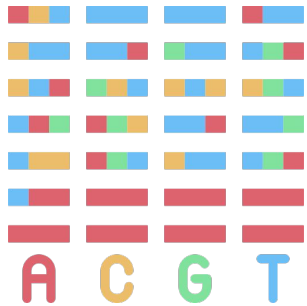


# **Alignement multiple heuristique par la méthode CLUSTAL**

Karine DUONG - M2 BI

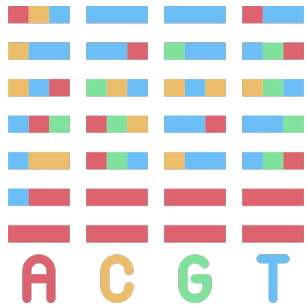


# L'alignement multiple de séquence (MSA)

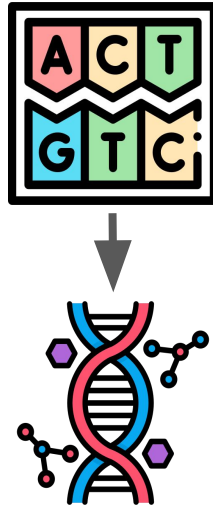


Trouver des régions  
conservées

# L'alignement multiple de séquence (MSA)

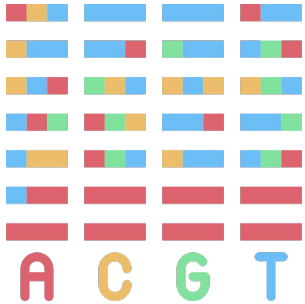


Trouver des régions  
conservées

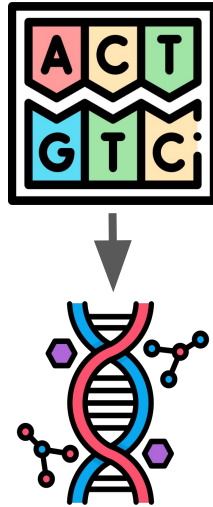


Relier la séquence  
à la structure  
et à la fonction

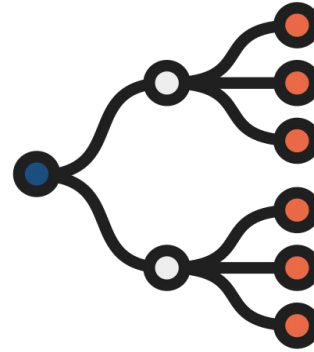
# L'alignement multiple de séquence (MSA)



Trouver des régions  
conservées

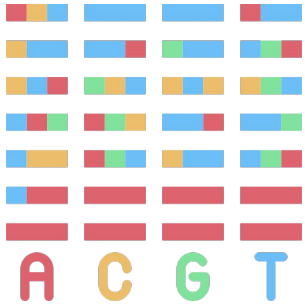


Relier la séquence  
à la structure  
et à la fonction

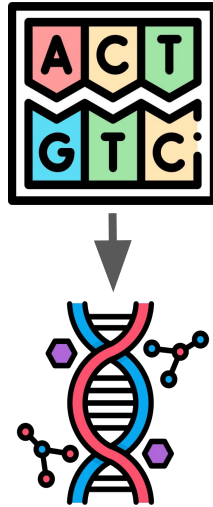


Construire  
un arbre  
phylogénétique

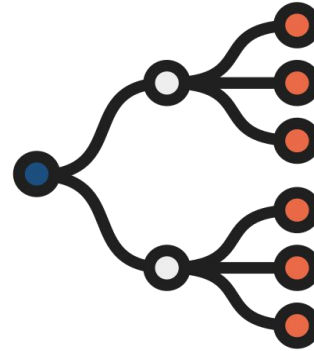
# L'alignement multiple de séquence (MSA)



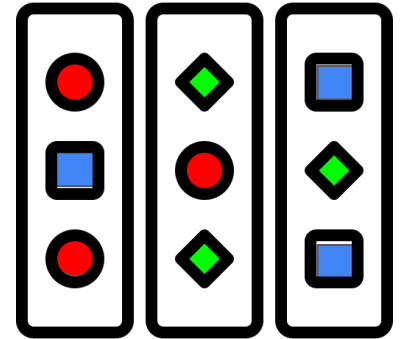
Trouver des régions  
conservées



Relier la séquence  
à la structure  
et à la fonction



Construire  
un arbre  
phylogénétique



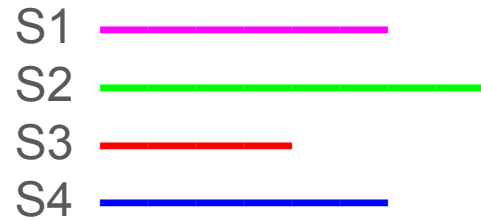
Etudier les  
variations  
génétiques

# Implémentation de mon projet






***NumPy***

# Algorithme



# Algorithme

S1   
S2   
S3   
S4 

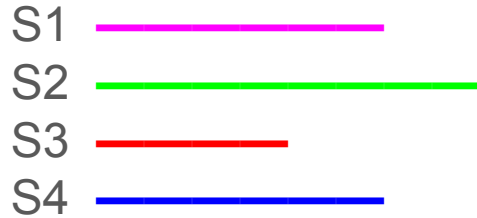
↓ Alignement  
par paire

	$S_1$	$S_2$	$S_3$	$S_4$
$S_1$		4	9	4
$S_2$			4	7
$S_3$				4
$S_4$				

Matrice de similarité



# Algorithme

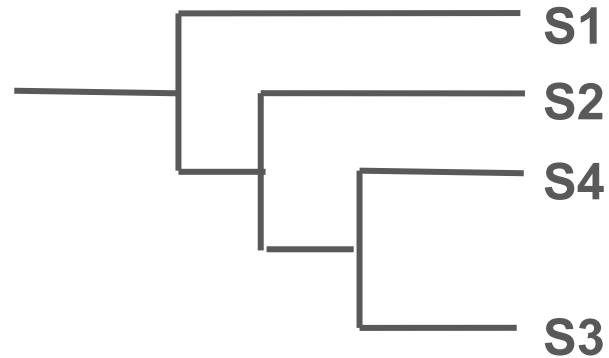


↓  
Alignement  
par paire





	S <sub>1</sub>	S <sub>2</sub>	S <sub>3</sub>	S <sub>4</sub>
S <sub>1</sub>		4	9	4
S <sub>2</sub>			4	7
S <sub>3</sub>				4
S <sub>4</sub>				

Matrice de similarité

→  
UPGMA



# Algorithme

S1   
S2   
S3   
S4 

↓  
Alignement  
par paire

	S <sub>1</sub>	S <sub>2</sub>	S <sub>3</sub>	S <sub>4</sub>
S <sub>1</sub>		4	9	4
S <sub>2</sub>			4	7
S <sub>3</sub>				4
S <sub>4</sub>				

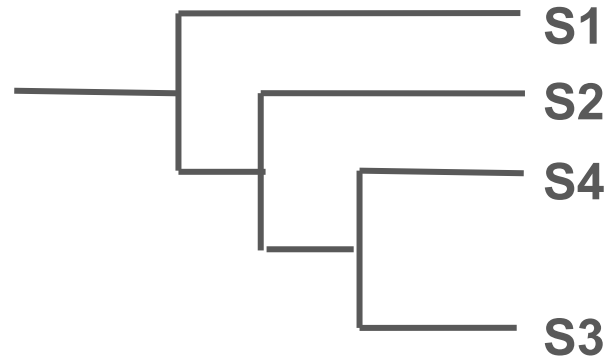
Matrice de similarité

→  
UPGMA

Alignement de:

- S4 et S3
- (S4, S3) et S2
- (S4, S3, S2) et S1

↑  
Alignement  
progressif



# Résultats

['QMS45321.1', 'XP\_004050475.2', 'NP\_000198.1'] (A)

QMS45321.1: MGSETIKPAGAQPPSALQDRLHQKRPSSRSVPRAFSGGLRIPGWLDPRLQCSREDVAGLVK  
 XP\_004050475.2: MGSETIKPVGTQPPSALQDRLHQKRPSSRSVPRAF  
 NP\_000198.1:

QMS45321.1: HVGVSPPGAPRQGTWPSACLSACLDPHCPASAMALWMRLPLLALLALWGPDPAAAFVNQHLG  
 XP\_004050475.2: A S DHCPASAMALWMRLPLLALLALWGPDPAAAFVNQHLG  
 NP\_000198.1: MALWMRLPLLALLALWGPDPAAAFVNQHLG

QMS45321.1: SHLVEALYLVCGERGFFYTPKTRREAEDLVGGQVELGGGPGAGSLQPLALEGSLQKRGIVEQC  
 XP\_004050475.2: SHLVEALYLVCGERGFFYTPKTRREAEDLVGGQVELGGGPGAGSLQPLALEGSLQKRGIVEQC  
 NP\_000198.1: SHLVEALYLVCGERGFFYTPKTRREAEDLVGGQVELGGGPGAGSLQPLALEGSLQKRGIVEQC

QMS45321.1: CTSICSLYQLENYCN  
 XP\_004050475.2: CTSICSLYQLENYCN  
 NP\_000198.1: CTSICSLYQLENYCN

Figure 2: **Comparaison des résultats** produits par notre programme sur le fichier *data/insuline\_sequence.fasta*. (A) L'ordre des clusters produit par UPGMA, qui se lit de gauche à droite pour avoir les embranchements successifs c'est-à-dire ((QMS, XP), NP). (B) L'alignement de ses trois séquences selon l'ordre donné par les clusters.

XP\_004050475.2 0.00816994  
 QMS45321.1 0  
 NP\_000198.1 0

(A)

XP\_004050475.2 MGSETIKPVGTQPPSALQDRLHQKRPSSRSVPRAFA----- 36  
 QMS45321.1 MGSETIKPAGAQPPSALQDRLHQKRPSSRSVPRAFSGGLRIPGWLDPRLQCSREDVAG 60  
 NP\_000198.1 ----- 0

(B)

XP\_004050475.2 -----SDHCPASAMALWMRLPLLALLALWGPDPAAAFV 69  
 QMS45321.1 LVKHVGVSPGAPRQGTWPSACLSACLDPHCPASAMALWMRLPLLALLALWGPDPAAAFV 120  
 NP\_000198.1 -----MALWMRLPLLALLALWGPDPAAAFV 26  
 \*\*\*\*\*

XP\_004050475.2 NQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLVGGQVELGGGPGAGSLQPLALEGSL 129  
 QMS45321.1 NQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLVGGQVELGGGPGAGSLQPLALEGSL 180  
 NP\_000198.1 NQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLVGGQVELGGGPGAGSLQPLALEGSL 86  
 \*\*\*\*\*

XP\_004050475.2 QKRGIVEQCCTICSILYQLENYCN 153  
 QMS45321.1 QKRGIVEQCCTICSILYQLENYCN 204  
 NP\_000198.1 QKRGIVEQCCTICSILYQLENYCN 110  
 \*\*\*\*\*

Figure 3: **Comparaison des résultats** produits par CLUSTAL Omega sur le fichier *data/insuline\_sequence.fasta*. (A) Le dendrogramme produit par l'étape de clustering. (B) L'alignement de ses trois séquences selon l'ordre donné par les clusters.

# Résultats

<i>Nombre de séquences</i>	<i>Temps d'exécution</i>
3	0.02 sec
9	0.3 sec
25	4min

Tableau 1: **Récapitulatif du temps d'exécution** pour un nombre de séquences variables et pour des séquences d'environ 150-200 acides aminés.

# Conclusion

Notre version simplifié de l'implémentation  
de Clustal fonctionne

Cependant, quelques points aurait pu être  
amélioré, tels que :



[https://github.com/KarinDuong/clustal\\_alignement](https://github.com/KarinDuong/clustal_alignement)

# Conclusion

Notre version simplifié de l'implémentation de Clustal fonctionne

Cependant, quelques points aurait pu être amélioré, tels que :

- **Paralléliser l'alignement par paire**
- **Utilisation d'un score de gap affine**
- **Améliorer l'affichage des résultats**



[https://github.com/KarinDuong/clustal\\_alignement](https://github.com/KarinDuong/clustal_alignement)