Université de Rennes

UFR SVE
Sciences de la Vie et de l'Environnement

MUSÉUM NAL HIST NATURELLE

Développement d'une méthodologie pour l'annotation structurale et fonctionnelle de séquences génomiques fonctionnellement inconnues

Jérémy Rousseau
Sous la supervision de Mathilde Carpentier et Lucie Bittner

Atelier de Bio-Informatique, Institut de SYstématique Evolution Biodiversité
Muséum National d'Histoire Naturelle

Annotation fonctionnelle et structurale

Objectif : attribuer une fonction ou une structure à une séquence

Les méthodes :

- **Alignement de séquences => BLAST**
- **Chaînes de Markov cachées => HMMER**

Annotation fonctionnelle et structurale

Objectif : attribuer une fonction ou une structure à une séquence

Les méthodes :

- Alignement de séquences => BLAST
- Chaînes de Markov cachées => HMMER

Biais des banques de données

Annotation fonctionnelle et structurale

Objectif : attribuer une fonction ou une structure à une séquence

Les méthodes :

- Alignement de séquences => BLAST
- Chaînes de Markov cachées => HMMER

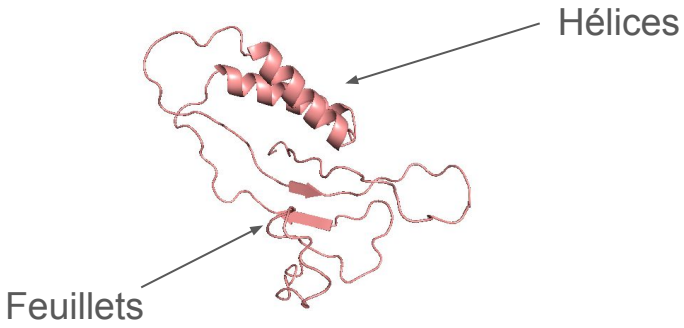
Biais des banques de données

Annotation fonctionnelle des **organismes “non-modèles”** : cela reste un défi !

Pourquoi étudier la structure tridimensionnelle des protéines ?

Une protéine est caractérisé par :

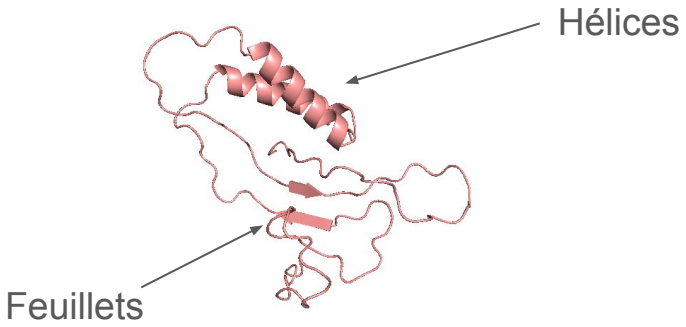
- Une **séquence** : **MKSRKKITPRPPAAQ**
- Une **structure** : appartient à une famille de **repliement**



Pourquoi étudier la structure tridimensionnelle des protéines ?

Une protéine est caractérisé par :

- Une **séquence** : **MKSRKKITPRPPAAQ**
- Une **structure** : appartient à une famille de **repliement**



Intérêt de la structure :

- Permet de **préciser** la fonction biologique
- La **structure est plus conservée** que la séquence
- Recherche d'une fonction aux séquences inconnues en **utilisant le repliement**

(Schaefer & Rost, 2012)

Pourquoi étudier la structure tridimensionnelle des protéines ?

Comment prédire la structure tridimensionnelle des protéines ?

Pourquoi étudier la structure tridimensionnelle des protéines ?

Comment prédire la structure tridimensionnelle des protéines ?

Par l'**expérimentation** (ex : cristallographie) => mais non applicable à grande échelle

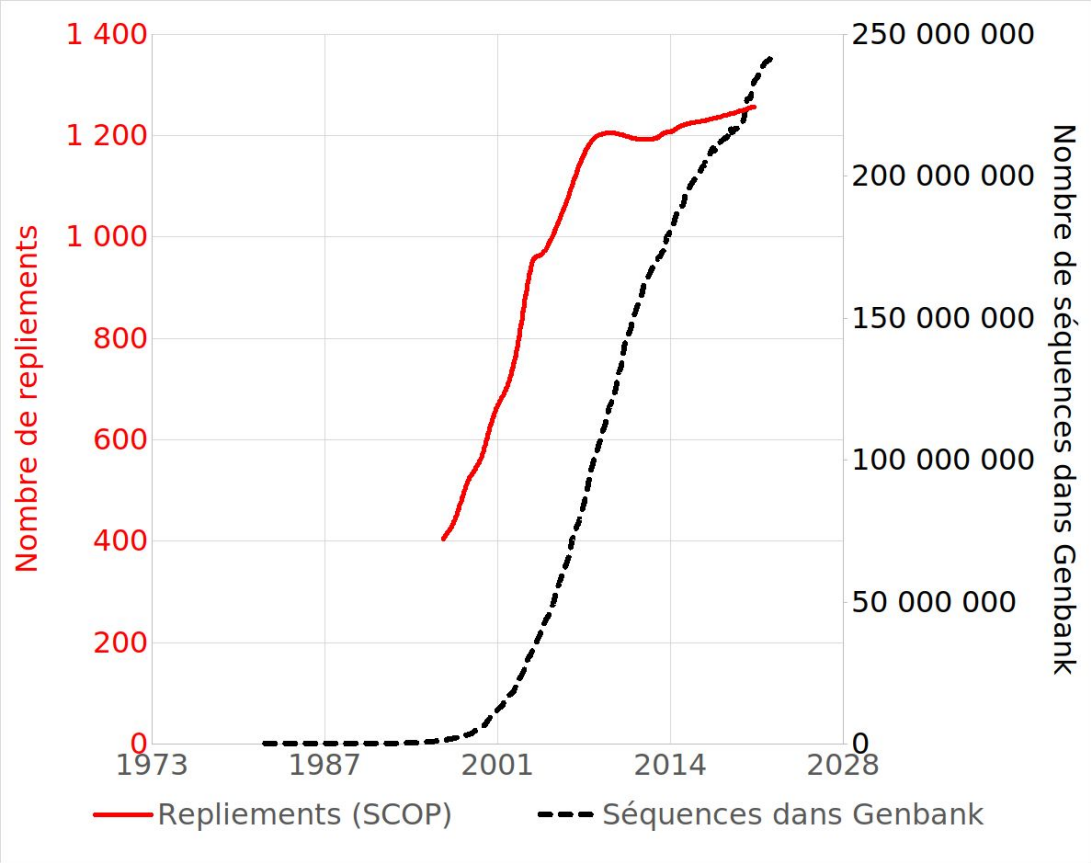
Pourquoi étudier la structure tridimensionnelle des protéines ?

Comment prédire la structure tridimensionnelle des protéines ?

Par l'**expérimentation** (ex : cristallographie) => mais non applicable à grande échelle

Prédire la structure, 2 méthodes :

1) Utilisation des **profils de séquences** => ne permet de trouver de nouveaux repliements
2) Utilisation des **réseaux de neurones profonds** => prédire la structure 3D


1 400
1 200
1 000
800
600
400
200
0
Nombre de repliements
250 000 000
200 000 000
150 000 000
100 000 000
50 000 000
0
Nombre de séquences dans Genbank
1973
1987
2001
2014
2028
Repliements (SCOP)
Séquences dans Genbank

Paradoxe

Augmentation du nombre de séquences disponibles

Peu de découverte de nouveaux repliements

Paradoxe

Augmentation du nombre de séquences disponibles

Peu de découverte de nouveaux repliements

Connaissons-nous uniquement les repliements présents chez les organismes modèles ?

Existe-il des repliements encore inconnus dans les données de métagénomiques ou les organismes non modèles?

Paradoxe

Augmentation du nombre de séquences disponibles

Peu de découverte de nouveaux repliements

Connaissons-nous uniquement les repliements présents chez les organismes modèles ?

Existe-il des repliements encore inconnus dans les données de métagénomiques ou les organismes non modèles?

Innovations structurales : repliement d'une protéine inconnu et spécifique à un taxon

Innovations fonctionnelles : Fonction d'une protéine encore inconnue et spécifique à un taxon

Questions de recherches

Quelle est la proportion de séquences annotées fonctionnellement et structurellement chez les organismes non modèles avec les outils classiques ?

Le deep learning peut-il nous permettre d'explorer l'innovation structurale et contribuer à l'annotation fonctionnelle ?

Ornithocercus quadratus

1) Organismes utilisés :

dinoflagellés en culture

Ornithocercus quadratus

1) Organismes utilisés :

dinoflagellés en culture

- **Rôles écologiques majeurs**
- **Large distribution** océaniques
- Fort **potentiel d'innovation évolutive** fonctionnelle

Ornithocercus quadratus

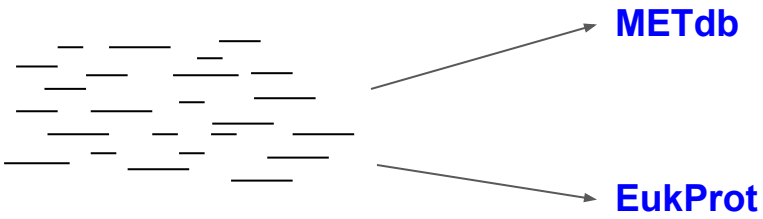
1) Organismes utilisés :

dinoflagellés en culture

- **Rôles écologiques majeurs**
- **Large distribution** océaniques
- Fort **potentiel d'innovation évolutive** fonctionnelle

99 transcriptomes et génomes

≈ 6,7 millions de séquences



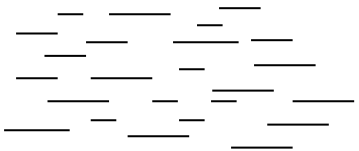
(Niang *et al.*, 2020 ; Richter *et al.*, 2022)

Ornithocercus quadratus

1) Organismes utilisés :

dinoflagellés en culture

≈ 6,7 millions de séquences



2) Annotations structurales et fonctionnelles

- Outil : **InterProScan**

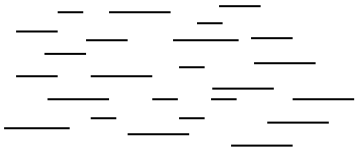
(Jones *et al.*, 2014)

Ornithocercus quadratus

1) Organismes utilisés :

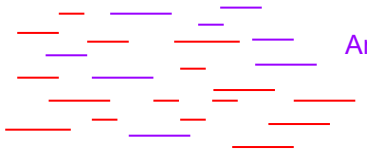
dinoflagellés en culture

≈ 6,7 millions de séquences



2) Annotations structurales et fonctionnelles

- Outil : **InterProScan**



Annotations : ~ 47%

Pas d'annotations : ~ 53%

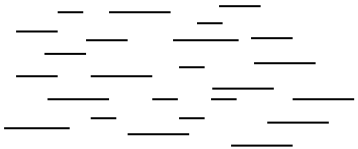
(Jones *et al.*, 2014)

Ornithocercus quadratus

1) Organismes utilisés :

dinoflagellés en culture

≈ 6,7 millions de séquences



2) Annotations structurales et fonctionnelles

- Outil : **InterProScan**

Annotations : ~ 47%

Pas d'annotations : ~ 53%

3) Construction des familles protéiques

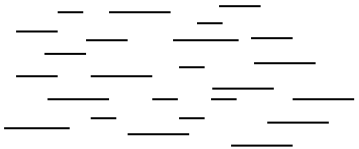
- Méthodes : **graphes**
- Regrouper les séquences selon la **similarité**

Ornithocercus quadratus

1) Organismes utilisés :

dinoflagellés en culture

≈ 6,7 millions de séquences



2) Annotations structurales et fonctionnelles

- Outil : **InterProScan**

Annotations : ~ 47%

Pas d'annotations : ~ 53%

3) Construction des familles protéiques

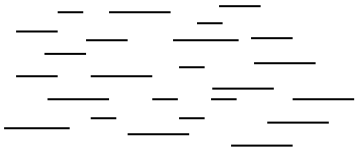
- Méthodes : **graphes**
- Regrouper les séquences selon la **similarité**
- **Choix des paramètres** :
  - pourcentage d'identité : **60%**
  - pourcentage d'overlap : **70%**
  - e-value : **$1e^{-19}$**

Ornithocercus quadratus

1) Organismes utilisés :

dinoflagellés en culture

≈ 6,7 millions de séquences



2) Annotations structurales et fonctionnelles

- Outil : **InterProScan**

Annotations : ~ 47%

Pas d'annotations : ~ 53%

3) Construction des familles protéiques

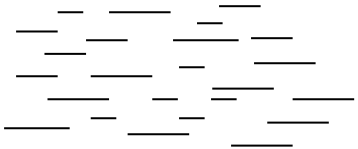
- Méthodes : **graphes**
- Regrouper les séquences selon la **similarité**
- **Choix des paramètres** :
  - pourcentage d'identité : **60%**
  - pourcentage d'overlap : **70%**
  - e-value : **$1e^{-19}$**
- **Objectif** :
  - Maximiser les gros clusters
  - Minimiser les petits clusters

Ornithocercus quadratus

1) Organismes utilisés :

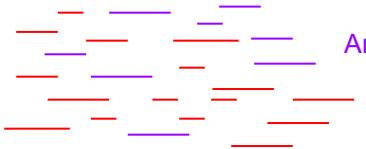
dinoflagellés en culture

≈ 6,7 millions de séquences



2) Annotations structurales et fonctionnelles

- Outil : **InterProScan**

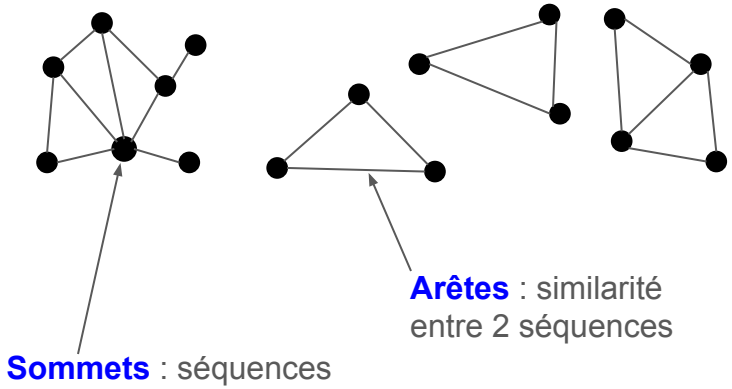


Annotations : ~ 47%

Pas d'annotations : ~ 53%

3) Construction des familles protéiques

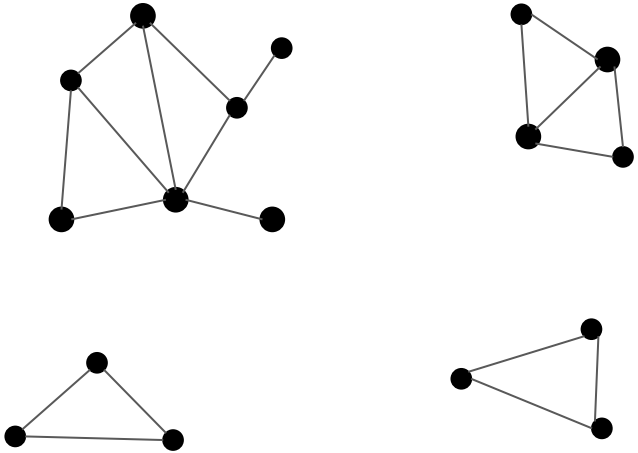
- Méthodes : **graphes**
- Regrouper les séquences selon la **similarité**
- **Clusters : ≈ 220 000** (≈ 2.9 millions de séquences)



Arêtes : similarité entre 2 séquences

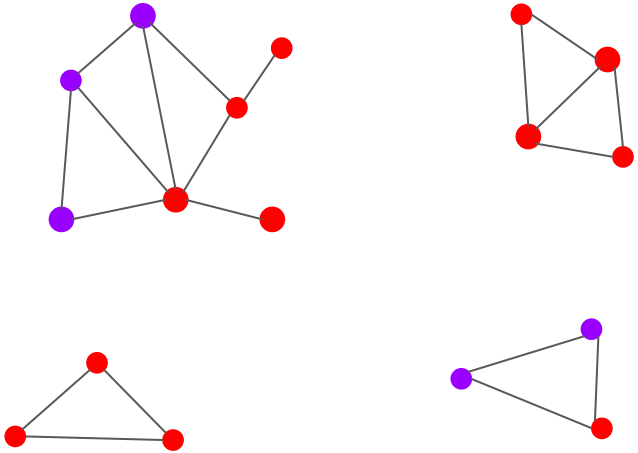
Sommets : séquences

4) Ajout des informations sur les noeuds :



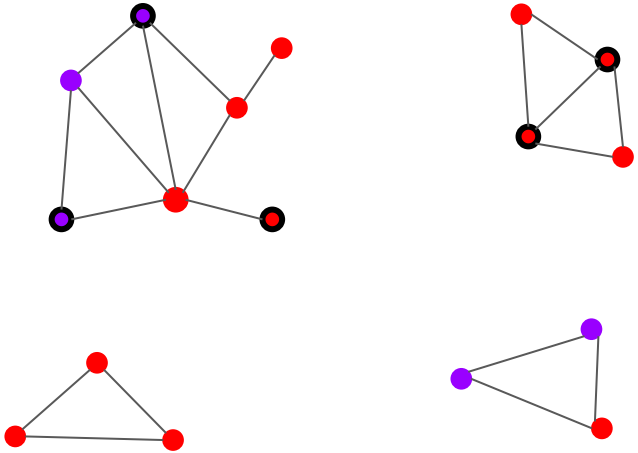
4) Ajout des informations sur les noeuds :

- **Annotations fonctionnelles**, méthode classique



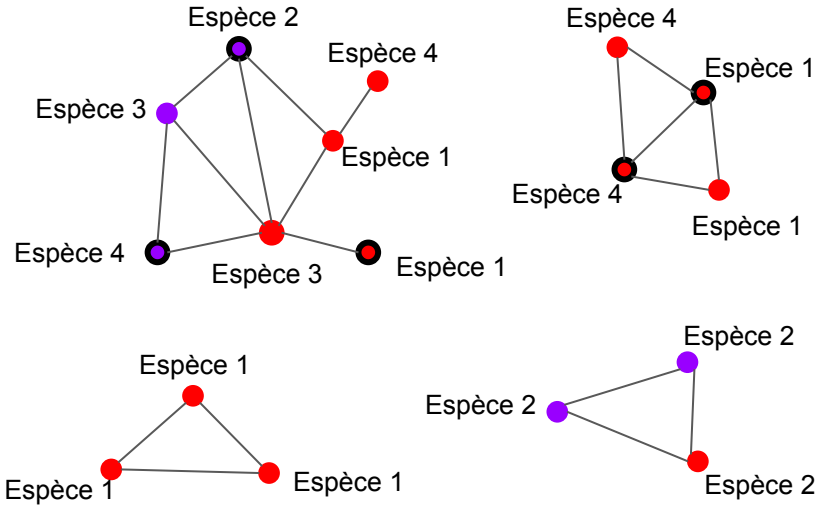
4) Ajout des informations sur les noeuds :

- **Annotations fonctionnelles**, méthode classique
- **Annotations structurales**, méthode classique : ○



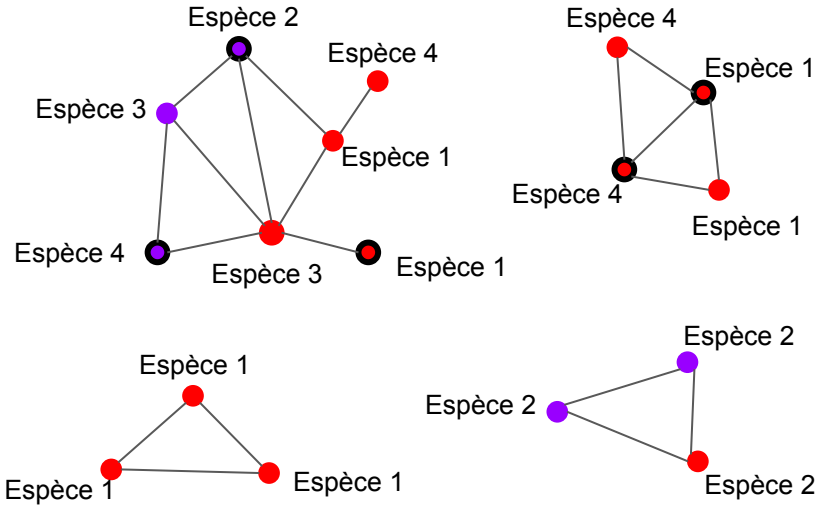
4) Ajout des informations sur les noeuds :

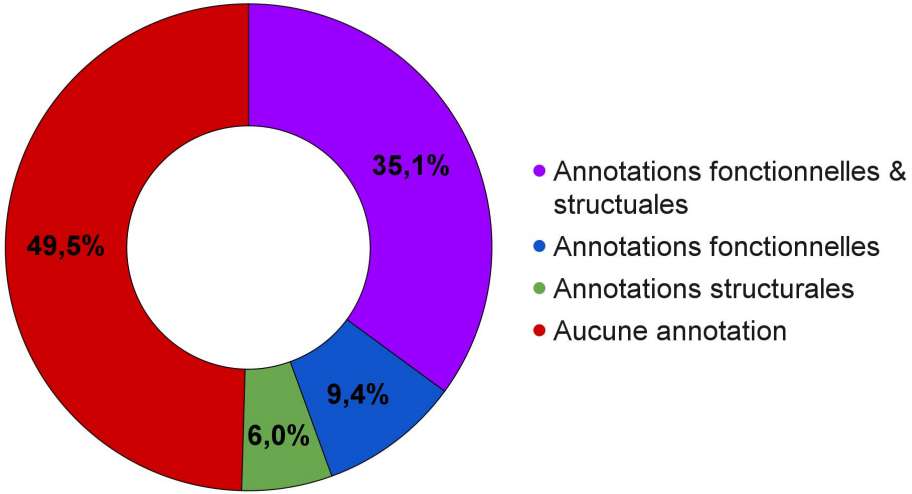
- **Annotations fonctionnelles**, méthode classique
- **Annotations structurales**, méthode classique : ⭘
- **Taxonomie** : *espèce 1, espèce 2, …*



4) Ajout des informations sur les noeuds :

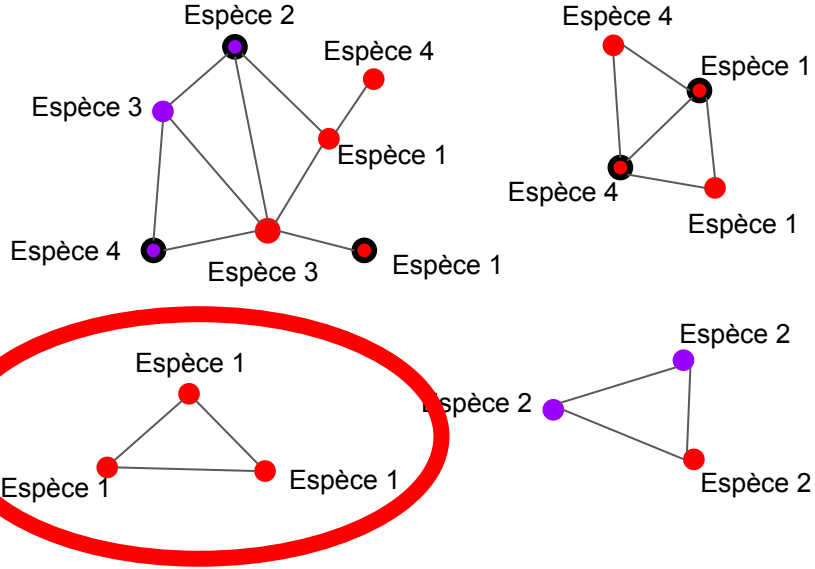
- **Annotations fonctionnelles**, méthode classique
- **Annotations structurales**, méthode classique : ○
- **Taxonomie** : *espèce 1, espèce 2, …*





4) Ajout des informations sur les noeuds :

- **Annotations fonctionnelles**, méthode classique
- **Annotations structurales**, méthode classique : O
- **Taxonomie** : *espèce 1, espèce 2, …*



⅓ des clusters sont spécifiques à une espèce et inconnus fonctionnellement et structurellement

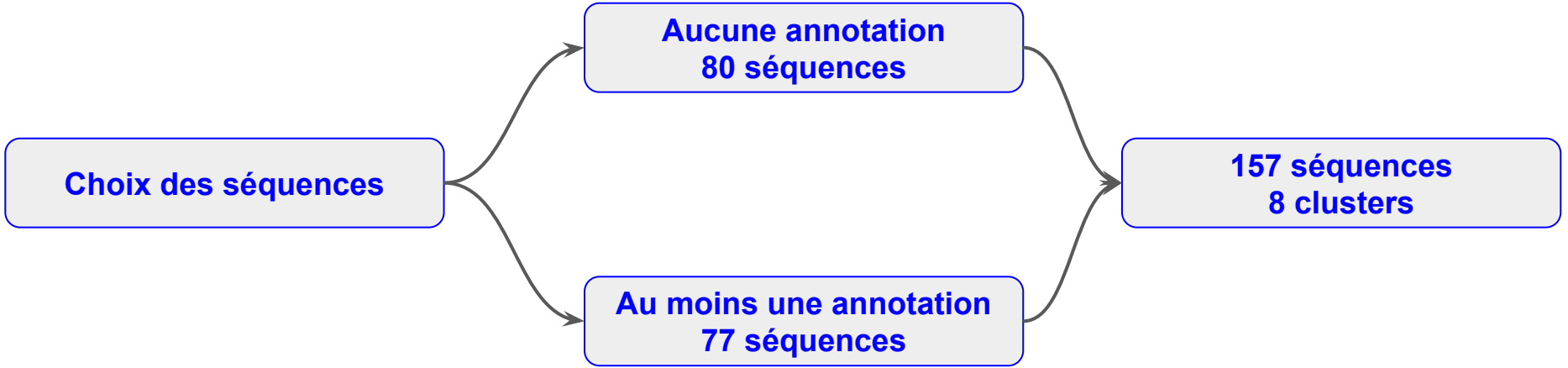
5) Prédiction structurale - AlphaFold2 :

Objectif : tester de la faisabilité de la méthode

Choix des séquences

5) Prédiction structurale - AlphaFold2 :

Objectif : tester de la faisabilité de la méthode



(Jumper *et al.*, 2021)

5) Prédiction structurale - AlphaFold2 :

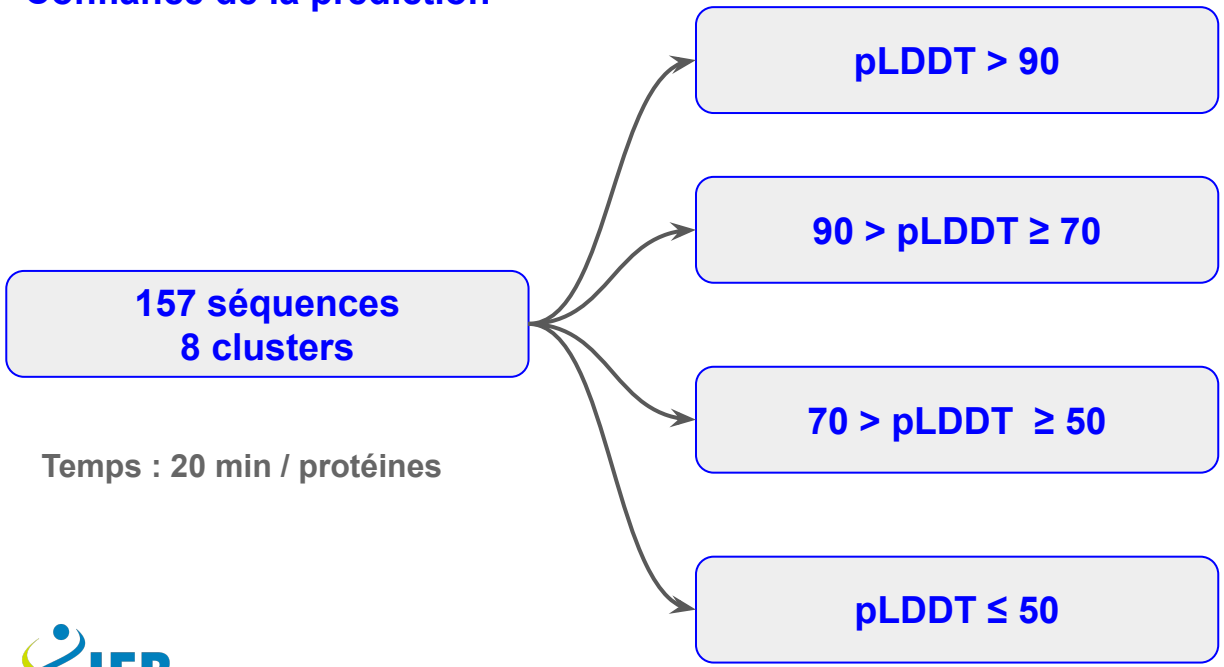
Confiance de la prédiction

157 séquences
8 clusters

Temps : 20 min / protéines

5) Prédiction structurale - AlphaFold2 :

Confiance de la prédiction

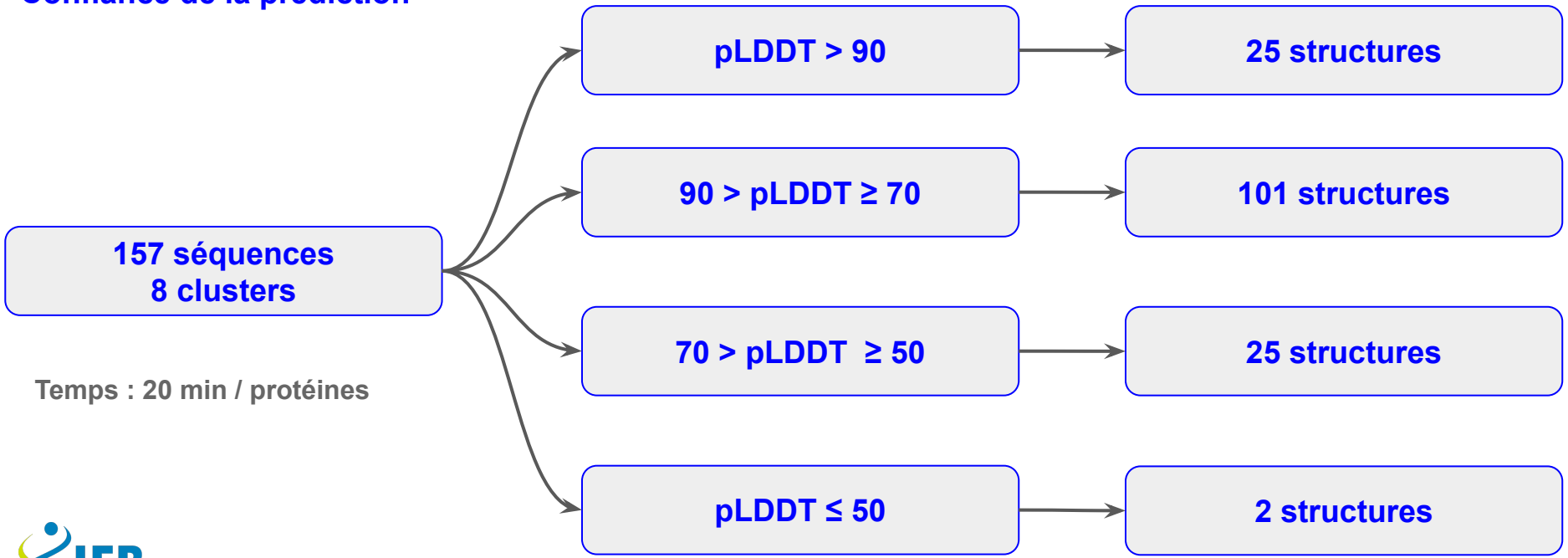


Temps : 20 min / protéines

(Jumper *et al.*, 2021)

5) Prédiction structurale - AlphaFold2 :

Confiance de la prédiction

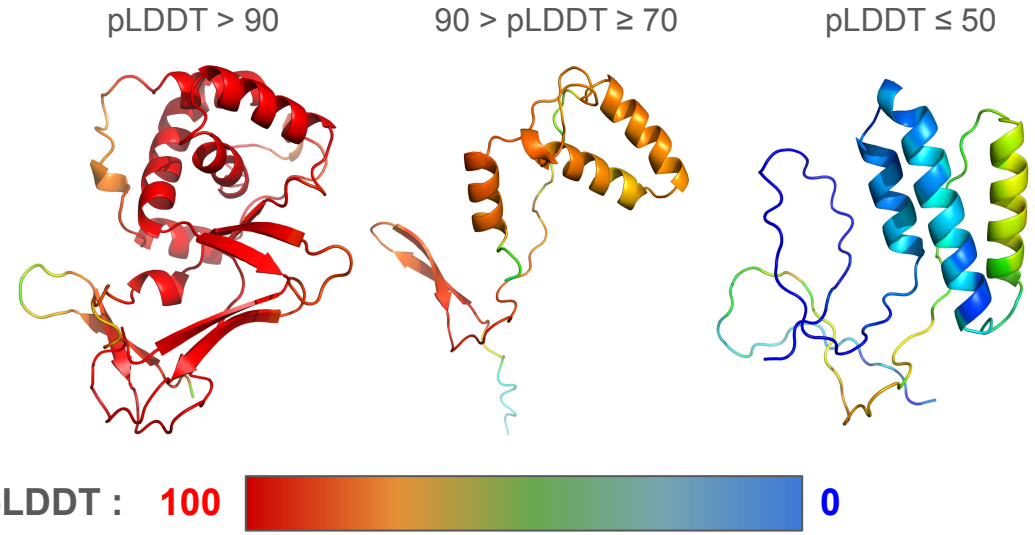


Temps : 20 min / protéines

5) Prédiction structurale - AlphaFold2 :

Confiance de la prédiction

La taille des séquences influence la qualité de la prédiction



(Jumper *et al.*, 2021)

5) Prédiction structurale - AlphaFold2 :

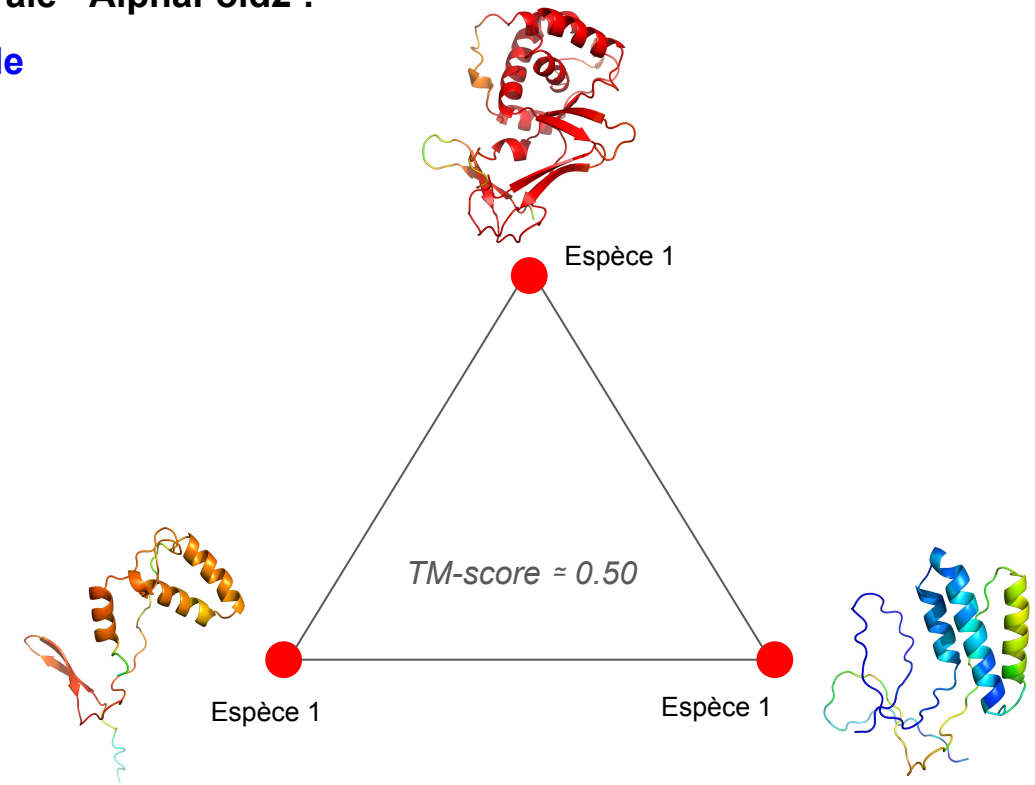
Cohérence structurale

Vérification de l'homogénéité des structures

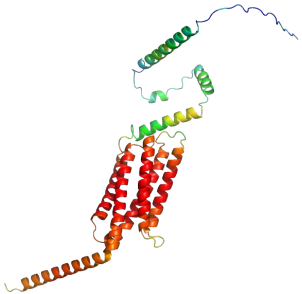
TM-score : permet de mesurer la similitude des structures prédites

5) Prédiction structurale - AlphaFold2 :

Cohérence structurale

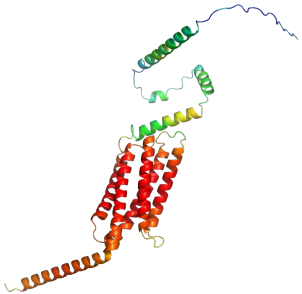


6) Recherche d’une fonction pour les repliements



157 structures

6) Recherche d'une fonction pour les repliements



157 structures

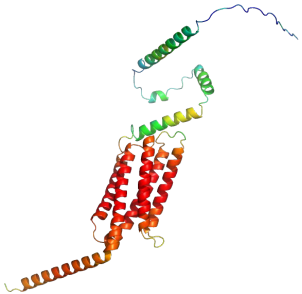
Foldseek

RCSB PDB
PROTEIN DATA BANK

150 000 structures

(Berman *et al.*, 2000 ; van Kempen *et al.*, 2023)

6) Recherche d'une fonction pour les repliements



157 structures

Foldseek

RCSB PDB
PROTEIN DATA BANK

150 000 structures

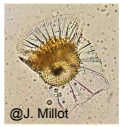
Similarité : 42%

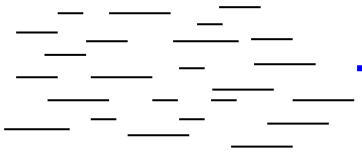
Pas de similarité : 58%

Nouveaux repliements ?

Nouvelles fonctions ?

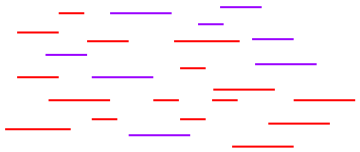
(Berman *et al.*, 2000 ; van Kempen *et al.*, 2023)

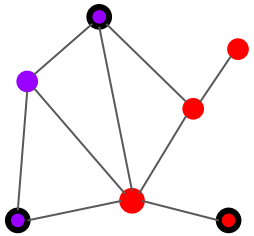




ABiMS

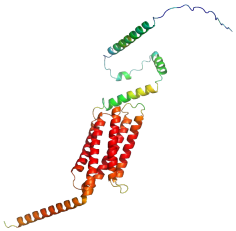
InterProScan





IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

AlphaFold2



RCSB PDB
PROTEIN DATA BANK

Repliement connu

Repliement inconnu

Limites

- **Qualité des prédiction d'AlphaFold**
- **Utilisation des annotations Pfam / CATH => utiliser la Gene Ontology**
- **Temps de calcul (20 minutes / protéine)**
- **Utilisation d'un seul GPU**

- Étude des **séquences non retenues** pour la construction du graphe
- Étude de la **composantes connexes** avec plus de **1 millions de séquences**
- Étude de l'**évolution des structures**
- Étude de la **biogéographie des structures**
- Utilisation d'un **supercalculateur** (ex. Jean Zay)

- Étude des **séquences non retenues** pour la construction du graphe
- Étude de la **composantes connexes** avec plus de **1 millions de séquences**
- Étude de l'**évolution des structures**
- Étude de la **biogéographie des structures**
- Utilisation d'un **supercalculateur** (ex. Jean Zay)

Les perspectives seront étudiées lors d'une thèse à l'Atelier de Bio-Informatique

Institut des sciences du calcul et des données
ALLIANCE SORBONNE UNIVERSITÉ

MUSÉUM NAL HIST NATURELLE

ISYEB
Institut de Systématique, Evolution, Biodiversité

ABI

PEPR ATLASea 2023-2030

Projet FORMAL 2023-2027

Institut des sciences du calcul et des données
ALLIANCE SORBONNE UNIVERSITÉ

Les perspectives seront étudiées lors d'une thèse à l'Atelier de Bio-Informatique

Institut des sciences du calcul et des données
ALLIANCE SORBONNE UNIVERSITÉ

Merci pour votre attention !

- Berman, H. M. et al. The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000).
- Jones, P. et al. InterProScan 5: genome-scale protein function classification. Bioinforma. Oxf. Engl. 30, 1236–1240 (2014).
- Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).
- Niang, G. et al. METdb: A GENOMIC REFERENCE DATABASE FOR MARINE SPECIES. F1000Research 9, (2020).
- Richter, D. J. et al. EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. Peer Community J. 2, (2022).
- Schaefer, C. & Rost, B. Predict impact of single amino acid change upon protein structure. BMC Genomics 13, S4 (2012).
- van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. Nat. Biotechnol. 1–4 (2023).