

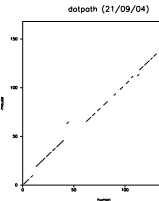
Comparaison de séquences deux à deux

Équipe Bonsai
<http://www.lifl.fr/bonsai>
 année 2013

Exemple 1 - heat shock protein beta 9

```
>human
MQRVGNTPSN ESRVASRCPS VGLAERNVA TMPVLLRDS FFAQEDMDHA RDGFQMKLDA
RQFAPFELWV QVDSQWLVMT CQQLDVRDP ERVYVRSQK VRRMLPSNL SPTAMTCLT
PSQLWVRGQ CVALALPEAQ TQSPRLQSL GSKASNLTR
>mouse
MQRVGSFST QREPCENRV ASRCPSVALA ERNQTALPV RLLRDEYQGN GCEQPSFOIK
VDAQFPED LVVRIDQNL TVTQGRQES NDFSRGRYM EQSVHRQML PPTLDPAAAT
CSLTPSGHLW LRGNKCLPP PEAGTQQSQK PRRGGPKSSL QNESVKNP
```

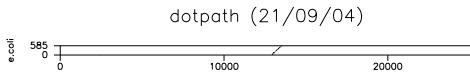
```
human 1 MQRVGNTPS-----NESRVASRCPSVGLAERNVATMPVLLRDSPPAAQ
      | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
mouse 1 MQRVGSFSTQREPCENRVASRCPSVALAERNQVATLVPVLLRDE---V
      | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
human 45 EKDHDHARDGFQHLDAHGFAPFELVYQVDGQLMVTQQLDVRDPERYVS
      | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
mouse 48 QHNGCEQPSFOIKVDAQFPEDLVVRIDQNLTVTQGRQESNDPSRGR
      | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
human 95 YMSQKVRHKM-LPSNL SPTAMTCLTQSPQLVVRGQCVLALPEAQTC
      | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
mouse 98 YRMEQSVHRQMLPPTLDPAAATCSLTPSGHLWLRGNKCLPPPEAQTCG
      | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
human 144 S--PRLQSLGSKASNLTR----- 159
      | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
mouse 148 SQQPRAG--GPKSSLQNESVKNP 168
```



⇒ similarité globale, alignement **global**

Exemple 2 - E.coli peptidyl tRNA hydrolase

AE008779.1 : *Salmonella typhimurium* LT2, section 83 of 220 of the complete genome. 25184 bp
 e.coli : *Escherichia coli* peptidyl tRNA hydrolase. 585 bp

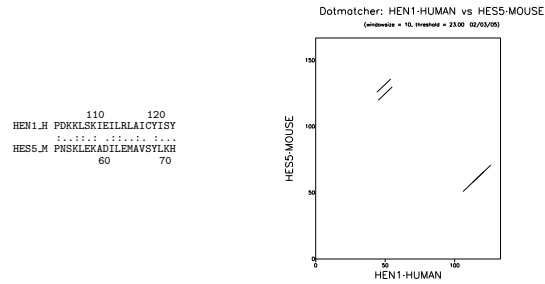


```
AE008779.1 12901 TCAGGCAAAAAAGCTGGCAATTAAATTTGTTGGTGTGGGAAATCC 15201 TCGATCTCTCTGCGGCGGGCTGGCGAAATTAACCTGCGCGGGGGCCAGCGC
      |||..|||||..|||||..|||||..||..|||||..||..|||||..||
e.coli 1 gtgagattaaattgattgtcgcttgcgacacccc 287 tggatctcctctccgtgcgccaaattaaattggcggctgacctggt
AE008779.1 12951 GGTGGGAAATATGCGGACGCGACCAATGCAAGCGCATGTCAGTCGAA 13251 GGGCAGCATGGTCTGAAAGACATCACTACGACGTCGGCATAATCCCAA
      |||..|||||..|||||..|||||..|||||..||..|||||..||..|||||..||
e.coli 37 gggcgataatgacgcaagcgcaatgctggtggctgagttgtga 337 ggtcaaatggctggaagcctcactcagtaabaattgggtataaccctaa
AE008779.1 13001 TTACTGGGAGGCGGTGCGGGCGCGGCTTGGTGAAGAGCAATTC 13301 CTTTCAGGATTACGGCTTGCAATTTGCTACGCAAGGCAATAAAAAAAG
      |||..|||||..|||||..|||||..|||||..||..|||||..||..|||||..||
e.coli 87 cttaactggcagagcttggcgtcccgctgpcgaaggcctaactct 387 cttcacgcttagcgtcggactcgtctccggcgataaanaaaag
AE008779.1 13051 TTGGCTATACTGCAAGCACTGCAAGCAAGCAAGCAAGCAAGCA 13351 TTGTTGGTTTTGCTGGTAAACCCCGCTTTCTTGACAAAAATAAT
      |||..|||||..|||||..|||||..|||||..||..|||||..||..|||||..||
e.coli 137 gttgtataacttgcgagtcactcttggagcgaagtgcggctgtga 437 ttgctggtttggtttagcgaacccctgttagtgaacagaagttaatt
AE008779.1 13101 GTACCCACAGCTTCAATAAAGCAAGTCAAGTCAAGTCAAGTCAAG 13401 GATGAGGCAATGACAAAGGCGCAGCTGTAAGCAAAATTTGGTCAAGA
      |||..|||||..|||||..|||||..|||||..||..|||||..||..|||||..||
e.coli 187 gtccgacacattatgaaattcagagcgaagcccttggcgtgagc 487 gttgaaagccttagcgaagctgtagtgaatgtagctagcaag
AE008779.1 13151 CAGTTTTAAGCATTAGCGCGGACGAAATTTTGGTGTGCAAGCAAGCA 13451 GGGTCTGGGCAAGCAAGCAAGCGGTTGGCATAAGCTTTAAGCGCATAAAG
      |||..|||||..|||||..|||||..|||||..||..|||||..||..|||||..||
e.coli 237 cagtttttccgacttaaccggcgaattctgtggcaccagcaaac 537 tggcttgaccaaaagcaagcaagcattgacagcctttaaacgcataaa
```

⇒ similarité globale sur une partie, alignement **semi-global**

Exemple 3 : domaine conservé

```
>HEN1_HUMAN
HMLNSDTPHELDPFTSESETSGFSDCGGAGPDGAGGGGQGGQARPQEPGRKDLHLREERRRRR
RATAKYTAHTATKIRIRVAFNLAFLAEKRLKPLTPDKKLKILELRLAICYSTLHVLDV
>HESS_MOUSE
MAPSTVAVEMLSPKKNLRLKPVVEKMRADRINSSEIQKLLLEQEFARHQVNSKLEKADILEMAVSYLKH
SRAFAAAAGPKSLHQDTSEYFSWCLQEAQVQLTLHAASDTQHKLLYHFQPPAPAFAPKEPPAPGAAPQ
PARSSAKAAAAYSTSRQFAQLRQV
```



⇒ similarité sur une sous-partie, alignement **local**

Pourquoi comparer des séquences ?

quelques problèmes

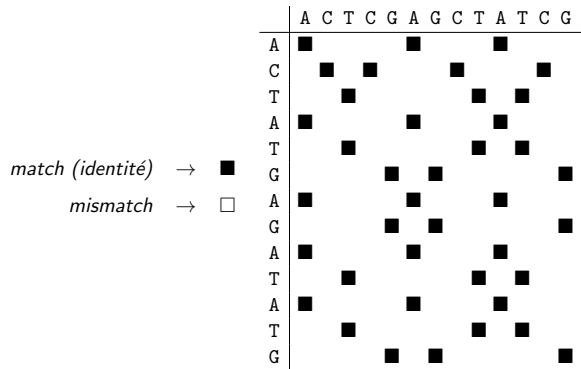
- rassembler un ensemble de fragments d'ADN séquencés (fragment assembly)
- recherche d'homologie (entre gènes)
- trouver des régions similaires (domaines protéiques)
- identifier les positions introns/exons (comparaison d'un gène à son ARNm)

Dotplot

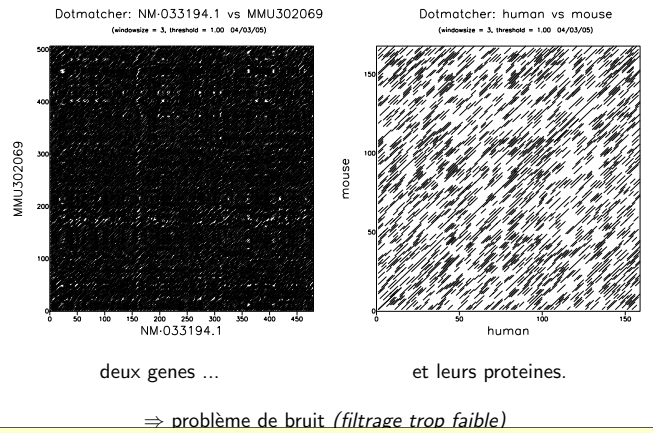
- un outil graphique pour la comparaison
- principe
 - mettre les séquences le long des axes d'une matrice
 - mettre un point là où il y a un match

une diagonale (une suite de points en diagonale) ⇒ une région similaire

Dotplot - exemple



Dotplot

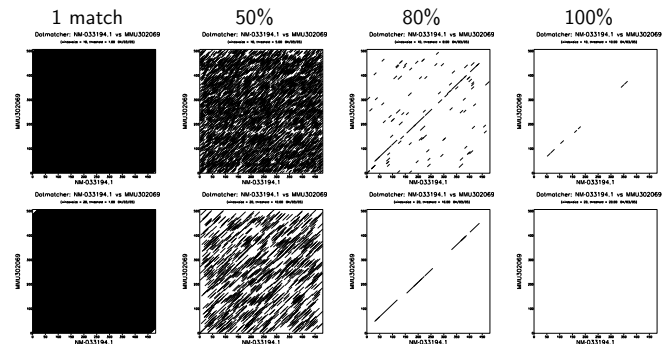


Dotplot - Filtrage

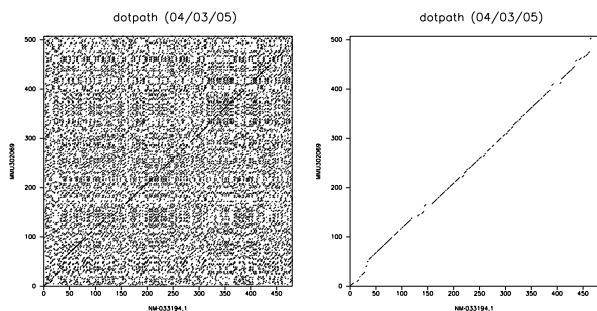
- blocs d'identité
 - idée : ne représenter que des fenêtres exactes et de longueur fixe
 - exemple de programme : **dotup**
 - très sélectif et peu sensible
- blocs de similarité avec un seuil de score (Maizel & Lenk - 1981)
 - idée : ne représenter que des fenêtres possédant un score supérieur à un seuil
 - exemple de programme : **dotmatcher**
 - bonne sélectivité et bonne sensibilité
- filtrer les blocs pour éliminer ceux qui se chevauchent
 - idée : observez la ressemblance globale
 - exemple de programme : **dotpath**
- accorder un poids physico-chimique aux matches (Staden - 1982)
 - variation de l'intensité, ne pas se contenter de ■ et □
 - prise en compte des propriétés des acides aminés

Dotplot : taille des mots et seuil

taille 10 et 20, seuil de 1% à 100%

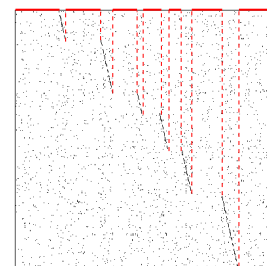


Dotplot : dotpath



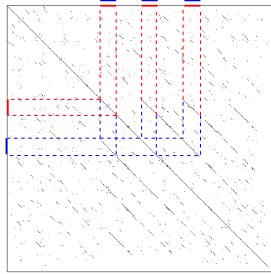
dotpath finds all matches of size wordsize or greater between two sequences. It then reduces the matches found to the minimal set of long matches that do not overlap. This is a way of finding the (nearly) optimal path aligning two sequences. It is not the true optimal path as produced by the algorithms used in water or needle, but for very closely related sequences it will produce the same result and will work well with very long sequences

Dotplot : repérer des similarités locales



horizontalement : séquence nucléaire du gène de l'actine de muscle de *Pisaster ochraceus*
 verticalement : cDNA de ce même gène

Dotplot : repérer des répétitions



protéine ribosomale S1 de *Escherichia Coli* sur elle-même

Dotplot : son information

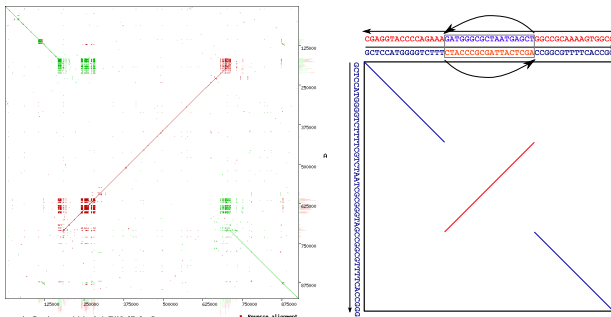
Sur une seule séquence et son **complémentaire inversé**



- 1 répétition en tandem (micro/mini-satellite)
- 2 palindrome

Dotplot : inv. centrées sur origine de réplication

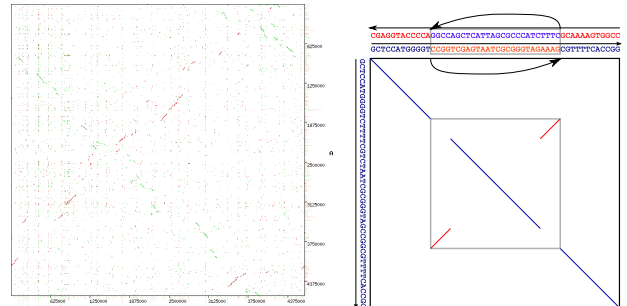
Entre deux séquences (même espèce ou espèces proches) :



inversion d'un segment (≠ palindrome)
(génomés de *Tropheryma whipplei*)

Dotplot : inv. centrées sur origine de réplication

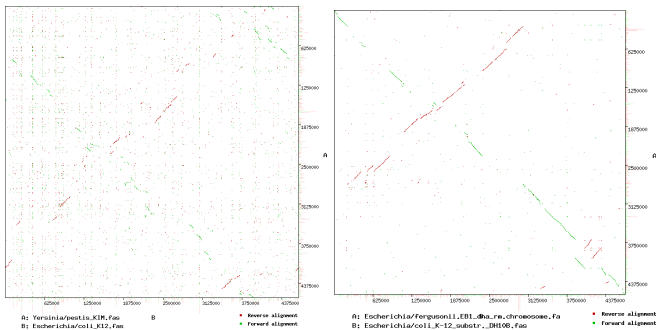
Entre deux séquences (même espèce ou espèces proches) :



(génomés de *Yersinia pestis* / *Escherichia coli*)

Dotplot : inv. centrées sur origine de réplication

Entre deux séquences (même espèce ou espèces proches) :



(génomés de *Yersinia pestis* / *Escherichia coli*)

(génomés de *Escherichia coli* / *Escherichia fergusonii*)

Dotplot : avantages et inconvénients

- les plus :
 - simple
 - très informatif
 - les moins :
 - interprétation ⇒ pas de mesure objective
 - identification ⇒ pas de méthode de détection automatique
- ⇒ besoin d'une mesure quantitative de similarité

Alignement

3 types d'alignement

- **global** \mapsto match sur deux séquences complètes.
- **local** \mapsto match sur des sous-séquences
- **semi-global** \mapsto chevauchements

Alignement

données :

- une paire de séquences (ADN / protéine)
- un schéma de score : comment compter ce qui se ressemble ?

but :

- déterminer le degré de similarité (meilleur score)
- montrer la similarité (meilleur alignement)

décrit la ressemblance grâce à 3 opérations (mutations ponctuelles)

- insertion
- délétion
- identité/substitution

mesure la ressemblance en donnant un poids à chaque opération

- poids positif ("récompense") aux *bonnes parties* de l'alignement
e.g appariement de deux lettres identiques ou proches
- poids négatif ou nul ("pénalité") aux *mauvaises parties* de l'alignement
e.g appariement de deux lettres non relatés, non-appariement

Composantes d'un schéma de scores

score (ou poids) pour une identité/substitution : matrice s de similarité

- exemple : $s(a, b)$ = score d'alignement des nucléotides a et b

$$\begin{pmatrix} +1 & -1 & -1 & -1 \\ -1 & +1 & -1 & -1 \\ -1 & -1 & +1 & -1 \\ -1 & -1 & -1 & +1 \end{pmatrix}$$

score (ou poids) d'un *indel* (insertion/délétion)

- exemple : score unitaire = -2 par *indel*

score de l'alignement = somme des scores des événements élémentaires

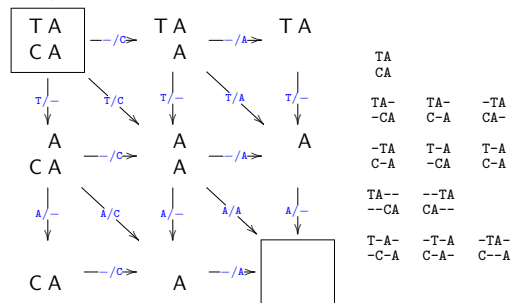
- exemple :

A	A	C	G	T	A	C	G	A	T	A
A	A	C	G	T	A	-	A	A	G	A
1	1	1	1	1	1	-2	-1	1	-1	1 = 4

Comment calculer le meilleur alignement ?

prenons 2 séquences de longueur n toutes les deux : alignement de longueur maximale $2n$

exemple avec les séquences TA et CA



- une suite de couples de lettres en suivant les flèches donne un alignement

Difficulté du problème

nombre max d'alignements (s'éq. de lg n)

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \sim \frac{2^{2n}}{\sqrt{2\pi n}}$$

pour deux séquences de longueur 100 : 2.10^{57} alignements

grâce à la représentation en tableau : complexité en temps et en espace $\mathcal{O}(n^2)$

(proportionnel au produit de la longueur des séquences)

pour deux séquences de longueur 100 : 10000 opérations

Alignement global

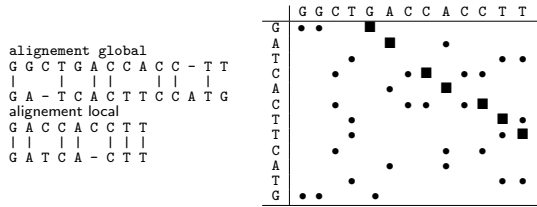
Needleman & Wunsch, 1970

- évaluation d'une ressemblance **globale** entre deux séquences
- recherche du meilleur alignement global.

Alignement local

Smith & Waterman, 1981

- évaluation d'une ressemblance **locale** entre deux séquences
- recherche de la région de plus forte similarité.



les séquences présentent une similarité que l'alignement global ne révèle pas

Alignement semi-global

```
CAGCACTGGATTCTCGG
| | | | | | | | | |
CAGC-----G-T----GG

CAGCA-CTGGATTCTCGG
| | | | | | | |
---CAGCGTGG-----
```

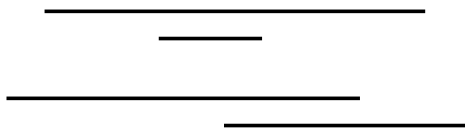
⇒ l'alignement global préfère le 1er alignement

⇒ l'alignement semi-global préfère le 2eme alignement

Alignement semi-global

Principe

- ne pénalise pas les *gaps* (séries d'indels) aux extrémités
- sinon, similaire à l'alignement global
- permet de détecter des similarités de ce type :



Pénalité associée aux *gaps*

- fonctions de *gaps* différentes ⇒ algorithmes différents
- la plus simple ⇒ fonction linéaire : $g \times l$
(l : longueur du *gap*)
- fonctions plus réalistes :
 - fonctions affines : $o + e \times l$
 - o : pénalité d'ouverture de *gap*
 - e : pénalité d'extension de *gap*
 - fonctions logarithmiques

Influence du jeu de scores sur l'alignement

	jeu de score			
	A	B	C	D
Match Cost	1	1	1	1
Mismatch Cost	-1	-1	-1	-1
Gap Open Penalty (o)	0	-1	-1	-1
Gap Extension Penalty (e)	0	-1	0	-0.1

- AL1 : ATGCGGGACATG
| | | | | | | | | |
AGGCG--CC-TG (7 matches, 2 mismatches, 1 gap of 1 bp, 1 gap of 2 bp)
- AL2 : ATGCGGGACATG
| | | | | | | | | |
AGGCGC--C-TG (7 matches, 2 mismatches, 1 gap of 1 bp, 1 gap of 2 bp)
- AL3 : ATGCGGGACATG
| | | | | | | | | |
AGGCG---CCTG (7 matches, 2 mismatches, 1 gap of 3 bp)
- AL4 : ATGCGGGACATG
| | | | | | | | | |
AGGCG-C-C-TG (7 matches, 2 mismatches, 3 gaps of 1 bp)

Comment bien choisir ?

- peu de connaissance a priori
- spécificité aux données
- valeurs typiques pour une fonction de *gap* affine
 - $0.5 < o < 5.0$
 - $0.05 < e < 1.0$
- toujours prendre (en valeur absolue) $o > \frac{1}{2} substitution$

```
ACCTG-A-CGTA-AGC      ACCTGACGTAAGC
| | | | | | | | | |
ACCT-C-TCGT-TAGC      ACCTCTCGTTAGC
```
- poids "naturels" (Thorne, Kishino & Felsenstein, 1991)
estimation de la vraisemblance que les deux séquences proviennent du même ancêtre

Evaluer la qualité d'un alignement ?

```

A C C T G A C G T A A G C       A C C A G T G C A G T - - T C
| | | | | | | | | | | | | | | |
A C C T G A C G T A A G C       A C C - - T G A C G T A A G C

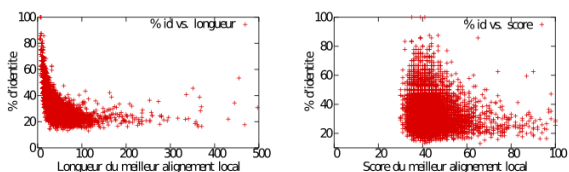
A C T T G A C G T - A G C       - - C T A C C T C G A C T - C A G C
| | | | | | | | | | | | | | | |
A C C T G A C G T A A G C       A C C T G A - - C G T A A G C - - -
    
```

Quelques principes informels

- **robustesse aux paramètres choisis pour le calcul du score**
On peut douter d'un alignement si de faibles changements (environ 10%) dans l'établissement des pénalités d'insertion-délétion modifient sensiblement cet alignement.
- **fréquence des indels**
On peut douter d'un alignement s'il nécessite plus d'une insertion en moyenne pour 20 acides aminés.
- **deux séquences nucléiques d'au moins 100 bases et identiques à 50% n'ont pas forcément de relation biologique.**
- **des séquences protéiques de 100 résidus ou plus, possédant au moins 25% d'identité entre elles ont certainement un ancêtre commun (Doolittle, 1990 - PDB).**

Le pourcentage d'identité

- dépend de la composition en bases, ou acides aminés
- dépend de la longueur des séquences



Une fausse bonne idée

Approche empirique

- test de la robustesse du score
- S : score de l'alignement entre U et V
- méthode
 1. Génération de 100 (200, 1000, ...) permutations de V (même longueur, même composition)
 2. Alignements avec U
 3. Distribution des scores d'alignement

Où se situe S dans cette distribution ?

Exemple

Alignement local pour ACCAGTGCAGTC et ACCTGACGTAAGC

```

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

score du meilleur alignement local : 16

```

score  s-w
0  0 :
4  138 :=====
8  166 :=====
12 146 :=====
16  33 :=====
20  7 :==
24  8 :==
28  2 :
32  0 :
36  0 :
40  0 :
44  0 :
48  0 :
    
```

500 séquences aléatoires

Exemple

```

Human alpha haemoglobin (141 aa) vs. Human myoglobin (153 aa)
V L S P A D K T N V K A A W G K V G A H A G E Y G A E A L E R M F L S F P T T K Y F P H P - D L S - - - - H G S A Q V K H G K K V A D A L T N A V A H V D D M P N A L S A L
: : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
G L S D S D E Q L V L W W K V E A D I P G H Q Q E V L I L F G H P E T L E K F O K F K H L K S E D E M K A S E D L A K H A T V L T A L G G L I K K G H E A E I K P L
: : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
S D L H A H K L R V D P V N F N K L L S H C L L V T L A A H L P A E F T P A V H A S L D K F L A S V T V L T S K Y R - - - - -
: : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
A Q S H A T R K H I P V K Y L E F I S E C I I Q V L Q S K H P D F G A D A Q A M N K A L E L F R K D M A S N Y K E L G F Q G
: : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
    
```

```

Chicken lysozyme (129 aa) vs. Bovine ribonuclease (124 aa)
K V F G R C E L A A A M K R G L D N Y R G Y S L G W V C A A R F E S N F W T Q A T N R N T D G S T D Y G I L Q I N S R W W C N D G R T P - - G S R N L C N I P C S A L L S S D
: : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
K E T A - - - - A A F E R Q H M D S T S A A S S S N Y C N Q M M K S R N L T K D R C K P V N T F V H E S L A D V Q A V - - C S Q K N V A C K N G Q T N C Y Q S Y S T M S I T D
: : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
I T A S V N C A K K I V S D G D M H A V A W N R C K G T D V Q A V I R C G L
: : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
C R E T - G S S K Y P N C A Y K T T Q A N K H I I V A C E G N P Y V P V H F D A S V
: : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
    
```

PRSS sur les globin

PRSS : Probability of Random Shuffle Sequence

```
< 20 0 0:
22 0 0:      one = represents 1 library sequences
24 0 0:
26 0 0:
28 0 0:
30 1 1:*
32 3 3:***
34 9 7:*****
36 25 15:*****
38 37 25:*****
40 29 34:*****
42 33 42:*****
44 51 46:*****
46 41 47:*****
48 32 45:*****
50 51 41:*****
52 31 36:*****
54 24 31:*****
56 30 26:*****
58 18 21:*****
60 24 17:*****
62 19 14:*****
64 4 11:*****
66 10 9:*****
68 5 7:*****
70 4 5:*****
72 7 4:*****
74 3 3:***
76 2 3:***
78 3 2:***
80 1 2:***
82 0 1:*
84 1 1:*
86 0 1:*
90 0 0:
92 1 0:      unshuffled s-w score: 177
94 0 0:      For 500 sequences, a score >= 177 is expected 3.096e-06 times
```

PRSS poulet/bovin

PRSS : Probability of Random Shuffle Sequence

```
< 20 0 0:
22 0 0:      one = represents 1 library sequences
24 0 0:
26 0 0:
28 0 0:
30 0 1:*
32 3 3:***
34 9 7:*****
36 17 15:*****
38 24 25:*****
40 35 36:*****
42 44 42:*****
44 57 46:*****
46 50 47:*****
48 44 45:*****
50 45 41:*****
52 25 36:*****
54 28 31:*****
56 20 26:*****
58 24 21:*****
60 17 17:*****
62 10 14:*****
64 8 11:*****
66 10 9:*****
68 8 7:*****
70 6 5:*****
72 6 4:*****
74 2 3:***
76 2 3:***
78 2 2:***
80 5 2:***
82 1 1:*
84 0 1:*
86 2 1:***
88 0 1:*
90 0 0:
92 0 0:      unshuffled s-w score: 43
94 1 0:      For 500 sequences, a score >= 43 is expected 267.1 times
96 0 0:
```

Approche statistique

Definition (E-value)

Nombre de fois attendu de trouver un alignement de score supérieur à S par hasard quand on aligne une séquence de longueur n avec une séquence de longueur m .

- décrit le bruit aléatoire qui existe lorsque on aligne des séquences
- croît de manière proportionnelle en fonction de n et de m
- décroît de manière exponentielle en fonction du score S
- **plus la E-value est proche de 0, plus la similarité est significative**