
①②	.	0	2.5
③	.	.	4
④	.	.	.



① TACCATGA
② TACCAT-A

MultAlin - exemple 1

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

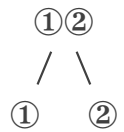
1. calcul des meilleurs alignements 2 à 2 :

scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

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

	①②	③	④
①②	.	0	2.5
③	.	.	4
④	.	.	.



① TACCATGA
② TACCAT-A

MultAlin - exemple 1

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

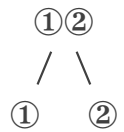
1. calcul des meilleurs alignements 2 à 2 :

scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

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

	①②	③	④
①②	.	0	2.5
③	.	.	4
④	.	.	.



① TACCATGA
② TACCAT-A

MultAlin - exemple 1

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

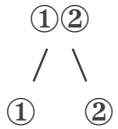
1. calcul des meilleurs alignements 2 à 2 :

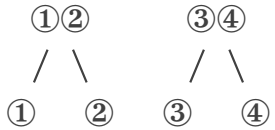
scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

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

	①②	③	④
①②	.	0	2.5
③	.	.	4
④	.	.	.





① TACCATGA
② TACCAT-A

MultAlin - exemple 1

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

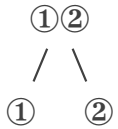
1. calcul des meilleurs alignements 2 à 2 :

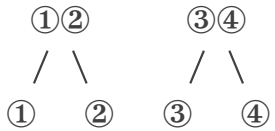
scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

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

	①②	③	④
①②	.	0	2.5
③	.	.	4
④	.	.	.





① TACCATGA
② TACCAT-A

③ GACGA-C-CA
④ GACCATCTCA

MultAlin - exemple 1

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

1. calcul des meilleurs alignements 2 à 2 :

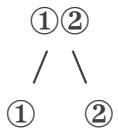
scores (Mach = 1, Mismatch =-1, Indel = -1)

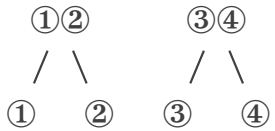
2. construction d'un arbre de clustering (et de l'alignement) :

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

	①②	③	④
①②	.	0	2.5
③	.	.	4
④	.	.	.

	① ②	③④
①②	.	1.25
③④	.	.





① TACCATGA
② TACCAT-A

③ GACGA-C-CA
④ GACCATCTCA

MultAlin - exemple 1

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

1. calcul des meilleurs alignements 2 à 2 :

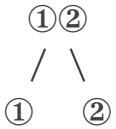
scores (Mach = 1, Mismatch =-1, Indel = -1)

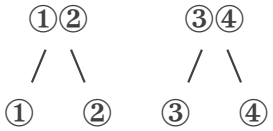
2. construction d'un arbre de clustering (et de l'alignement) :

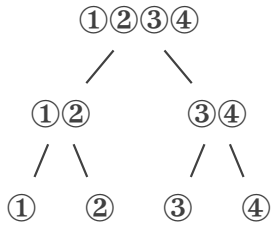
	①	②	③	④
①	.	6	0	2
②	.	.	0	3
③	.	.	.	4
④	.	.	.	.

	①②	③	④
①②	.	0	2.5
③	.	.	4
④	.	.	.

	① ②	③④
①②	.	1.25
③④	.	.







① TACCATGA
② TACCAT-A

③ GACGA-C-CA
④ GACCATCTCA

MultAlin - exemple 1

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

1. calcul des meilleurs alignements 2 à 2 :

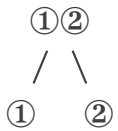
scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

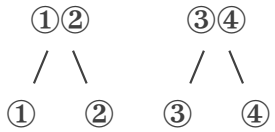
	①	②	③	④
①	.	6	0	2
②	.	.	0	3
③	.	.	.	4
④	.	.	.	.

	①②	③	④
①②	.	0	2.5
③	.	.	4
④	.	.	.

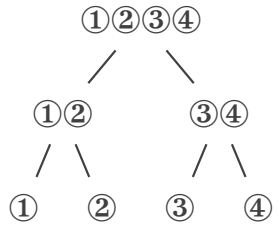
	① ②	③④
①②	.	1.25
③④	.	.



① TACCATGA
② TACCAT-A



③ GACGA-C-CA
④ GACCATCTCA



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

MultAlin - exemple 1

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

1. calcul des meilleurs alignements 2 à 2 :

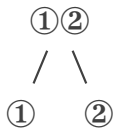
scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

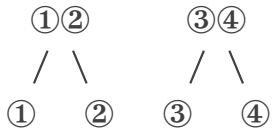
	①	②	③	④
①	.	6	0	2
②	.	.	0	3
③	.	.	.	4
④	.	.	.	.

	①②	③	④
①②	.	0	2.5
③	.	.	4
④	.	.	.

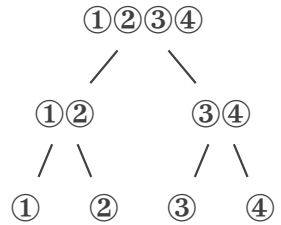
	① ②	③④
①②	.	1.25
③④	.	.



① TACCATGA
② TACCAT-A



③ GACGA-C-CA
④ GACCATCTCA



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

3. nouvelle matrice des scores et on recommence :

MultAlin - exemple 1

① TACCATGA ② TACCATA ③ GACGACCA ④ GACCATCTCA

1. calcul des meilleurs alignements 2 à 2 :

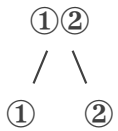
scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

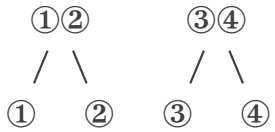
	①	②	③	④
①	.	6	0	2
②	.	.	0	3
③	.	.	.	4
④	.	.	.	.

	①②	③	④
①②	.	0	2.5
③	.	.	4
④	.	.	.

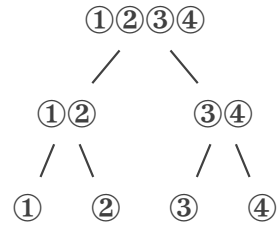
	① ②	③④
①②	.	1.25
③④	.	.



① TACCATGA
② TACCAT-A



③ GACGA-C-CA
④ GACCATCTCA



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

3. nouvelle matrice des scores et on recommence :

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

MultAlin - exemple 2

① TACCATGA ② TACCATA ③ GACGACCA ④ TACGATCGA

MultAlin - exemple 2

① TACCATGA ② TACCATA ③ GACGACCA ④ TACGATCGA

1. calcul des meilleurs alignements 2 à 2 :

scores (Mach = 1, Mismatch =-1, Indel = -1)

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

MultAlin - exemple 2

① TACCATGA ② TACCATA ③ GACGACCA ④ TACGATCGA

1. calcul des meilleurs alignements 2 à 2 :

scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

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

MultAlin - exemple 2

① TACCATGA ② TACCATA ③ GACGACCA ④ TACGATCGA

1. calcul des meilleurs alignements 2 à 2 :

scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

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

MultAlin - exemple 2

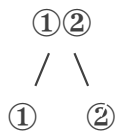
① TACCATGA ② TACCATA ③ GACGACCA ④ TACGATCGA

1. calcul des meilleurs alignements 2 à 2 :

scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

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



MultAlin - exemple 2

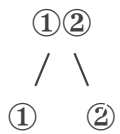
① TACCATGA ② TACCATA ③ GACGACCA ④ TACGATCGA

1. calcul des meilleurs alignements 2 à 2 :

scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

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



① TACCATGA
② TACCAT-A

MultAlin - exemple 2

① TACCATGA ② TACCATA ③ GACGACCA ④ TACGATCGA

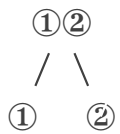
1. calcul des meilleurs alignements 2 à 2 :

scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

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

	①②	③	④
①②	.	0	4
③	.	.	3
④	.	.	.



① TACCATGA
② TACCAT-A

MultAlin - exemple 2

① TACCATGA ② TACCATA ③ GACGACCA ④ TACGATCGA

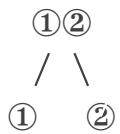
1. calcul des meilleurs alignements 2 à 2 :

scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

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

	①②	③	④
①②	.	0	4
③	.	.	3
④	.	.	.



① TACCATGA
② TACCAT-A

MultAlin - exemple 2

① TACCATGA ② TACCATA ③ GACGACCA ④ TACGATCGA

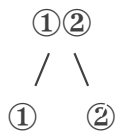
1. calcul des meilleurs alignements 2 à 2 :

scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

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

	①②	③	④
①②	.	0	4
③	.	.	3
④	.	.	.



① TACCATGA
② TACCAT-A

MultAlin - exemple 2

① TACCATGA ② TACCATA ③ GACGACCA ④ TACGATCGA

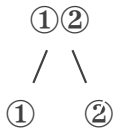
1. calcul des meilleurs alignements 2 à 2 :

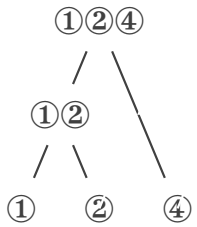
scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

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

	①②	③	④
①②	.	0	4
③	.	.	3
④	.	.	.





① TACCATGA
② TACCAT-A

MultAlin - exemple 2

① TACCATGA ② TACCATA ③ GACGACCA ④ TACGATCGA

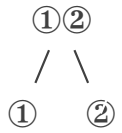
1. calcul des meilleurs alignements 2 à 2 :

scores (Mach = 1, Mismatch =-1, Indel = -1)

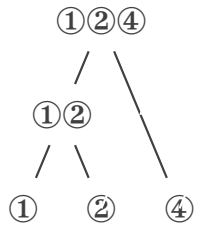
2. construction d'un arbre de clustering (et de l'alignement) :

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

	①②	③	④
①②	.	0	4
③	.	.	3
④	.	.	.



① TACCATGA
② TACCAT-A



① TACCAT-GA
② TACCAT--A
④ TACGATCGA

MultAlin - exemple 2

① TACCATGA ② TACCATA ③ GACGACCA ④ TACGATCGA

1. calcul des meilleurs alignements 2 à 2 :

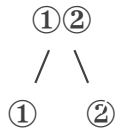
scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

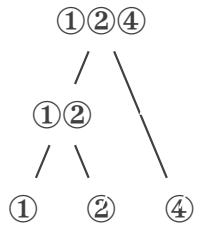
	①	②	③	④
①	.	6	0	5
②	.	.	0	3
③	.	.	.	3
④	.	.	.	.

	①②	③	④
①②	.	0	4
③	.	.	3
④	.	.	.

	①②④	③
①②④	.	1
③	.	.



① TACCATGA
② TACCAT-A



① TACCAT-GA
② TACCAT--A
④ TACGATCGA

MultAlin - exemple 2

① TACCATGA ② TACCATA ③ GACGACCA ④ TACGATCGA

1. calcul des meilleurs alignements 2 à 2 :

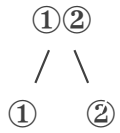
scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

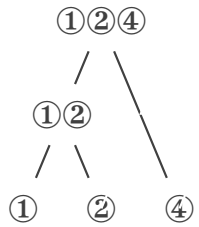
	①	②	③	④
①	.	6	0	5
②	.	.	0	3
③	.	.	.	3
④	.	.	.	.

	①②	③	④
①②	.	0	4
③	.	.	3
④	.	.	.

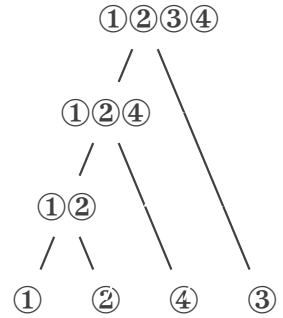
	①②④	③
①②④	.	1
③	.	.



① TACCATGA
② TACCAT-A



① TACCAT-GA
② TACCAT--A
④ TACGATCGA



MultAlin - exemple 2

① TACCATGA ② TACCATA ③ GACGACCA ④ TACGATCGA

1. calcul des meilleurs alignements 2 à 2 :

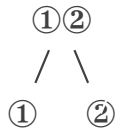
scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

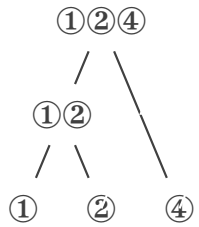
	①	②	③	④
①	.	6	0	5
②	.	.	0	3
③	.	.	.	3
④	.	.	.	.

	①②	③	④
①②	.	0	4
③	.	.	3
④	.	.	.

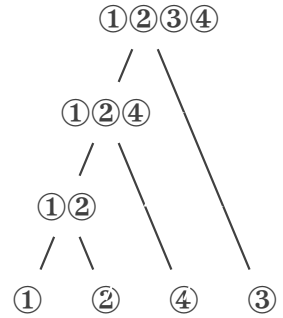
	①②④	③
①②④	.	1
③	.	.



① TACCATGA
② TACCAT-A



① TACCAT-GA
② TACCAT--A
④ TACGATCGA



① TACCAT-GA
② TACCAT--A
④ TACGATCGA
③ GACGACC-A

MultAlin - exemple 2

① TACCATGA ② TACCATA ③ GACGACCA ④ TACGATCGA

1. calcul des meilleurs alignements 2 à 2 :

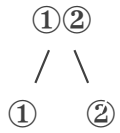
scores (Mach = 1, Mismatch =-1, Indel = -1)

2. construction d'un arbre de clustering (et de l'alignement) :

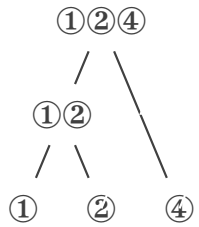
	①	②	③	④
①	.	6	0	5
②	.	.	0	3
③	.	.	.	3
④	.	.	.	.

	①②	③	④
①②	.	0	4
③	.	.	3
④	.	.	.

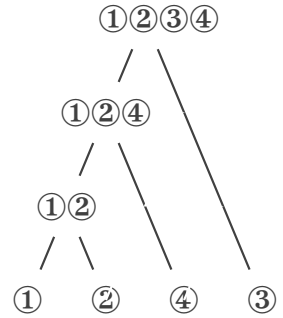
	①②④	③
①②④	.	1
③	.	.



① TACCATGA
② TACCAT-A



① TACCAT-GA
② TACCAT--A
④ TACGATCGA



① TACCAT-GA
② TACCAT--A
④ TACGATCGA
③ GACGACC-A

CLUSTAL

Thompson et al., 1994

principe :

1. calcule une matrice de similarité des paires par prog. dyn.
2. convertit les similarités en distances
3. construit l'arbre guide (méthode du Neighbor-Joining)
4. aligne progressivement les noeuds de l'arbre par ordre décroissant de similarité

CLUSTAL - exemple

- 4 séquences

s_1 `cgatgagtcattgtgactg`
s_2 `cgagccattgtagctactg`
s_3 `cgaccattgtagctacctg`
s_4 `cgatgagtcactgtgactg`

- jeu de scores :

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

Etape 1

calcul des scores de similarité de tous les alignements

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

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

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

```
s2  cgagccattgtagctac-tg
    ||| |||||||||||||| ||
s3  cga-ccattgtagctacctg

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

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

tableau des scores d'alignement :

	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

$\downarrow$

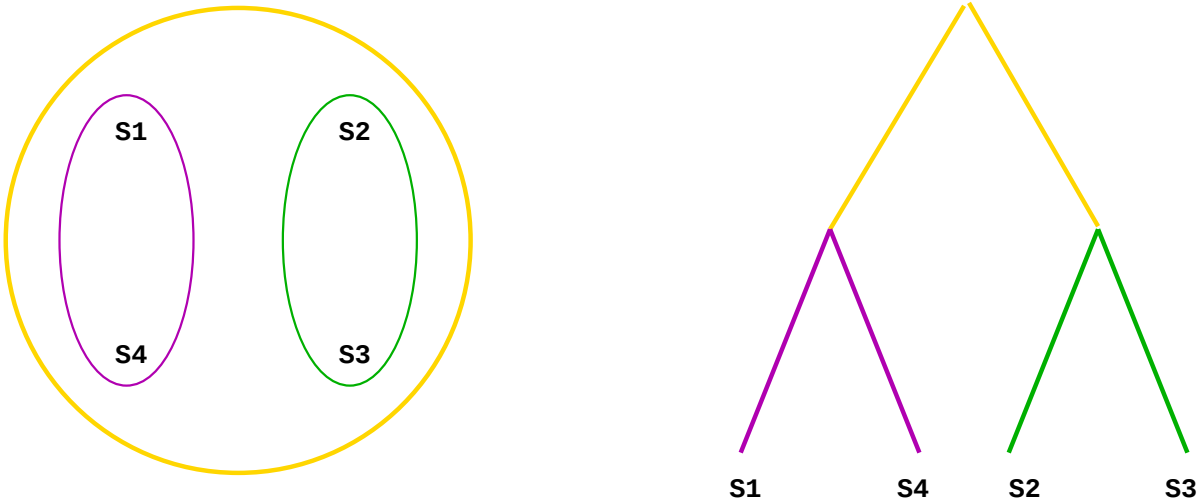
$\frac{n(n-1)}{2}$ calculs

Etape 2

construction de l'arbre guide

arbre obtenu avec l'algorithme de Neighbor-Joining

dendrogramme



regroupement des séquences suivant leur similarité à partir de la matrice des scores 2 à 2.

Etape 3

construction de l'alignement multiple final

s1
cgatgagtcattgtgactg

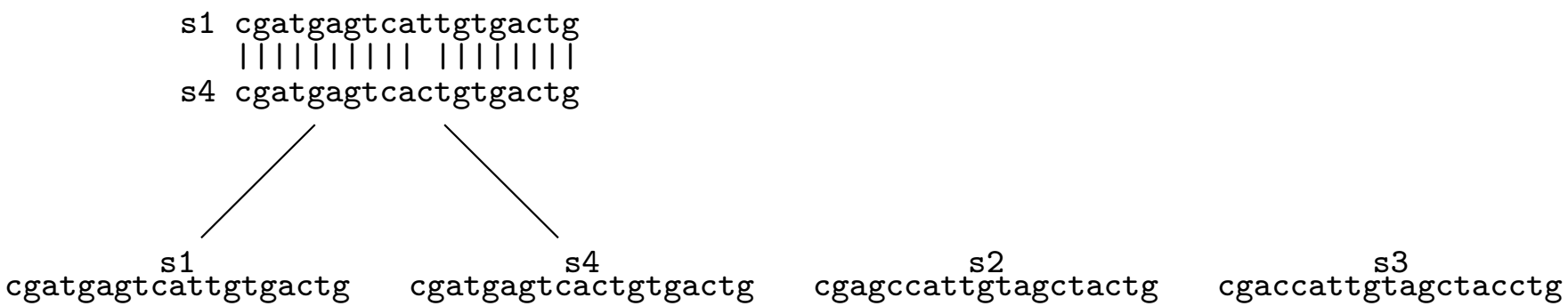
s4
cgatgagtcactgtgactg

s2
cgagccattgtagctactg

s3
cgaccattgtagctacctg

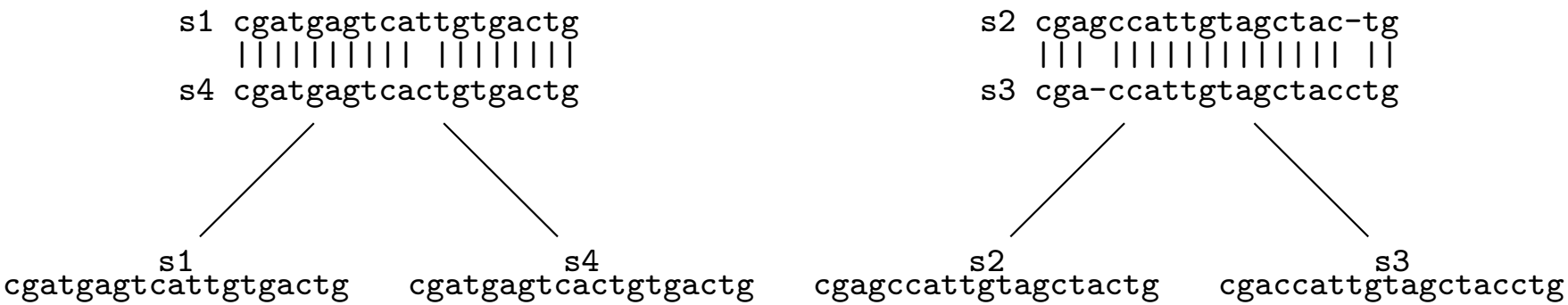
Etape 3

construction de l'alignement multiple final



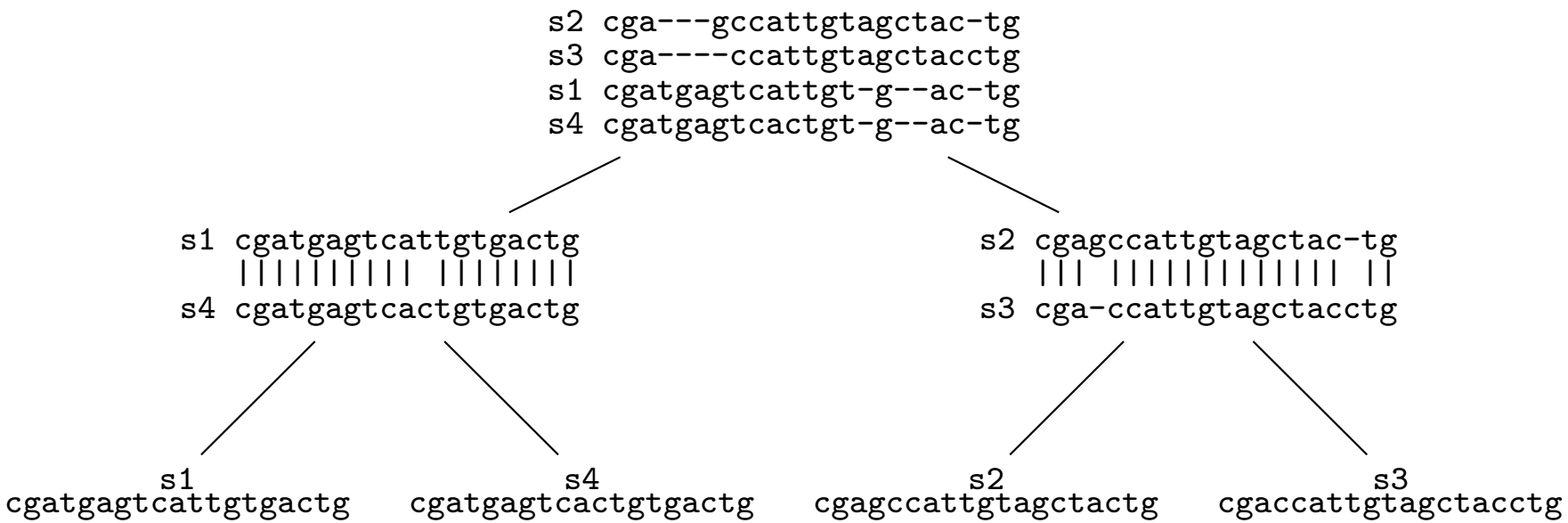
Etape 3

construction de l'alignement multiple final



Etape 3

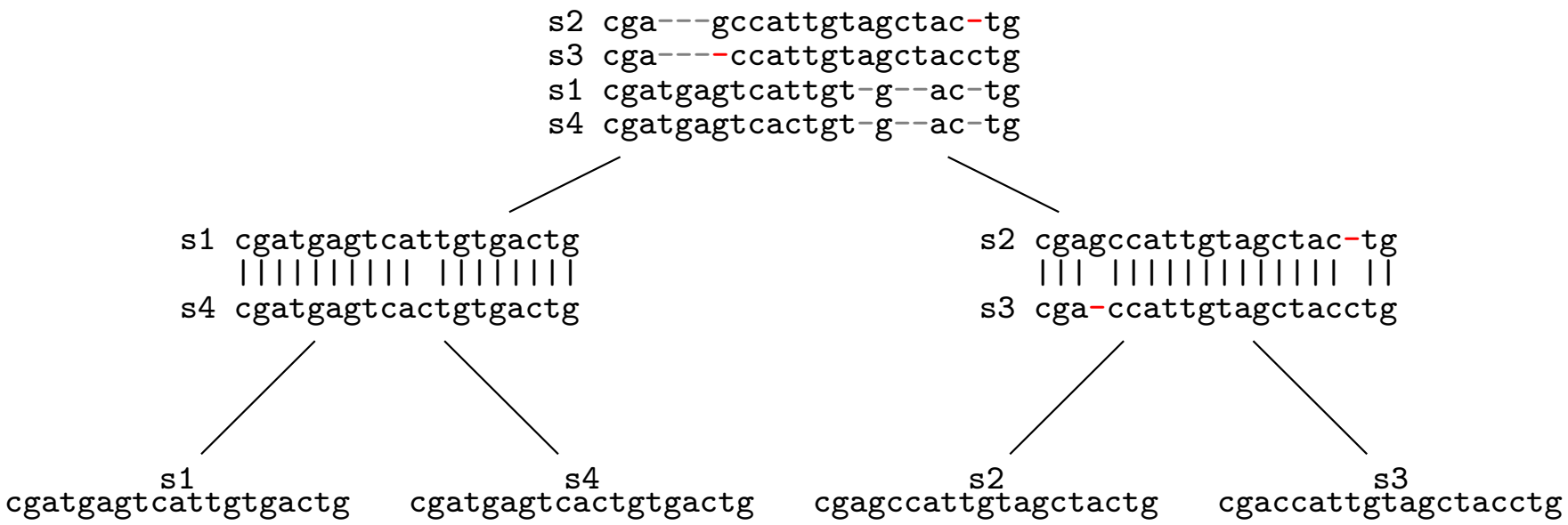
construction de l'alignement multiple final



Etape 3

construction de l'alignement multiple final

"Once a gap, always a gap."



ClustalW est optimisé pour les protéines

- Pondération des séquences en fonction de leur sur- ou sous-représentation
- Adaptation des matrices de similarité au fil de l'algorithme en fonction de la divergence des séquences à aligner
  BLOSUM 80 pour aligner les séquences proches,

  BLOSUM 50 pour aligner des séquences distantes, par exemple.
- Pénalités de gaps spécifiques à chaque résidu.

  Par exemple, les glycines sont davantage susceptibles d'avoisiner un gap que les valines.
- Pénalités de gaps réduites dans les régions hydrophiles

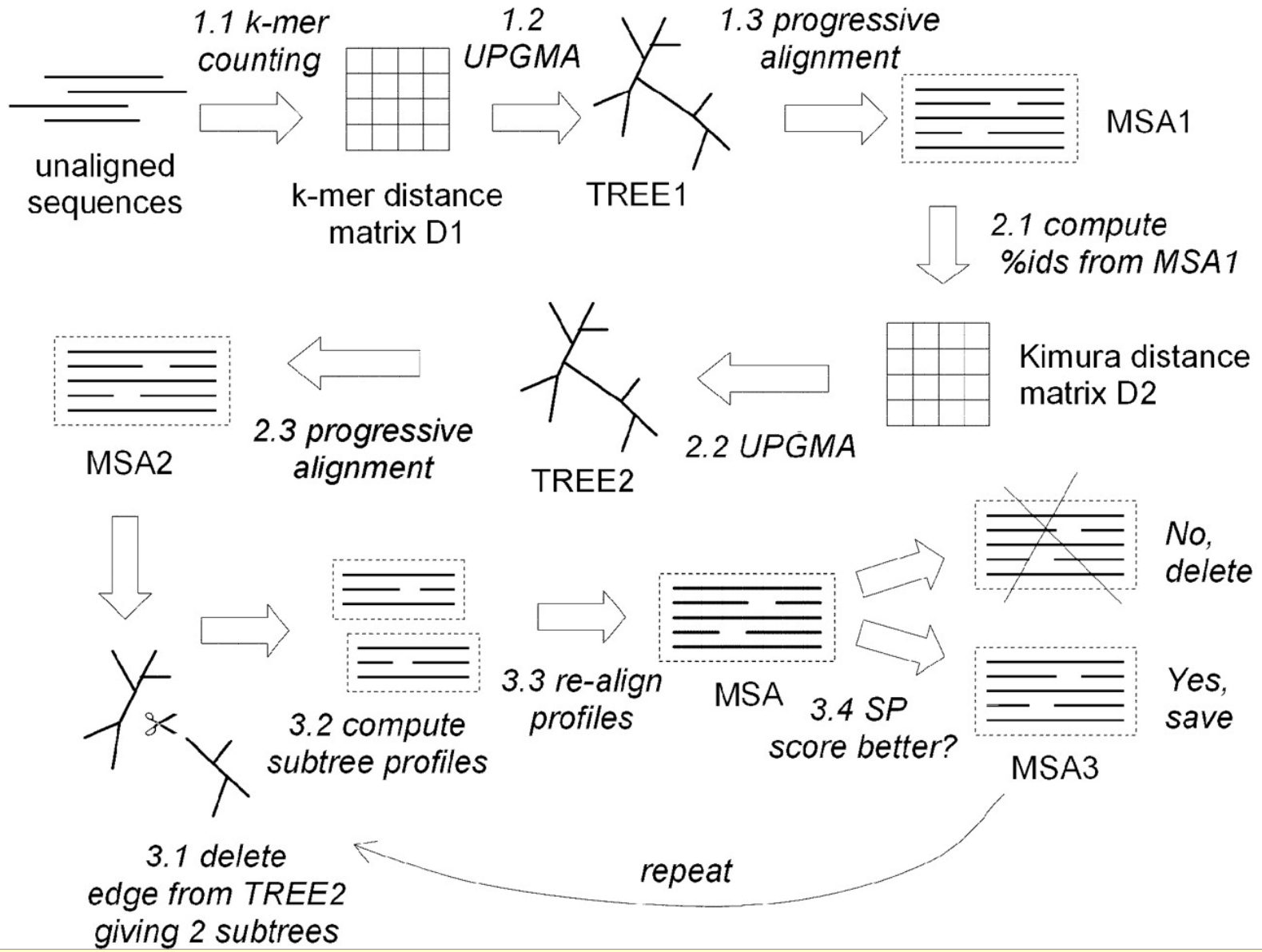
  Encourage la formation de gaps dans des boucles plutôt que dans des régions structurées.
- Pénalités de gaps augmentées dans le voisinage d'autres gaps

  Evite la formation de petits gaps voisins, au profit de longs gaps

Paramètres de Clustal

- *Slow/Fast* : qualité des alignements 2 à 2
- Matrice de similarité (PAM, BLOSUM, Gonnet)
- Pénalités de gaps :
  - Ouverture et d'extension de gaps
  - Distance de voisinage entre deux gaps
  - Gaps hydrophiles
  - Ouvertures de gaps spécifiques

Arbre guide : autres méthodes - Muscle



Arbre guide : autres méthodes - MA-FFT

FFT : Fast Fourier Transform

- *Progressive alignment* :
  - 1/2 arbre UPGMA à l'aide d'un distance rapide (k-mots).
  - 2/2 construction guidée d'un 1er alignement multiple (FFT-NS1).
- *Iterative refinement* :
  1. réutiliser la matrice de distance de FFT-NS1 pour refaire un alignement multiple (FFT-NS2).
  2. heuristique de réalignement par groupes (FFT-NSi).

Arbre guide : autres méthodes - Clustal-Omega

alignement multiple de n séquences

Arbre guide : autres méthodes - Clustal-Omega

alignement multiple de n séquences

construction de l'arbre guide $\Longrightarrow$ $\frac{n(n-1)}{2}$ *"comparaisons"* (alignements, distances par k-mots)

Arbre guide : autres méthodes - Clustal-Omega

alignement multiple de n séquences

construction de l'arbre guide $\Longrightarrow$ $\frac{n(n-1)}{2}$ *"comparaisons"* (alignements, distances par k-mots)

Construction très longue si $n > 1000$ séquences

(l'alignement est souvent plus rapide que la construction de l'arbre guide).

Arbre guide : autres méthodes - Clustal-Omega

alignement multiple de n séquences

construction de l'arbre guide $\Longrightarrow$ $\frac{n(n-1)}{2}$ *"comparaisons"* (alignements, distances par k-mots)

Construction très longue si $n > 1000$ séquences

(l'alignement est souvent plus rapide que la construction de l'arbre guide).

1. Eviter ces $\mathcal{O}(n^2)$ comparaisons $\rightarrow \mathcal{O}(n \log(n))$
2. Utilisation d'un alignement simple de *profils-HMM (voir prochain cours)* plutôt que de *blocks d'alignements*.

À partir des alignements locaux

- idée : repérer des similarités locales fortes entre les séquences
- typiquement : les diagonales du dotplot
- incorporer les diagonales dans l'alignement multiple
- conséquence : les gaps inter-diagonales sont considérés moins importants

DIagonal + ALIGNment $\Rightarrow$ DIALIGN

DIALIGN

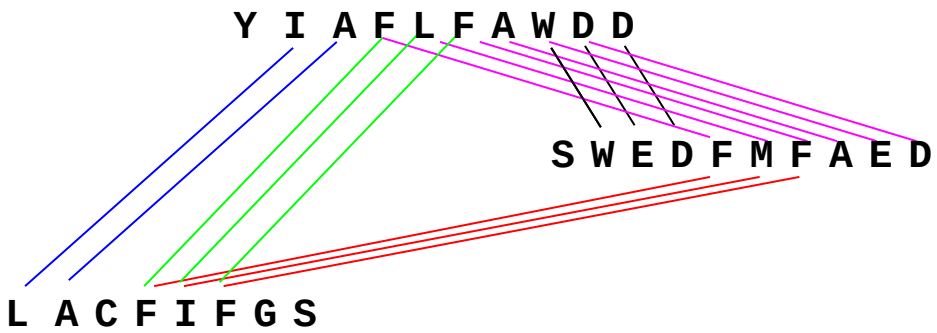
Morgenstern et al., 1996

principe :

1. alignement des paires avec optimisation des poids des diagonales
2. tri des diagonales selon leur poids et leur chevauchement
3. reconstruction gloutonne
    1. insertion des diagonales par poids décroissants
    2. vérification de la consistance avec les diagonales déjà introduites
4. recommencer

DIALIGN

- Étape 1 : détection des diagonales dans les paires de séquences



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

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

CLUSTAL vs. DIALIGN

Exemple (C. Notre-Dame)

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

CLUSTAL vs. DIALIGN

Exemple (C. Notre-Dame)

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

Alignement fourni par Clustal

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

CLUSTAL vs. DIALIGN

Exemple (C. Notre-Dame)

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

Alignement fourni par Clustal

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

Alignement fourni par Dialign2

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

Quelle méthode utiliser ? (1/2)

$\Rightarrow$ cela dépend du type de séquences à aligner . . .

BaliBASE : base de données d'alignements multiples pour *benchmark*

- plus de 150 familles de protéines
- alignements basés sur la structure secondaire

Référence 1	séquences équidistantes avec différents niveaux de conservation
Référence 2	protéines homologues + 1 séquence orpheline
Référence 3	sous-groupes avec moins de 25% d'identité entre les groupes
Référence 4	extensions N/C-terminales
Référence 5	insertions internes
Référence 6	répétitions internes
Référence 7	protéines transmembranaires
Référence 8	permutations de domaines

- *Réf. 1, 2 et 3 :* préférer **Clustal** à **Dialign2**
- *Réf. 4 et 5 :* préférer **Dialign2** à **Clustal**

Quelle méthode utiliser ? (2/2)

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

Pas de méthode universelle
Pas de confiance aveugle vis-à-vis du résultat obtenu

Exemple : domaine SH3

SH3 (Src homology 3) domains are often indicative of a protein involved in signal transduction related to cytoskeletal organization. The SH3 domain has a characteristic fold which consists of five or six beta- strands arranged as two tightly packed anti-parallel beta sheets. The linker regions may contain short helices.

Prosite PS50002

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

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

SH3 - Véritable Alignement

basé sur l'alignement des éléments de structure secondaire

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

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

SH3 - Alignement fourni par Clustal

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

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

SH3 - Alignement fourni par Dialign2

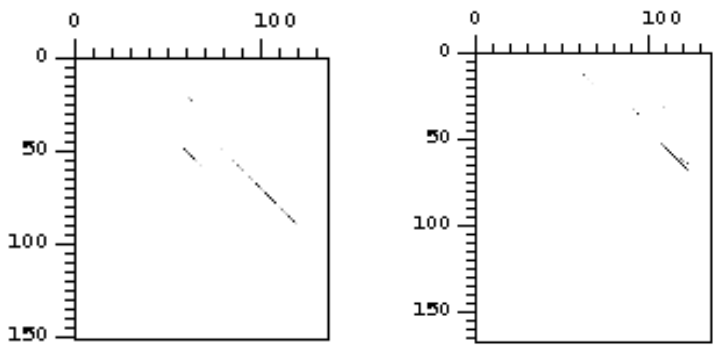
```
1aboA   n-LFVALYDFVASGDNTLSITKGEKLRVL--------------------
1ycsB   kgVIYALWDYEPQNDDELPMKEGDCMTIIhr----EDEDEI--------
1pht    gyQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDgqearpeei
1ihvA   --NFRV---YYRDSRDPVWKGPAKLLWKGEGAVVIQDNSDI--------
1vie    ---------------DRVRKKSGaa-W----------QGQI--------
 1aboA   ----GYNhngEWCEAQTKNGQGWV------PSNYItp---------VN
 1ycsB   ----------EWWWARLNDKEGYV------PRNLLgLYP---------
 1pht    gwlnGYN--------ETTGERGDF------PGTYV-EYigRKKIsp--
 1ihvA   ----------Kv-----------V------PRr-------KAKIIRd-
 1vie    ----------VGWYCTNLTPEGYAveseahPGSVQ-IYPv-AALERIN
```

Exemple : 5 protéines, domaine HLH

domaine *helix - loop - helix*

Séquences à aligner:	longueur:
1) HEN1-Human	133
2) CBF1-Yeast	351
3) HES5-Mouse	167
4) IN04-Yeast	151
5) ESC1-Yeast	413

- longueurs dissemblables
- similarité locale



helix-loop-helix, alignement Clustal

```
-----------------------------------------------MMLNSDTMELD---------LPPTHSETESG------FSDCGGG
--MNSLANNNKLSTEDEEIHSARKRGYNEEQNYSEARKKQRDQGLLSQESNDGNIDSALLSEGATLKGTQSQYESG------LTSNKDE
MSSYALPSMQPTPTSSIPLRQMSQPTTSAPSNSASSTPYSPQQVPLTHNSYPLSTPSSFQHGQTRLPPINCLAEPFNRPQPWHSNSAAP
-----------------------------------------------MAPSTVAVEMLSPKEKNRLRKPVVEKMRRDR------INSSIEQ
-----------------------------------------------MTNDIKEIQTIQPGLSEIKEIKGELANVKKR-------------

AGPD-------------------------------------------------------------------------GAGPGG-------------
KGSDDEDASVAEAAVAATVNYTDLIQGQE---------DSSDAHTSNQTNANGEHKDSLNGERAITPSNEGVKPNTSLEGMTSSPMEST
ASSSPTSATLSTAAHPVHTNAAQVAGSSSSYVYSVPPTNSTTSQASAKHSAVPHRSSQFQSTTLTPSTTDSSSTDVSSSDSVSTSASSS
LKLL---------------------------------------------------------------------------------------
---------------------------------------------------------------------------------------------

---------PGGGQARGPEPGEPGRKD------------LQHLSREERRRRRRAT---------AKYRTA-------------------
QQSKNDMLIPLAEHDRGPEHQQDDEDNDDADID-----LKKDISMQPGRRGRKPTTLATTDEWKKQRKDS-------------------
NASNTVSVTSPASSSATPLPNQPSQQQFLVSKNDAFTTFVHSVHNTPMQQSMYVPQQQTSHSSGASYQNESANPPVQSPMQYSYSQGQP
---------LEQEFARHQPNSKLEKAD------------ILEMAVSYLKHSKAFAAAAGPKSLHQDYSEG-------------------
-----------KRRSKKINKLTDGQIR------------INHVSSEKKRRELERAIFDELVAVVPDLQPQ-------------------

----------------------HATRERIRVEAFNLAFA--ELRKLLPTLPP-----------------------------DKKLSKIEILR
----------------------HKEVERRRRENINTAIN--VLSDLLPVRESSKAAILARAAEYIQKLKETDEANIEKWTLQKLLSEQNASQ
FSYPQHKNQSFSASPIDPSMSYVYRAPESFSSINANVPYGRNEYLRRVTSLVPNQPEYTGPYTRNPELRTSHKLAERKRRKEIKELFDDLKDA
----------------------YSWCLQEAVQFLTLHAASDTQMKLLYHFQRPP-------------------------APAAPAKEPPA
--------------------------ESRSELIIYLKSLSYLSWLYERNEKLR-------------------------KQIIAKHEAKT

LAIC------------YISYLNHVLDV-----------------------------
LASANEKLQEELGNAYKEIEYMKRVLRKEGIEYEDMHTHKKQENERKSTRSDNPHEA
LPLDKSTKSSKWGLLTRAIQYIEQLKSEQVALEAYVKSLEENMQSNKEVTKGT----
PGAAPQPARSSAKAAAAAVSTSRQPACGLWRPW----------------------
GSSSSSDPVQEQNGNIRDLVPKELIWELGDGQSGQ----------------------
```

helix-loop-helix, alignement Dialign2 (1/2)

```
mml-------------------------------------------------------------------------------------------
m-----------------------------------------------------------------------------------------NSLANNNKLS
MA------------------------------------------------------------------------------------------
MT------------------------------------------------------------------------------------------
mssyalpsmqptptssiplrqmsqpttsapsnsasstpyspqqvplthnsyplstpssfqhgqtrlppinclaepfnrpqpwhsnsaapaSSSPTSATLS

---------------NSDTMELD----------------------------LPPTHSETESGFSDCGGGAGPDgagpggpgggqarg------------
TEDEEIHSARKRGYNEEQNYsearkkqrdqgllsqesndgnidsallsegatLKGTQSQYESGLTSNKDEKGSDdedasvaeaavaatvnytdliqgQED
------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------
TAAHPVHTNAAQVAGSSSSYVYS-------------------------VPPTNSTTSQAsakhsavphrssqfqsttltpstt----------DSS


-----------------------------------PEPGEPGRK-------------------------------------------------
SSDAHTSNQTNANGEHKDSLNGERAITPSNEGVKP-------------------------NTSLEGMTSSPMESTQQSKNdmliplaehdrg-------
------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------
STDVSSSDSVSTSASSSNASNTVSVTSPASSSATPLPNQPSQQqflvskndafttfvhsvhNTPMQQSMYVPQQQTSHSSGasyqnesanppvqspmqys
```

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

```
------------------------------------------------------------DLQHL---SREERRRRRATA-------------K
---------PEHQqddednddadidlkkdismqpgrrgrkPTTLAttdew-KKQR-----------------------------------------
-------------------------------------PSTVAVEMLSPKEKN-------------------------------------------
-------------------------------------NDIKEIQTIQPGLSEIKEIKGELANVKKR---KRRSKKINKLTDG-------------Q
ysqgqpfsyPQHK-------------------------NQSFSASPIDPSMSYVYRAPESFSSINANvpyGRNEYLRRvtslvpnqpeytgpytrnpE

YRTAHATRERIRVEAFNLAFAELRKLLPTL----PPDKKLSKIEILRLAICYISYLNHVldv-------------------------------------
-KDSHKEVERRRRENINTAINVLSDLLP-V----RESSKAA---ILARAAEYIQKLKETDEanieKWTLQKLLSEQNASQLASANEKLQEELGNaykeie
-RLRKPVVEKMRRDRINSSIEQLKLLLeqefarhQPNSKLEKADILEMAVSYLKHSKAFAA----Aag-----------------------------P
IRINHVSSEKKRRELERAIFDELVAVVPDL----QPQESRSELIIYLKSLSYLSWLYERNE----KLRKQIIAKHEAKTGSSSSSDPVQEQNGNirdlvP
LRTSHKLAERKRRKEIKELFDDLKDALP-L----DKSTKSSKWGLLTRAIQYIEQLKSEQV----ALEAYVKSLEEnmqsnkevtkgt-----------

-------------------------------------------------------------------------------------------
ymkrvlr--------KEGIEYEDMHThkkqenerkstrsdnphea-----------------------------------------------
KSLHQDYSEGYSwclQEAVQFLTLHAasdtqmkllyhfqrppapaapakeppapgaapqparssakaaaaavstsrqpacglwrpw
KELIWELGDGQSgq-------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
```

Résumé des méthodes

Méthode	Idée	Stratégie
MSA DCA	Extension de l'algorithme de Needlemen et Wunsh	Simultanée
Clustal PIMA PILEUP MULTALIGN Dialign	Ajout successif de séquences ou groupes de séquences	Progressive
Saga/Coffee PRRN HMMT MUSCLE MA-FFT	Réalignment lors de l'ajout successif de séquences ou groupes de séquences	Itérative

Passage à l'échelle

la comparaison de génomes

1. taille de séquences bien plus longue (1000bp $\rightarrow >$ 1000000bp)
2. présence de réarrangements/duplications (combinées)

La comparaison de génomes

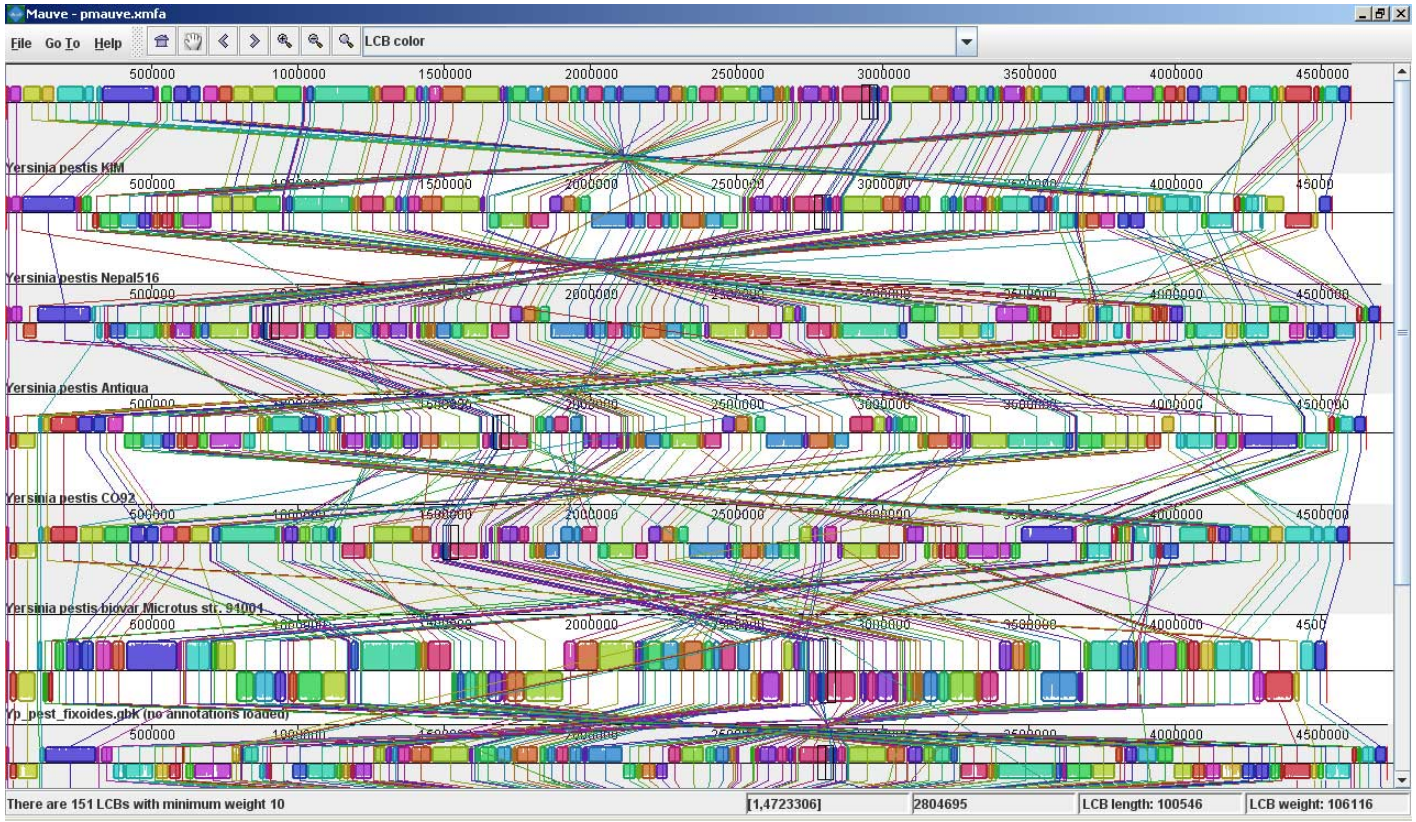
$\Rightarrow$ autant de programmes qui produisent des alignements différents !

Différentes écoles :

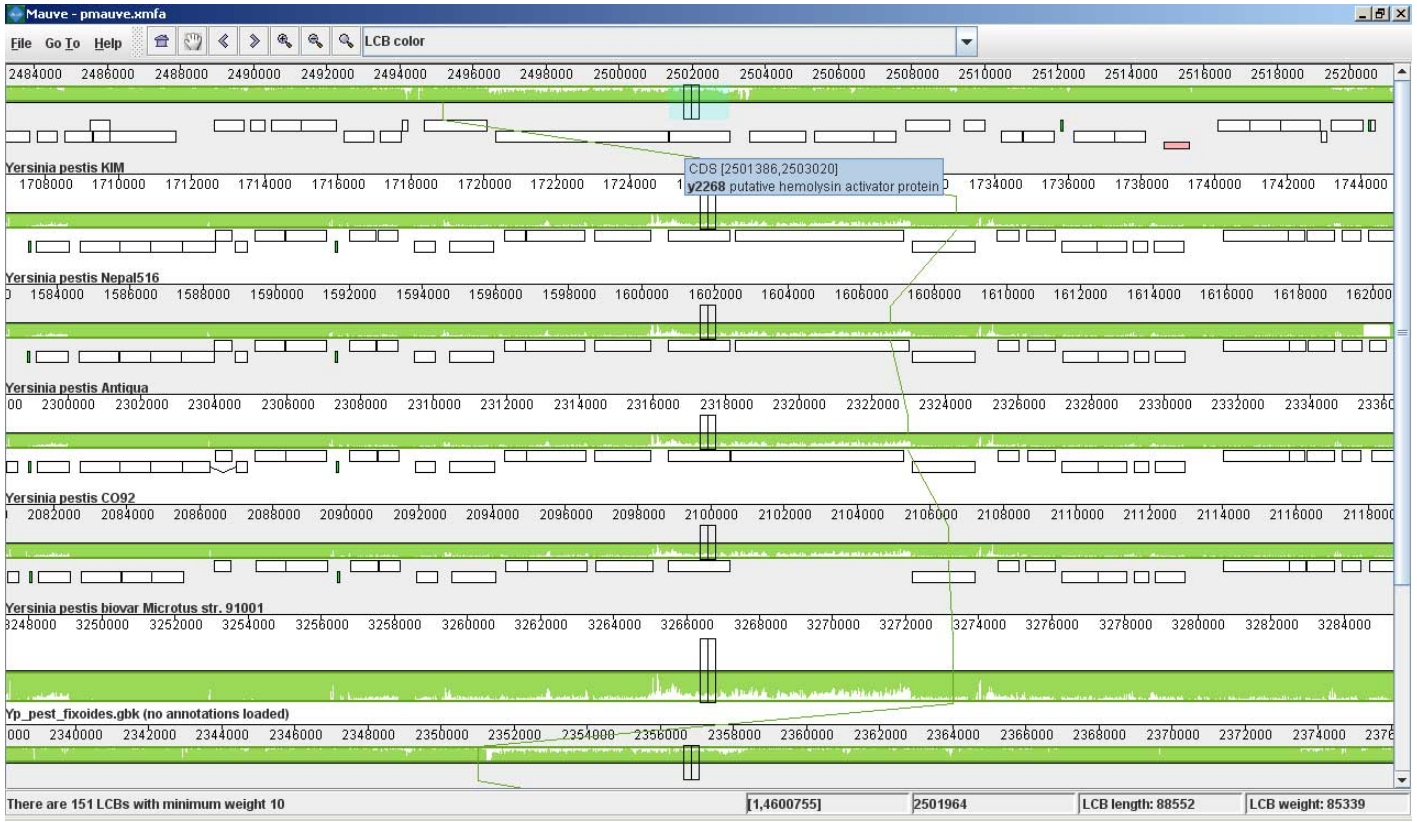
- MGA (Bielefeld)
- MUMmer (Baltimore/Celera genomics)
- Lagan, Multilagan (Standford)
- MAUVE (Wisconsin-Madison)
- GLASS AVID (Berkeley)
- et bien d'autres ...

La comparaison de génomes - exemple de Mauve

La comparaison de génomes - exemple de Mauve



La comparaison de génomes - exemple de Mauve



La comparaison de génomes - exemple de Mauve