M1 Bioinformatique, Connaissances et Données
Master Sciences et Numérique pour la Santé
Année 2017-2018

HMSN206 - Partie Alignement

Partie I-2 : Matrices de scores, BLAST, FASTA

Anne-Muriel Chifolleau

http://www.lirmm.fr/∼arigon/enseignement/HMSN206/

LIRMM - UM

Similarité et homologie (1/3) 1

- Tendance à les utiliser de manière interchangeable alors qu'ils ont des sens différents et impliquent des relations biologiques différentes
- Similarité
  - → Mesure quantitative de la ressemblance entre 2 séquences
  - → Mesure toujours basée sur de l'observable, en général sur l'alignement de ces 2 séquences
  - → Pourcentage d'identité : pourcentage de résidus identiques
  - → Pourcentage de similarité : pourcentage de résidus relativement proches sans forcément être identiques (utilisation de matrice de substitution)

⇒ Haut degrés de similarité *peut impliquer* une origine évolutive commune

Similarité et homologie (2/3) 2

- Tendance à les utiliser de manière interchangeable alors qu'ils ont des sens différents et impliquent des relations biologiques différentes
- Homologie
  - → Conclusion supposée basée sur l'examen d'un alignement optimal entre 2 séquences et de leur similarité
    (ex pour les protéines : 30 % d'identité sur une longueur de 100 a.a. peut impliquer une homologie)
  - → Les gènes (ou les protéines corr.) sont ou ne sont pas homologues, l'homologie ne se mesure pas en degrés
  - → Le concept d'homologie implique des relations évolutives (ancêtre commun) et peut s'employer pour 2 types de rel. : orthologie et paralogie

⇒ Voir la partie phylogénie

Similarité et homologie (3/3) 3

Mais ...

- on peut avoir similarité sans homologie
  - → Convergence ou simple hasard pour de courtes séquences
  - → Existence de régions de faible complexité (régions riches en quelques a.a.)
- on peut avoir homologie sans similarité
  - → Similarité faible entre des protéines qui ont des données fonctionnelles et biochimiques qui montrent qu'elles sont homologues.

Alignement global contre alignement local

- Ce sont les deux grands types d'alignements
- Ils permettent d'estimer la similarité entre séquences

→ GLOBAL : 2 séquences, comparaison sur leur totalité, trouve le meilleur alignement sur toute leur longueur
Utilisé pour des séq. très similaires et approx. de même lg
[Needleman & Wunsh, 1970]

→ LOCAL : chercher les régions les plus similaires dans les 2 séq.
Permet de trouver des sous-séq. ayant des relations biologiques
[Smith & Waterman, 1981]

Les alignement locaux sont meilleurs pour des séquences avec un faible degrés de similarité ou de tailles différentes

Plan

- Donner un score aux alignements
  - → Les matrices de scores, généralités
  - → Les matrices protéiques
  - → Les matrices nucléiques
  - → Pénalités de gaps
- Les heuristiques
- BLAST
- FASTA
- Comparaison BLAST/FASTA
- La suite ...

Les matrices de scores ou substitution

- Utilisées, notamment, dans la comparaison de deux ou plusieurs séquences (alignement) : fonction de scores construite en utilisant une matrice de score
- Pour chaque élément, une matrice de scores définit le score élémentaire à affecter
  - → à la conservation d'un élément
  - → au remplacement d'un élément par un élément
- De nombreuses matrices de scores : unitaire, code génétique, propriétés physico-chimiques, empiriques déduites de l'évolution, ...
- Comment sont-elles construites ? Comment les choisir ?

Choix important influençant fortement les résultats

[Henikoff & Henikoff, 2000]

Facteurs biologiques importants

- Conservation
  Conservation absolue et substitutions conservatives (identité / similarité)
  → *Quels résidus peuvent être substitués par d'autres sans affecter la fonction de la protéine ?*
- Fréquence
  Les résidus rares ont un poids plus fort que les résidus communs
  → *Les matrices doivent refléter le nombre de fois qu'un résidu apparaît dans l'ensemble des protéines*
- Évolution
  Les matrices représentent implicitement les schémas évolutifs
  → *Le choix des matrices à utiliser dépend de la distance évolutive*

⇒ De nombreuses matrices nucléiques / protéiques existent

Les matrices protéiques déduites de l'évolution 8

- Construites sur les substitutions entre acides aminés intervenues au cours de l'évolution

→ basées sur des substitutions observées dans des alignements de protéines homologues (familles de protéines)

→ utilisent les fréquences de substitutions observées et théoriques

- Exemples : BLOSUM [Henikoff et Henikoff, 1992], PAM [Dayhoff et al., 1978], JTT [Jones et al., 1992], WAG [Whelan et Goldman, 2001], ...

Les matrices protéiques déduites de l'évolution 9

- Score = logarithme d'un "rapport de chances" (*odds ratio*) : prend en compte le nb de fois où un résidu particulier

→ est remplacé par un autre dans la nature

→ serait remplacé par un autre si ça se produisait par chance

- Log odds ratio : $S_{i,j} = \lambda \log[\frac{q_{i,j}}{p_i p_j}]$
  λ coefficient de normalisation
  $q_{i,j}$ fréquence de substitution i par j (à partir d'alignements)
  p_i prob. d'occurence de i parmis toutes les protéines

- Le log odds ratio est le rapport de l'observé sur l'aléatoire

> Les substitutions observées fréquemment ont un score positif et celles peu fréquentes, un score négatif

Exemple : BLOSUM62 10

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Les matrices PAM 11

- Les premières matrices utiles pour l'analyse de séquences [Dayoff et al., 1978]
- Idée : comparer des séquences très proches permettant d'inférer des relations d'évolution des protéines au sein des familles (phylogénie) et extrapoler ces relations à des distances d'évolution plus grandes.
- Basées sur l'étude des motifs de substitution d'un groupe de protéines avec plus de 85% d'identité (71 fam. de prot. homologues)
  $\Rightarrow$ Table contenant la fréquence de substitution d'un a.a. par un autre à une position donnée
- L'unité de mesure résultant de cette analyse est le **P***oint* **A***ccepted* **M***utation* ou unité **PAM** (normalisation de la table)
- Une unité PAM correspond à 1 changement d'a.a. pour 100 résidus $\Rightarrow$ PAM1

Les matrices PAM : les hypothèses

- Le remplacement d'un a.a. est indépendant des mutations précédentes à la même position
  ⇒ la matrice originale a été extrapolée pour prédire les fréquences de substitution pour des distances évolutives plus élevées
  Ex : PAM1 multipliée par elle-même 100 fois donne PAM100
- PAMn : n représente le nombre de substitutions pour 100 résidus
  Ex : PAM100 = 100 substitutions pour 100 résidus
- Tous les sites sont mutables de manière équivalente
- Les remplacements sont indépendants des résidus voisins
- Pas de considération de bloc ou de motif
- Autres biais dûs par exemple aux nb et caractéristiques des protéines connues en 1978, ...

Les matrices BLOSUM

- Approche différente par [Henikoff & Henikoff, 1992]
- Identifier les motifs conservés à l'intérieur des familles de protéines (2000 domaines conservés de 500 fam. de prot. + ou - divergentes)
  ⇒ Création de la base de données BLOCKS

→ Motif : suite conservée d'a.a. qui confère une fonction ou une structure spécifique à une protéine
→ Blocs : généralisation des motifs

- À l'aide de ces blocs, ils ont regardé les schémas de substitution dans les régions les plus conservées des protéines

  ⇒ Génération des **BL***ocks* **SU***bstitution* **M***atrices* ou matrices **BLOSUM**

Les matrices BLOSUM (2)

- Plus de protéines disponibles en 1992 qu'en 1978, donc un jeu de données plus robuste pour dériver des matrices
- Les matrices BLOSUM sont directement calculées à différentes échelles évolutives, elles ne sont pas extrapolées
- BLOSUMn : n représente le niveau de conservation des séquences utilisées pour construire la matrice

  Ex : BLOSUM62 est calculée à partir de séq. qui partagent plus de 62 % d'identité
- Réduire la valeur de n permet d'analyser des séquences plus divergentes

Quelle matrice ? Quand ?

Matrices	Meilleur usage	Similarité (%)
PAM40	Alignements courts qui sont très similaires	70-90
PAM160	Détecter les membres d'une famille de protéines	50-60
PAM250	Ali. plus longs de séq. plus divergentes	~ 30
BLOSUM90	Alignements courts qui sont très similaires	70-90
BLOSUM80	Détecter les membres d'une famille de protéines	50-60
BLOSUM62	La matrice "tout-terrain"	30-40
BLOSUM30	Ali. plus longs de séq. plus divergentes	~ 30

- Rmq : Les numéros des 2 familles de matrices vont en sens opposé
- PAM250∼BLOSUM45 ; PAM160∼BLOSUM62 ; PAM120∼BLOSUM80
- Il existe d'autres matrices : JTT, WAG, LG, des matrices spécialisées (ex : espèces spécifiques, classes de protéines), ...

Quelle matrice ? Quand ? 16

- Le choix par défaut des logiciels n'est pas forcément approprié
- Les matrices de scores affectent les résultats d'un alignement et il est difficile de juger de la qualité d'un alignement de deux séquences.
- Le choix de la matrice dépend de la divergence qu'ont les deux séquences étudiées, les meilleurs résultats étant obtenus lorsque on utilise la matrice la plus sensible par rapport au niveau de divergence réel des séquences.

Les matrices de scores nucléiques 17

- 4 nucléotides au lieu de 20 a.a.
- Hypothèse simple de fréquence identique (25% chacun), avec éventuellement un poids différents pour les transitions (A $\leftrightarrow$ G, C $\leftrightarrow$ T), plus fréquentes que les transversions (A $\leftrightarrow$ C, G $\leftrightarrow$ T)

	A	C	G	T
A	1			
C	0	1		
G	0	0	1	
T	0	0	0	1

	A	C	G	T
A	3			
C	0	3		
G	1	0	3	
T	0	1	0	3

- Plus d'études sur les séquences protéiques que sur les séquences nucléiques $\Rightarrow$ matrices nucléiques moins perfectionnées

$\Rightarrow$ Chercher sur les séq. protéiques est toujours plus efficace

Nucléique ou protéique ? 18

- Alphabet réduit dans le cas des séquences nucléiques
  $\Rightarrow$ distinction entre le signal et le bruit plus difficile
- Redondance du code génétique
  $\Rightarrow$ une mutation au niveau de l'ADN ne se traduit pas forcément au niveau de la protéine
  $\Rightarrow$ similarité plus longtemps observable au niveau des séquences protéiques
- Séquences protéiques plus courtes
  $\Rightarrow$ comparaison plus rapide
- En phylogénie, possible manque de signal dans les séquences protéiques (utilisation de méthodes adaptées aux séquences nucléiques : MACSE, ...)

Gaps et pénalités de gap (rappel) 19

- Les gaps servent à améliorer les alignements
- Mais pour ne pas donner lieu à des scénarios biologiquement invraisemblables, il faut les garder en nombre raisonnable :
  $\rightarrow$ 1 indel pour 20 a.a.
- La méthode la plus utilisée pour pénaliser les gaps est la "pénalité de gap affine"

  $o + e \times (l - 1)$ où $o > e$

  o pénalité d'ouverture de gap
  e pénalité d'extension de gap
  l longueur du gap
- Autres types de pénalités possibles :
  $\rightarrow$ linéaires : $e \times l$
  $\rightarrow$ logarithmiques : $o + e \times log(l)$
  $\rightarrow$ pondération de e par la distance évolutive ou la nature des résidus
  $\rightarrow$ ...

Plan

- Donner un score aux alignements
- Les heuristiques
- BLAST
- FASTA
- Comparaison BLAST/FASTA
- La suite ...

Algorithmes d'alignement heuristiques

- Les algorithmes considérés (NW, SW) donnent des solutions exactes mais ils ne sont pas très rapides (complexité $O(n \times m)$)

  $\Rightarrow$ Nécessité d'algorithmes plus rapides : heuristiques
- Algorithmes heuristiques, les plus connus : BLAST [Altschul et al., 1990] et FASTA [Pearson et Lipman, 1988]
  - $\rightarrow$ Comparaison d'une séquence requête avec toutes les séquences d'une BD
  - $\rightarrow$ Score (normalisé) = mesure de similarité entre les séquence
  - $\rightarrow$ E-value = nb de fois que 2 seq auront par chance un score $\geq$ au score trouvé $\Rightarrow$ indication de la fiabilité du score
  - $\rightarrow$ Idée : les séquences similaires ont des segments communs de taille k quasiment identiques, ces petits segments vont servir de graînes
  - $\rightarrow$ Compromis entre sensibilité et sélectivité / spécificité

Sensibilité et sélectivité

- Par exemple : on regarde si des séquences sont membres d'une famille de protéine en fonction d'un certain seuil

	Séq. membres famille	Séq. non membres
Séq. au-dessus du seuil	Vrais positifs	Faux positifs
Séq. en dessous du seuil	Faux négatifs	Vrai négatifs

- Sensibilité $= \frac{\text{Vrais positifs}}{\text{Vrais positifs+Faux négatifs}}$

  $\Rightarrow$ capacité à détecter les vraies instances de l'objet recherché (VP) = capacité à détecter tout ce qui est intéressant/vrai sur le plan biologique au risque d'avoir beaucoup d'intrus

  $\Rightarrow$ minimiser les faux négatifs
- Sélectivité $= \frac{\text{Vrais positifs}}{\text{Vrais positifs+Faux positifs}}$

  $\Rightarrow$ capacité à rejeter les fausses instances (FP) = capacité à ne détecter que la réalité biologique et rien de plus au risque de ne pas retenir certaines bonnes informations

  $\Rightarrow$ minimiser les faux positifs

Plan

- Donner un score aux alignements
- Les heuristiques
- BLAST
  - $\rightarrow$ L'algorithme
  - $\rightarrow$ Un exemple d'utilisation
- FASTA
- Comparaison BLAST/FASTA
- La suite ...

BLAST [Altschul *et al.*, 1990] 24

Basic Local Alignment Search Tool

- But : trouver un score d'alignement local élevé entre une séquence requête et une base de données cible
- Capable de détecter précisément et rapidement des similarités entre des séquences nucléiques ou protéiques, sans sacrifier la sensibilité (= capacité à détecter tout ce qui est intéressant sur le plan biologique)
- Idée : les séquences similaires ont des segments communs de taille k quasiment identiques, ces petits segments vont servir de graînes
  (ADN $k = 11$, protéines $k = 3$)

BLAST : L'algorithme (1/2) 25

- On découpe S la séquence requête, en mots de longueur fixée (petite) : les mots requêtes (*query words*)
  ex S=TLSHAWRLSNETDKRPFIETALRDQHKKDYPEYKYQPRRRKNGKP
  query words : TLS, LSH, SHA, ...
- Pour les séquences protéiques :
  → Pour chaque mot requête, on détermine son voisinage

  ex pour RDQ

RDQ 16	QDQ 12	EDQ 11	RDN 11	RDB 11	BDQ 10	RDP 10
RBQ 14	REQ 12	HDQ 11	RDD 11	ADQ 10	XDQ 10	RDT 10
RDZ 14	RDR 12	ZDQ 11	RDH 11	MDQ 10	RQQ 10	RDY 10
KDQ 13	RDQ 12	RNQ 11	RDM 11	SDQ 10	RSQ 10	RDX 10
RDE 13	NDQ 11	RZQ 11	RDS 11	TDQ 10	RDA 10	DDQ 9 ...

  → Score seuil T , par exemple $T = 11$, seuls les mots voisins dont le score d'ali. (matrice de scores) avec le *query word* est $\geq T$ constituent la liste de mots similaires qui passent à l'étape suivante

BLAST : L'algorithme (2/2) 26

- Recherche des occurences des mots requêtes/similaires dans la banque

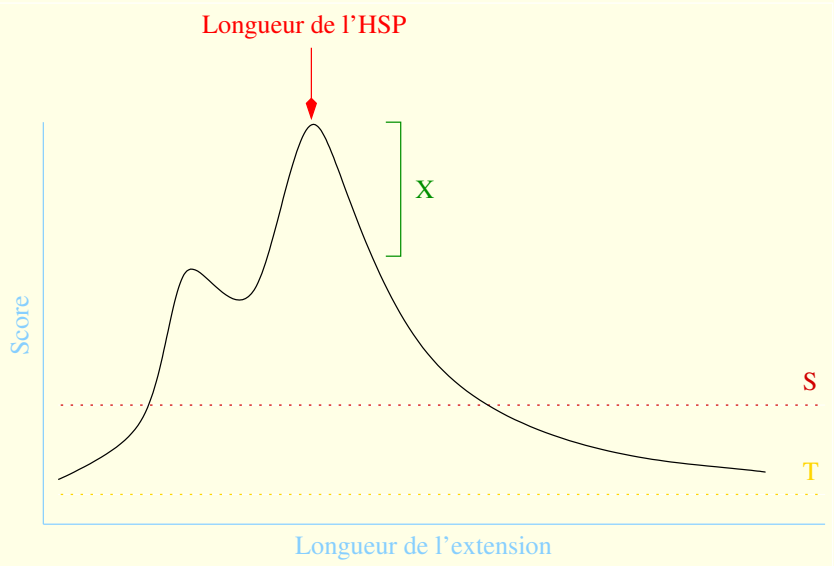
  ⇒ Alignement du mot requête (de la séquence requête) avec un mot requête/similaire (trouvé dans une séquence de la banque) $\geq T$

  ⇒ Extension jusqu'à la construction du meilleur alignement local dont le score est $\geq$ score seuil S (compromis entre longueur et score maximal)

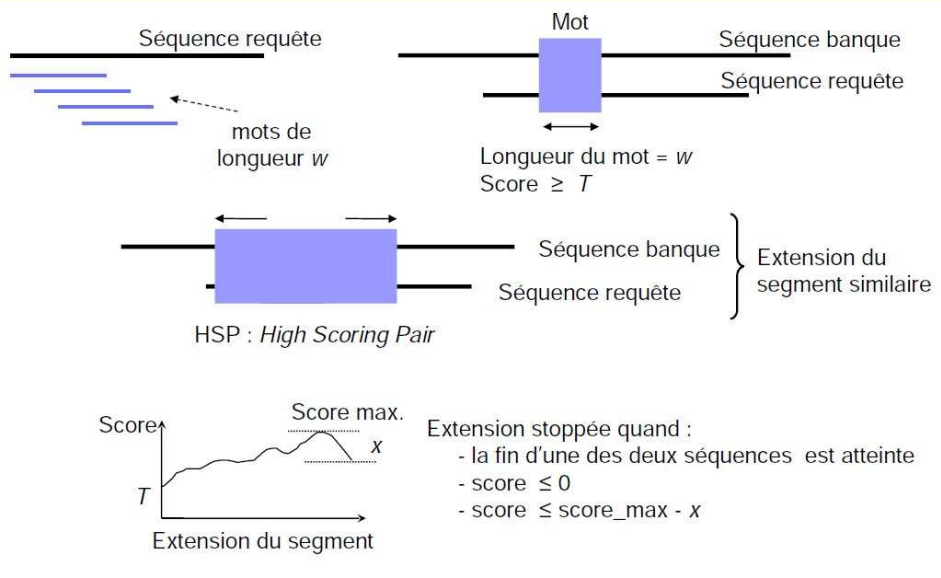
```
seq req          TLSHAWRLSNETDKRPFIETALRDQHKKDYPEYKYQPRRRKNGKP
                 ||    | ||| |   +||||+|  |||+|||||+|||||||||||+ |
seq de la banque TLESGWRLENPGDKRPFVEGALREQHKKDHPEYKYQPRRRKSVKN
                 ←←←←     →→→→
```

- L'alignement résultant est l'Hight-scoring Segment Pair ou HSP (plusieurs HSP possibles par paire de séquences)

BLAST : *Hit Extension* 27



BLAST : Résumé de l'algorithme



BLAST : Mesure de significativité

- Une fois qu'un HSP est identifié il est important de déterminer s'il est vraiment significatif
- BLAST calcule une valeur E en utilisant le score de l'alignement et d'autres paramètres : $E = kmNe^{-\lambda S}$

  m / N : nb de résidus seq req / seq de la banque ; S : score HSP

  λ et k : constantes estimées en fonction de la taille de la banque, des matrices/pénalités et composition en a.a. des séquences (normalisation)
- Pour chaque résultat, E représente le nb d'HSP ayant un score de S ou plus, que BLAST aurait pu trouver par chance
- La valeur de E fournit une mesure indiquant si l'HSP est un faux positif

Plus la valeur de E est petite, plus la similarité est significative

Les différentes versions de BLAST

Banque \ Séquences	nucléiques	protéiques
nucléique	**BLASTN - TBLASTX**	**TBLASTN**
protéique	**BLASTX**	**BLASTP**

= traduction dans les 6 phases

- BLASTN ⇒ recherche de séquences répétées, d'éléments régulateurs, . . .
- BLASTP ⇒ recherche de protéines homologues, étude de fonction de prot., . . .
- BLASTX ⇒ trouver les phases de lecture dans une séquence codante, analyse d'EST (Expressed Sequence Tag), . . .
- TBLASTN ⇒ localiser un gène dans un génome, rechercher des similitudes entre une prot. et une séquence génomique pas ou mal annotée, . . .
- TBLASTX (~ 36 BLASTP) ⇒ combine les avantages de tblastn et blastx mais la recherche est + longue

Les différentes versions de BLAST

- Gapped BLAST [Altschul *et al.*, 1997] : prise en compte des insertions et délétions (étape d'extension)
  ⇒ améliore sensiblement l'alignement et augmente la vitesse d'exécution
- Psi-BLAST [Altschul *et al.*, 1997] : itération de BLAST (*Position-Specific-Iterated Blast*)
  ⇒ augmente la sensibilité
- Phi-BLAST : motifs (*Pattern-Hit Initiated BLAST*)
  ⇒ augmente la sélectivité
- MEGABLAST : comparer plusieurs séquences à une base de données
  ⇒ pour séquences longues ou très similaires (> 95 %) : beaucoup plus rapide que BLAST, mais moins précis
- . . .

PSI-BLAST

- PSI-BLAST : Position-Specific-Iterated BLAST
- Permet d'identifier des protéines très divergentes
- PSI-BLAST utilise des PSSM (Position-Specific Scoring Matrices)
- Une PSSM est une représentation numérique d'un alignement multiple :

```
BIRC6_HUMAN   RRLAQEAVTLST.....S.........LPLSSSSSVFVRCde.......eRLDIMKVLITGP...ADTPY
COP10_ARATH   KRIQREMAELNI.....D.........PPPDCSAGPKGD-..........NLYHWIATIIGP...SGTPY
FTS1_HUMAN    YSLLAEFTLVVK.....Q.........KLPGVYVQPSYRS..........ALMWFGVIFI--...RHGLY
...
consensus     RRLMKELVLLLT.....Q.........KLPGVYVQPSYRS..........ALVWFGVIFI--...RHGLY
```

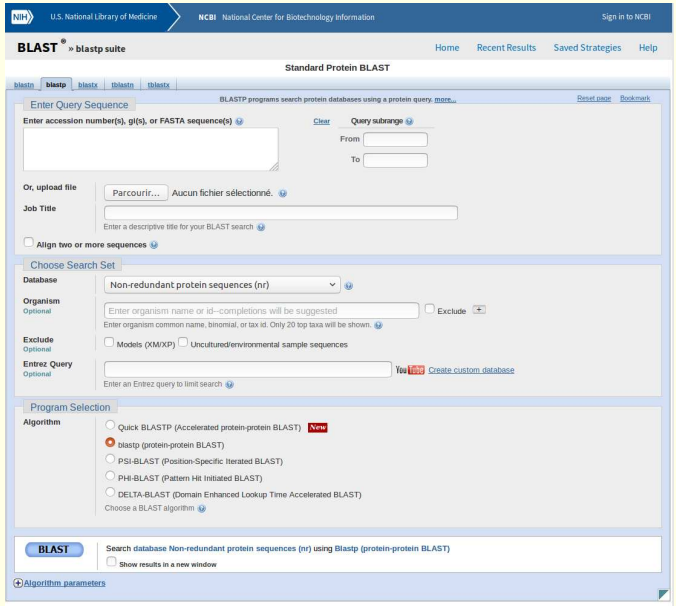
PSSM associée :

```
      A   B   C   D   E   F   G   H   I   K   L   M   N   P   Q   R   S   T   V   W   Y   Z
'R'; -10, -6,-25, -9, -2,-13,-19, -7,-16, 16,-14, -6, -1,-17,  1, 18, -6, -4,-13,-20, -4, -2;
'R'; -15,-12,-26,-13, -4,-13,-20, -5,-21, 16,-10, -6, -4,-20,  3, 47, -9, -6,-14,-21, -8, -4;
'L';  -9,-28,-22,-30,-20,  4,-32,-22, 25,-26, 35, 17,-26,-26,-19,-20,-24, -9, 16,-21, -2,-20;
'M';  -7,-14,-20,-19, -8, -8,-23, -7,  0, -5,  6, 14,-13,-19,  5, -3,-13, -7, -4,-19, -3, -2;
...
```

PSI-BLAST : la méthode

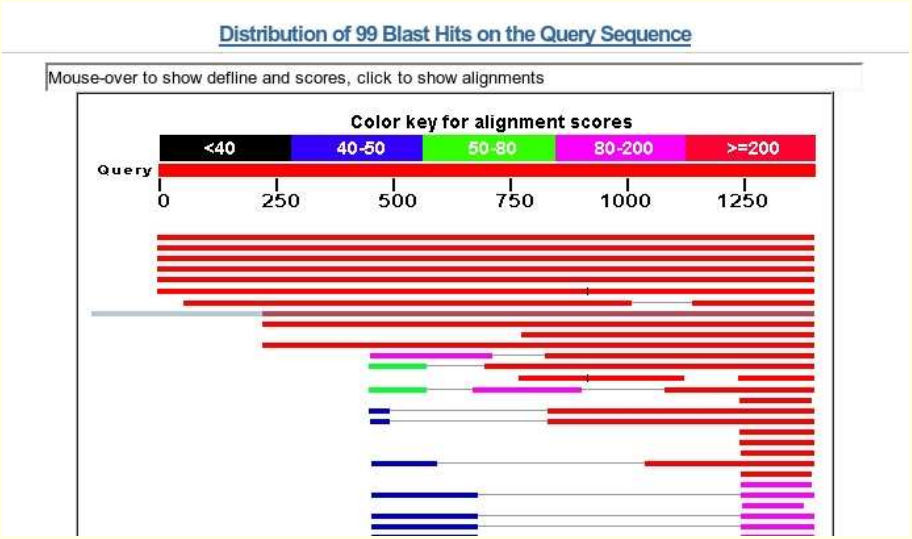
- Commence d'abord par un BLASTP avec la protéine requête
- Les séquences obtenues avec une valeur E inférieure à un certain seuil et la protéine requête sont alignées et une PSSM est construite
- La PSSM sert alors de requête pour rechercher de nouvelles séquences
- Une PSSM est reconstruite en intégrant ces séquences et une nouvelle recherche est effectuée
  → Cette étape est itérée plusieurs fois
- Fin : quand la recherche converge ou que la limite du nombre d'itération est atteint

BLAST : Exemple d'utilisation de BLASTP



BLAST : Résultats grapiques (1)

BLAST : Résultats grapiques (2) 36



BLAST : Liste des résultats 37

Sequences producing significant alignments:	Score (Bits)	E Value	
ref|NP_524317.2| prospero CG17228-PC, isoform C [Drosophila m...	1851	0.0	UG
dbj|BAA01464.1| prospero [Drosophila melanogaster]	1848	0.0	U
ref|NP_788636.1| prospero CG17228-PD, isoform D [Drosophila m...	1785	0.0	UG
gb|AAF05703.1|AF190403_1 homeodomain transcription factor Pro...	1775	0.0	
gb|AAA28841.1| Pros protein	1767	0.0	U
ref|NP_731565.2| prospero CG17228-PA, isoform A [Drosophila m...	1175	0.0	UG
sp|Q9U6A1|PROS_DROVI Protein prospero >gb|AAF06660.1|AF190405...	813	0.0	
ref|XP_309606.3| ENSANGP00000010936 [Anopheles gambiae str. PEST	714	0.0	G
ref|XP_001655942.1| homeobox protein prospero/prox-1 [Aedes a...	710	0.0	G
ref|XP_001359985.1| GA14403-PA [Drosophila pseudoobscura] >gb...	684	0.0	G
gb|EAA05345.4| AGAP004052-PA [Anopheles gambiae str. PEST]	625	6e-177	
ref|XP_971664.1| PREDICTED: similar to CG17228-PD, isoform D ...	474	3e-131	G
ref|XP_001602599.1| PREDICTED: similar to homeobox protein pr...	456	7e-126	G
pdb|1XPX|A Chain A, Structural Basis Of Prospero-Dna Interact...	345	2e-92	S
ref|XP_392355.3| PREDICTED: similar to prospero CG17228-PA, i...	330	3e-88	UG
pdb|1MIJ|A Chain A, Crystal Structure Of The Homeo-Prospero D...	312	1e-82	S
dbj|BAE87100.1| Prospero [Achaearanea tepidariorum]	299	1e-78	
emb|CAE00181.1| prospero protein [Cupiennius salei]	284	4e-74	
gb|AAL28228.1| GH11848p [Drosophila melanogaster]	240	5e-61	U
ref|XP_001666659.1| Hypothetical protein CBG22984 [Caenorhabd...	226	1e-56	
ref|NP_498760.1| C.Elegans Homeobox family member (ceh-26) [C...	226	1e-56	UG
gb|AAB30541.1| Prox 1=homeobox gene prospero homolog [mice, e...	219	9e-55	
ref|XP_781578.1| PREDICTED: similar to prospero-related homeo...	216	1e-53	UG
emb|CAF92934.1| unnamed protein product [Tetraodon nigroviridis]	202	2e-49	
emb|CAG04605.1| unnamed protein product [Tetraodon nigroviridis]	198	3e-48	

BLAST : Alignements 38

```
>ref|XP_392355.3| UG PREDICTED: similar to prospero CG17228-PA, isoform A [Apis mellifera]
Length=1146

 Score =  330 bits (847),  Expect = 3e-88, Method: Composition-based stats.
 Identities = 208/320 (65%), Positives = 234/320 (73%), Gaps = 50/320 (15%)

Query  1083  MMPVSLPTSVAIPNPSLHESKVFSPYSPFFNPHAAAGQATAAQLHQHHQQHHPHHQSMQL  1142
              M+PVSLPTSVAIPNPSLHES+VFSPYSPFFNPHA              H    P     L
Sbjct  876   MLPVSLPTSVAIPNPSLHESQVFSPYSPFFNPHAG------------HPGQVPPPGPHHL  923

Query  1143  SSSPPGSLGALMDSRDSPPLPHPPSMLHPALLAAAHHGGSPDYKTCLRAVMDAQDRQSEC  1202
               +SPP   G  +D RDSP  P P   LHPALLAAA H GSPDY        MD+ +R ++C
Sbjct  924   PASPP---GGGVDVRDSP--PLPHMPLHPALLAAAQH-GSPDYG---HLRMDSNERPNDC  974

Query  1203  NSADMQFDGMAPTISFYKQMQLKTEHQESLMAKHCESLTPLHSSTLTPMHLRKAKLMFFW  1262
              NS D+ +DG+ PT                                SS LTP+HLRKAKLMFFW
Sbjct  975   NSDDISYDGIQPT----------------------------SSMLTPIHLRKAKLMFFW  1005

Query  1263  VRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIKTPDDLL  1322
              VRYPSS++LKMYFPDI+FNKNNTAQLVKWFSNFREFYYIQMEKYARQAV+EG+K  DDL
Sbjct  1006  VRYPSSSILKMYFPDIRFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVSEGVKNADDLR  1065

Query  1323  IAGDSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKKSIYKII  1382
              + GDSE+YRVLNLHYNRNNHIEVP NFR+VVE TL+EFF+AIQGGKDTEQSWKKSIYK+I
Sbjct  1066  VGGDSEIYRVLNLHYNRNNHIEVPSNFRYVVEQTLKEFFKAIQGGKDTEQSWKKSIYKVI  1125

Query  1383  SRMDDPVPEYFKSPNFLEQL  1402
              SR+DDPVPEYFK+PNFL+QL
Sbjct  1126  SRLDDPVPEYFKTPNFLQQL  1145


 Score = 85.1 bits (209),  Expect = 4e-14, Method: Composition-based stats.
 Identities = 87/236 (36%), Positives = 127/236 (53%), Gaps = 33/236 (13%)

Query  674  GHPALPQGFP----PLLQHMGDMSHAAAMYQQFFFEQEARMAKEAAEQQQQQQQQQQQQ  729
             G PALP   P        + HMG          Q+ + E          QQQ   ++ +Q
Sbjct  488  GLPALPTEHPHAAAAAMYHMG---------QKLYLE------------QQQAALERMKQ  525
```

Les artefacts de la recherche avec BLAST 39

- Valeur E dépend de la taille de la banque et du score de l'alignement
  ⇒ Plus la banque est grande, plus la valeur E est grande
  ⇒ Plus les séquences sont courtes (score petit), plus la valeur E est grande
- k et λ ont été estimés et sont fixes
  ⇒ Valeurs adaptées à la majorité des données mais pas toutes
- Score élevé dû à des régions de faible complexité ou d'éléments répétés comme les séquences LINE, SINE ou Alu
  ⇒ Masquage de ces régions (par N ou X pour faible compl.)
- Gènes inconnus (gène "orphelin")
- Attention à la fiabilité des résultats qui peuvent être des protéines hypothétiques (issues de prédiction) ou des EST (moins fiables)

Conclusions sur BLAST (1/2) 40

- Résultat d'un BLAST
  - → Un résumé des séquences de la banque produisant un alignement significatif ordonnées selon le Score (Bits)
    - ▷ Score normalisé (Score (Bits))
    - ▷ Valeur E (E-Value , Expect)
    - ▷ Liens vers les banques de données
  - → Les alignements par paire de chaque HSP
    - ▷ Identites : nb paires identiques / nb paires
    - ▷ Positives : nb paires avec poids positif / nb paires
    - ▷ Gaps : nb insertions ou délétions / nb paires
    - ▷ Longueurs alignement et séquence (couverture)

Conclusions sur BLAST (2/2) 41

- Où couper dans la liste des résultats ?
  - → Nucléotides : valeur E $< 10^{-6}$ et identité > 70 %
  - → Protéines : valeur E $< 10^{-3}$ et identité > 25 %

Ne pas utiliser ces limites sans réfléchir !
Est-ce que la matrice utilisée est correcte ? Regarder l'alignement. Quel sens biologique ? . . .

- Différents paramétrages possibles en fonction des données
  - → Filtrage des données : régions de faible complexité, régions répétées
  - → Recherche de motifs : paramètres adaptés à l'analyse de séquences courtes
  - → . . .

Plan 42

- Donner un score aux alignements
- Les heuristiques
- BLAST
- FASTA
- Comparaison BLAST/FASTA
- La suite ...

FASTA [Pearson et Lipman, 1988] 43

FAST Alignment

- Le premier programme largement utilisé pour la recherche de similarité dans les banques
- Plusieurs versions disponibles

Programme	Requête	Base de données	équiv.
FASTA	Nucléotide	Nucléotide	BLASTN
	Protéine	Protéine	BLASTP
FASTX/FASTY	ADN	Protéine	BLASTX
TFASTX/TFASTY	Protéine	ADN traduit	TBLASTN

- Similaire à BLAST : méthode à graînes

FASTA : L'algorithme

1. Localiser les mots de lg $ktup$ matchant exactement entre la séquence requête et les séquences de la BD
   (ADN $ktup = 4$ ou 6, protéines $ktup = 1$ ou 2)

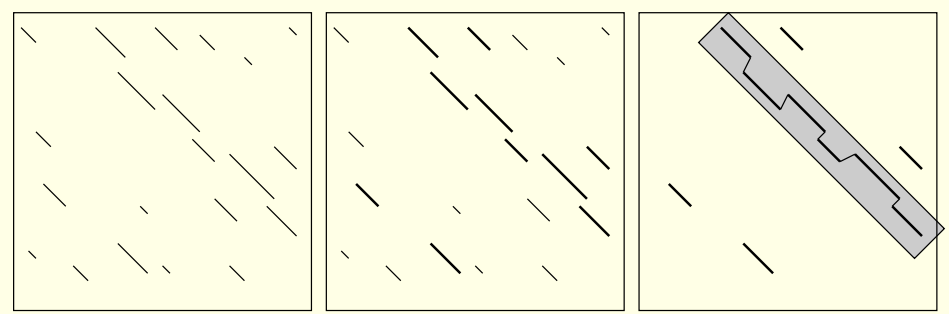
2. Regarder les 10 diagonales de meilleurs scores pour chaque alignement 2 à 2 $\Rightarrow$ Score max = init1

3. Essayer de joindre les diagonales entre elles ($\sim$ extension de hit de BLAST) $\Rightarrow$ 2 diagonales sont réunies si score résultat = initn $\geq$ init1

4. Sélectionner les candidats parmi tous les ali. 2 à 2 avec les meilleurs scores et réaligner leurs diagonales avec le principe de programmation dynamique (banded S&W et full S&W pour les séquences protéiques) $\Rightarrow$ Scores alignement = opt (optimized - banded S&W) et S&W score (full S&W)

5. Calculer les significativités des résultats : calcul d'une valeur E représentant la probabilité que le résultat trouvé le soit par chance

FASTA : Phases 1, 2 & 3



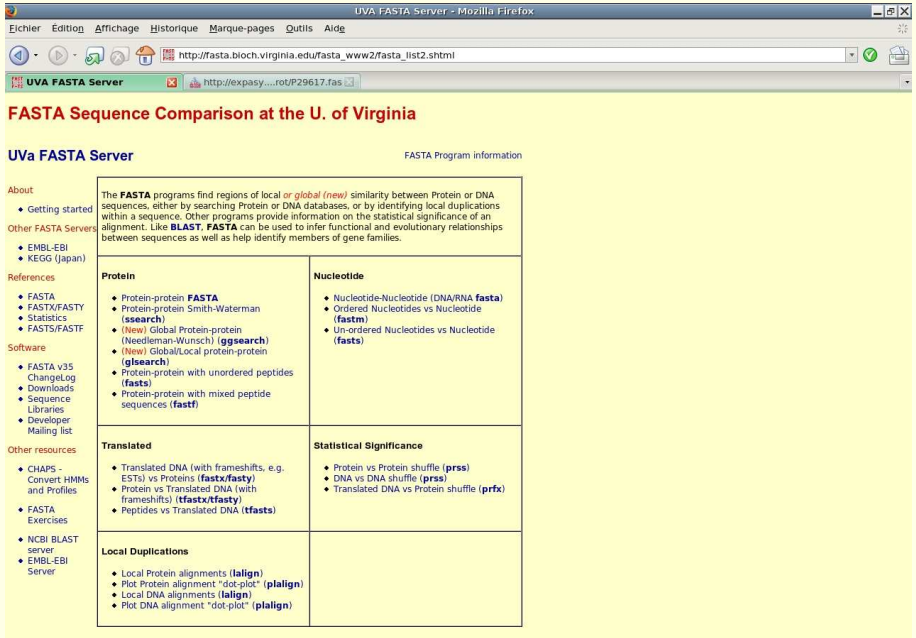
Score max = init1

Score diagonales réunies = initn

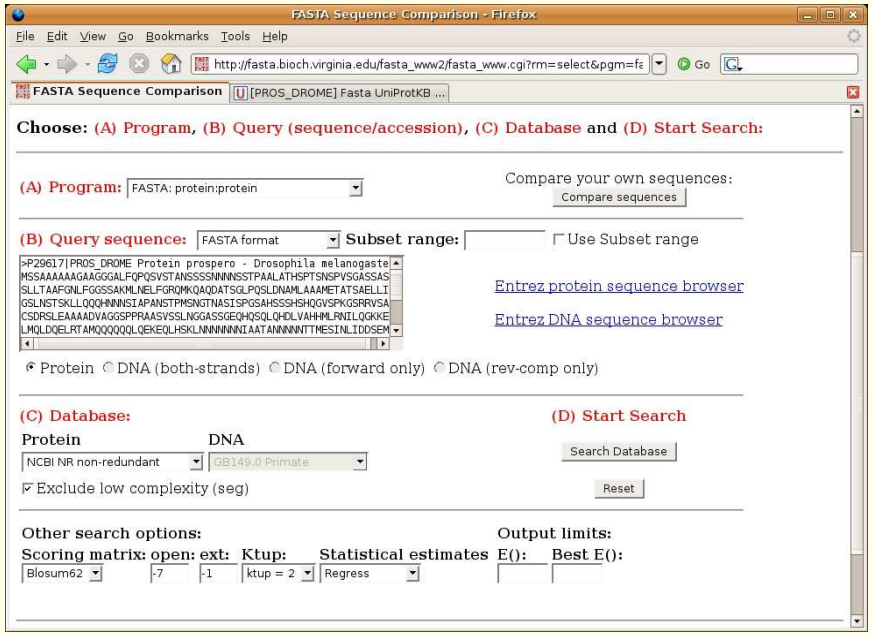
$\Downarrow$

Score alignement (S&W) = opt et S&W score

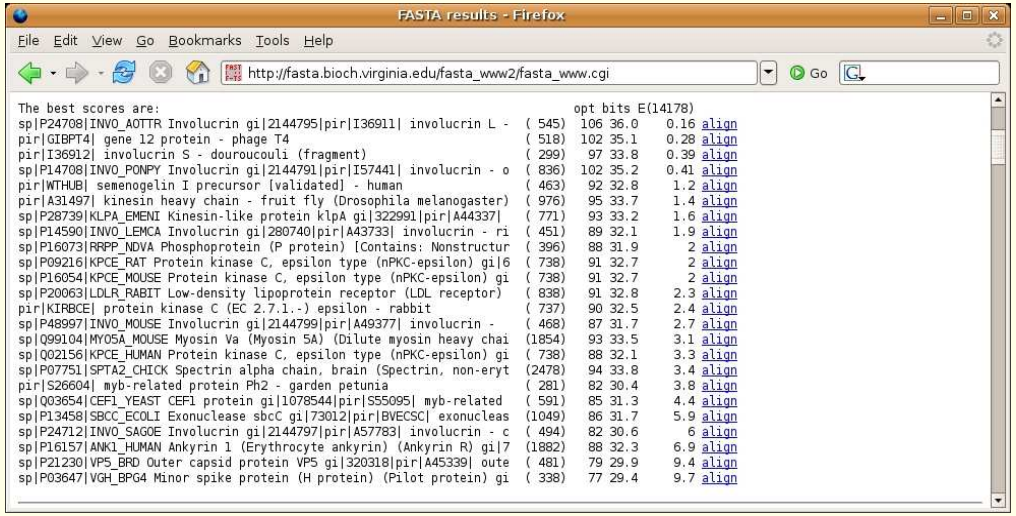
FASTA : Page d'accueil



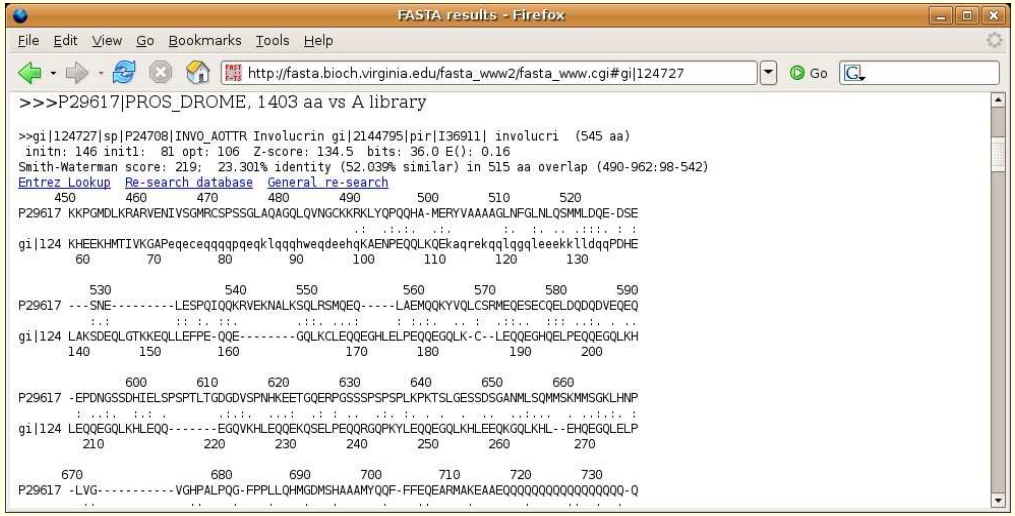
FASTA : Exemple d'utilisation



FASTA : Liste des résultats



FASTA : Exemple d'alignement



Conclusions sur FASTA (1/2)

- Résultat obtenu avec FASTA
  - → Un résumé des séquences de la banque produisant un alignement significatif ordonnées selon le score Initn : longueurs des séquences, score optimal (opt) et normalisé (bits) de l'alignement , valeur E (E())
  - → Les alignements par paire entre la séquence requête et les séquences de la banque : Init1, Initn, Z-score, S&W score, Identity / Similarity, overlap
- Z-score : score maximum attendu normalisé (dérivé du score opt avec une correction en fontion de la longueur de la séquence)
- Valeur E : nombre statistique de séquences pouvant donner ce Z-score par hasard

Conclusions sur FASTA (2/2)

- Interprétations
  - → init1 = initn ⇒ 1 seul segment homologue détecté
  - → initn > init1 ⇒ plusieurs régions homologues existent
  - → opt > initn ⇒ homologie diluée le long de la séquence
  - → opt < initn ⇒ régions homologues séparées par une insertion importante
  - → opt > z-score et valeur E faible ⇒ alignement significatif
  - → valeur E < 0.01 ⇒ séquences homologues
  - → valeur E entre 1 et 10 ⇒ similarité plus lointaine
- Comme pour BLAST, attention aux artefacts de recherche

Plan

- Donner un score aux alignements
- Les heuristiques
- BLAST
- FASTA
- Comparaison BLAST/FASTA
- La suite ...

Comparaison BLAST/FASTA

- Quel est le meilleur ? pas de bonne réponse

→ FASTA commence la recherche en cherchant des mots exacts alors que BLAST autorise les substitutions conservatives

→ FASTA et BLAST retournent plusieurs alignements / HSP par paire requête/séquences résultats.
⇒ Ce n'était pas vrai pour les versions antérieures à FASTA36

→ FASTA utilise un alignement local plus rigoureux, donc ses alignements finaux sont plus fiables

→ BLAST est plus rapide que FASTA

Références

- Fiches 13 à 18 dans "Bio-informatique Principes d'utilisation des outils"
  [Tagu & Risler,2010]

- Chap 11 dans "Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins, 3rd Edition"
  [Baxevanis & Ouellette,2005]

- Illustrations :

  → BLAST : http://www.ncbi.nlm.nih.gov/blast/

  → FASTA :
  http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml

Plan

- Donner un score aux alignements
- Les heuristiques
- BLAST
- FASTA
- Comparaison BLAST/FASTA
- La suite ...

- Et si l'on veut comparer plusieurs séquences à la fois ?
- Plusieurs séquences $\Rightarrow$ meilleure fiabilité
- Multiple Sequence Alignement ou MSA
- Comme pour les alignements de 2 séquences, il existe des MSA globaux et des MSA locaux