Introduction à la bioinformatique (UE SSV3U15)
TP2. Du gène à la protéine
Diaporama d'accompagnement du TP

Jacques van Helden (Aix-Marseille Université)
ORCID 0000-0002-8799-8584

Objectifs du TP 2 – Du gène à la protéine

- **Mettre en pratique les concepts** présentés au CM “Du gène au génome”
  - Gène, transcrit, exon, intron, UTR, région codante
- Apprendre à utiliser l’interface de la base de données de gènes du NCBI, qui est l’une des grandes références internationales (partenaire américain du International Nucleotide Sequence Database Consortium, INSDBC)
  - **Consulter les annotations** d’un gène dans une base de données de référence
  - **Extraire les différents types de séquences** associées (gène, ARNm, protéine)
  - Comprendre le format **FASTA** (le plus utilisé pour les séquences macromoléculaires)
- Effectuer des **alignements par paire** entre ces trois types de séquences
  - Utilisation de différentes modalités de BLAST pour aligner des séquences nucléiques (traduites) et des séquences protéiques

Déroulé du TP

Etapes

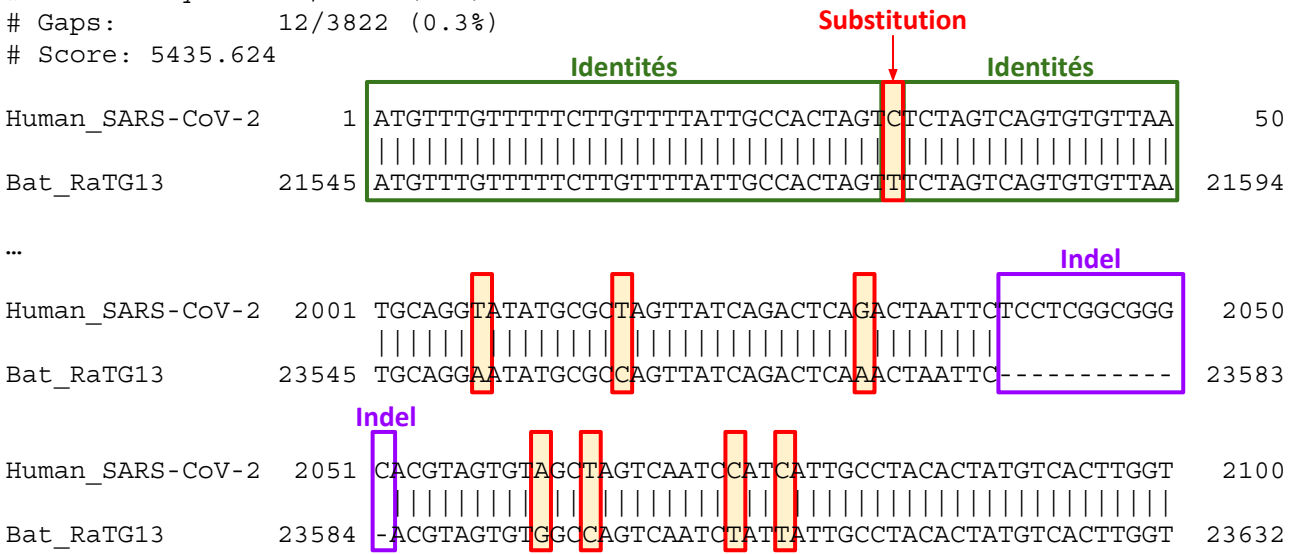
- Requête pour trouver un gène sur la base de données NCBI-Gene
- Exploration des annotations de ce gène sur l'interface graphique de NCBI-Gene
- Exercice 1: Extraction des séquences du gène, de l'ARNm et de la protéine
- Exercice 2 : Comparaison d'un gène et de son ARNm - alignement global avec needle
- Exercice 3 : Comparaison de l'ARNm et de la protéine - alignement local avec BLAST
- Exercice 4 - Comparaison de l'ARNm et protéine - alignement global avec needle
- Exercice 5 - Comparaison d'un gène et de son ARNm - alignement local avec BLAST

Complétion

- Tous les exercices doivent être réalisés par chaque étudiant.
- En principe, les exercices 1 à 3 devraient être faits en séance (avec explications par les enseignants).
- L'exercice 1 prend plus de temps que les exercices suivants.
- Si nécessaire, les exercices 4 et 5 peuvent être terminés ultérieurement.

Alignement de séquences – Gènes S de SARS-CoV-2 et RaTG13

```
# Aligned_sequences: 2
# 1: Human_SARS-CoV-2_BetaCoV/Wuhan/IPBCAMS-WH-01/2019
# 2: Bat_RaTG13
#
# Length: 3822
# Identity:    3549/3822 (92.9%)
# Similarity:    NA/3822 (NA%)
# Gaps:          12/3822 (0.3%)
# Score: 5435.624
```



Note

- "**Indel**" signifie "Insertion ou délétion"
- Sur base de ce seul résultat, one ne peut pas déterminer si la différence observée provient d'une insertion chez un ancêtre de SARS-CoV-2, ou d'une délétion chez un ancêtre de RaTG13

Matrice de substitutions

- Une ***matrice de substitution*** associe un score à chaque paire de résidus qu'on peut trouver dans un alignement.
  - Chaque ligne et chaque colonne représente l'un des résidus (4 nucléotides, 20 acide aminés).
  - La diagonale correspond aux identités.
  - Le triangle inférieur correspond à des substitutions.
  - Le triangle supérieur est symétrique au triangle inférieur, il n'est pas nécessaire d'indiquer les nombres.
- Les ***scores négatifs*** sont considérés comme des pénalités associées à certaines substitutions qu'on n'observe que rarement dans les alignements. Les algorithmes d'alignements tenteront donc d'éviter ces substitutions.
- Les ***scores positifs*** correspondent à des substitutions qu'on observe plus souvent que prévu, dans les alignements d'un grand nombre de séquences. Ceci suggère que ces substitutions particulières sont moins dommageable que d'autres, et on les qualifie donc de « ***substitutions conservatives*** » ou encore de « ***mutations ponctuelles acceptées*** » (***PAM***).
- Au sein d'un alignement, le terme ***similarité*** désigne les positions où se superposent des résidus ayant un score positif dans la matrice de substitution (identité ou substitution conservative).

Matrice de substitutions entre nucléotides

	A	C	G	T
A	1	-2	-1	-2
C	-2	1	-2	-1
G	-1	-2	1	-2
T	-2	-1	-2	1

Scores	
1	identité
-1	transition
-2	transversion

Matrice de substitutions entre acides aminés

Ala	A	4																		
Arg	R	-1	5																	
Asn	N	-2	0	6																
Asp	D	-2	-2	1	6															
Cys	C	0	-3	-3	-3	9														
Gln	Q	-1	1	0	0	-3	5													
Glu	E	-1	0	0	2	-4	2	5												
Gly	G	0	-2	0	-1	-3	-2	-2	6											
His	H	-2	0	1	-1	-3	0	0	-2	8										
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4									
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4								
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5							
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5						
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6					
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7				
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4			
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5		
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Utilisation d'une matrice de substitution pour calculer le score d'un alignement

- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement (*i de 1 à L*) , des scores des paires de résidus ($r_{1,i}$ et $r_{2,i}$).
- Les "gaps" sont traités par une règle spécifique reposant sur deux paramètres de pénalité:
  - Pénalité d'ouverture de gap (***go***)
    Valeur typique : -10
  - Pénalité d'extension de gap (***ge***)
    valeur typique: -1

Exercice

Dans l'alignement ci-dessous,

- identifiez les identités, les transitions et les transversions
- en vous basant sur la matrice de substitution, calculez le score de l'alignement

	A	C	G	T
A	1	-2	-1	-2
C	-2	1	-2	-1
G	-1	-2	1	-2
T	-2	-1	-2	1

Scores	
1	identité
-1	transition
-2	transversion

$$S = \sum_{i=1}^{L} s_{r_{1,i} r_{2,i}}$$

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
	A	T	A	A	C	T	T	A	G	G	-	-	-	-	-	C	C	C	A	T	G
	A	A	T	C	G	T	G	G	G	C	A	T	T	A	A	T	T	A	G	T	G

Utilisation d'une matrice de substitution pour calculer le score d'un alignement

	A	C	G	T
A	1	-2	-1	-2
C	-2	1	-2	-1
G	-1	-2	1	-2
T	-2	-1	-2	1

Scores	
1	identité
-1	transition
-2	transversion

- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement (*i de 1 à L*) , des scores des paires de résidus ($r_{1,i}$ et $r_{2,i}$).
- Les "gaps" sont traités par une règle spécifique reposant sur deux paramètres de pénalité:
  - Pénalité d'ouverture de gap (***go***)
    Valeur typique : -10
  - Pénalité d'extension de gap (***ge***)
    valeur typique: -1

$$S = \sum_{i=1}^{L} s_{r_{1,i} r_{2,i}}$$

```
i    1  2  3  4  5  6  7  8  9  10 11 12 13 14 15 16 17 18 19 20
     A  T  A  A  C  T  T  A  G  G  -  -  -  -  C  C  C  A  T  G
     |  |           |  |     |  |  |  GO ge ge ge |  |        |  |
     A  T  G  C  C  T  G  A  G  G  A  T  T  A  C  C  A  G  T  G
     1 +1 -1 -2 +1 +1 -2 +1 +1 +1 -10 -1 -1 -1 +1 +1 -2 -1 +1 +1 = -10
```

Utilisation d'une matrice de substitution pour calculer le score d'un alignement

$$S = \sum_{i=1}^{L} s_{r_{1,i} r_{2,i}}$$

Ala	A	4																		
Arg	R	-1	5																	
Asn	N	-2	0	6																
Asp	D	-2	-2	1	6															
Cys	C	0	-3	-3	-3	9														
Gln	Q	-1	1	0	0	-3	5													
Glu	E	-1	0	0	2	-4	2	5												
Gly	G	0	-2	0	-1	-3	-2	-2	6											
His	H	-2	0	1	-1	-3	0	0	-2	8										
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4									
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4								
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5							
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5						
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6					
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7				
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4			
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5		
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement (*i de 1 à L*) , des scores des paires de résidus ($r_{1,i}$ et $r_{2,i}$).
- Les "gaps" sont traités par une règle spécifique reposant sur deux paramètres de pénalité:
  - Pénalité d'ouverture de gap (***go***)
    Valeur typique : -10
  - Pénalité d'extension de gap (***ge***)
    valeur typique: -1

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
	R	L	A	S	V	E	T	D	M	P	-	-	-	-	-	L	T	L	R	Q	H
	T	L	T	S	L	Q	T	T	L	K	N	L	K	E	M	A	H	L	G	T	H

Utilisation d'une matrice de substitution pour calculer le score d'un alignement

$$S = \sum_{i=1}^{L} s_{r_{1,i} r_{2,i}}$$

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	4																			
Arg	R	-1	5																		
Asn	N	-2	0	6																	
Asp	D	-2	-2	1	6																
Cys	C	0	-3	-3	-3	9															
Gln	Q	-1	1	0	0	-3	5														
Glu	E	-1	0	0	2	-4	2	5													
Gly	G	0	-2	0	-1	-3	-2	-2	6												
His	H	-2	0	1	-1	-3	0	0	-2	8											
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

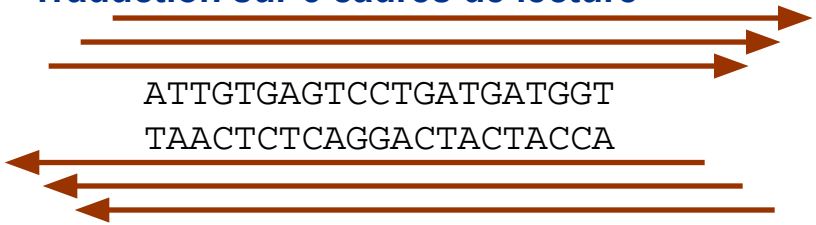
- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement (*i de 1 à L*) , des scores des paires de résidus ($r_{1,i}$ et $r_{2,i}$).
- Les "gaps" sont traités par une règle spécifique reposant sur deux paramètres de pénalité:
  - Pénalité d'ouverture de gap (***go***)
    Valeur typique : -10
  - Pénalité d'extension de gap (***ge***)
    valeur typique: -1

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
	R	L	A	S	V	E	T	D	M	P	-	-	-	-	-	L	T	L	R	Q	H	
	.	|	.	|	:	:	|	.	:	.	go	ge	ge	ge	ge	.	.	|	.	.	|	
	T	L	T	S	L	Q	T	T	L	K	N	L	K	E	M	A	H	L	G	T	H	
S	-1	+4	+0	+4	+1	+2	+5	-1	+2	-1	-10	-1	-1	-1	-1	-1	-2	+4	-2	-1	+8	= 7

Traduction d'une séquence nucléique dans les 6 cadres de lectire

- Si l'on dispose d'une séquence nucléique, on peut facilement déduire la séquence de la protéine qui pourrait être produite par sa traduction, sur chacun des 6 brins.
- Si cette séquence n'est pas codante, on s'attend à trouver des codons stop assez fréquemment (3 codons sur 64).
- Cependant, rien n'empêche d'aligner les 6 séquences ainsi produites avec d'autres séquences peptidiques.

Traduction sur 6 cadres de lecture



Résultat

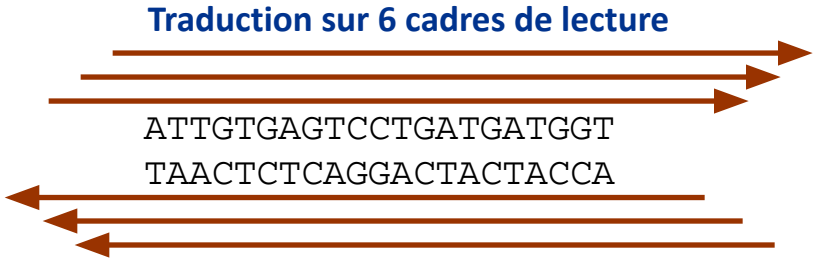
```
F1       I   V   S   P   D   D   G
F2        L   *   V   L   M   M   V
F3          C   E   S   *   *   W   X
        1 ATTGTGAGTCCTGATGATGGT 21
          ----:----|----:----|-
        1 TAACACTCAGGACTACTACCA 21
F6         X   T   L   G   S   S   P
F5       X   Q   S   D   Q   H   H
F4         N   H   T   R   I   I   T
```

- Outil: http://www.ebi.ac.uk/Tools/st/emboss_sixpack/

Modalités de BLAST

Le logiciel BLAST présente 5 modalités différentes en fonction du type des séquences (peptidique ou nucléotidique) de requête et de la base de données.

Pour les comparaisons entre séquences nucléotidiques et peptidiques, la séquence nucléotidique est traduite dans les 6 phases de lecture (3 par brin), et on lance ensuite une recherche de similarité “protéine *versus* protéine”.



Séquence requête	Base de données	Outil	Exemples d’applications
peptidique	peptidique	**blastp**	En partant d’une protéine de fonction connue, collecter les protéines similaires dans la base de données Uniprot afin de constituer la famille de protéine supposées homologues.
nucléique	nucléique	**blastn**	Comparer les séquences d’ARNm aux séquences génomiques.
			Aligner un ARN d’interférence (ARNi) sur un génome pour détecter ses cibles potentielles.
nucléique (traduite dans les 6 cadres)	peptidique	**blastx**	Après avoir séquencé un morceau d’ADN, chercher des fragments potentiellement codants (susceptibles de produire un polypeptide similaire à des protéines connues) dans cette séquence même si on ne connaît pas la position des gènes.
peptidique	nucléique (traduite dans les 6 cadres)	**tblastn**	Identifier une région génomique susceptible de coder pour un homologue d’une protéine d'intérêt.
			Identifier dans un génome les pseudo-gènes (gènes défectifs, qui peuvent contenir un ou plusieurs codons stop) correspondant à une protéine d’intérêt.
nucléique (traduite dans les 6 cadres)	nucléique (traduite dans les 6 cadres)	**tblastx**	A partir d’une séquence d’ADN, identifier des segments de régions codantes ayant une contrepartie dans un génome ou une base de données de référence