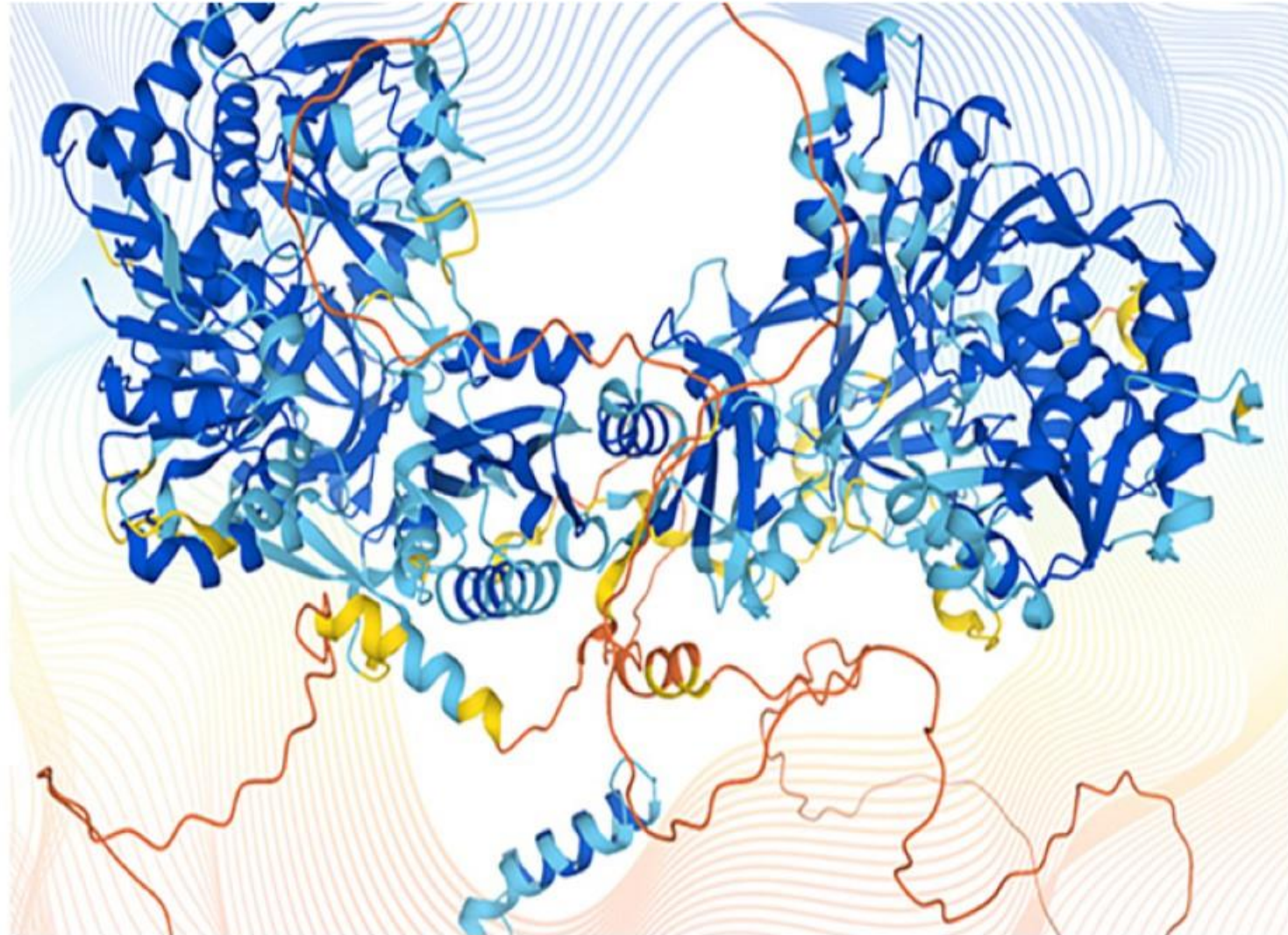


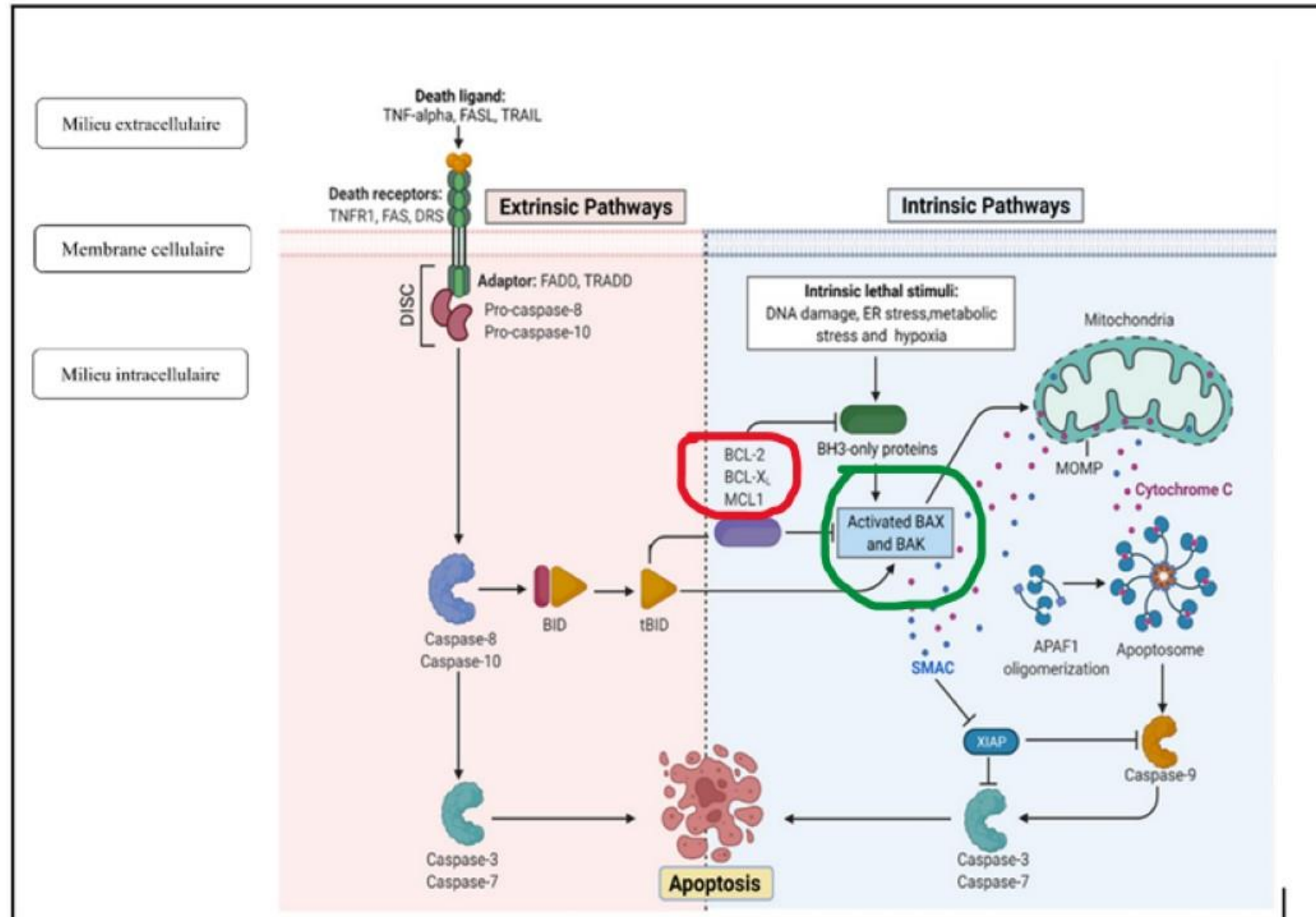
# Identification de Protéines basée sur des méthodes d'Homologie Persistante (IPHP)

Tidjani CISSE  
Adam BOUSSIF  
29 mars 2024



# Introduction

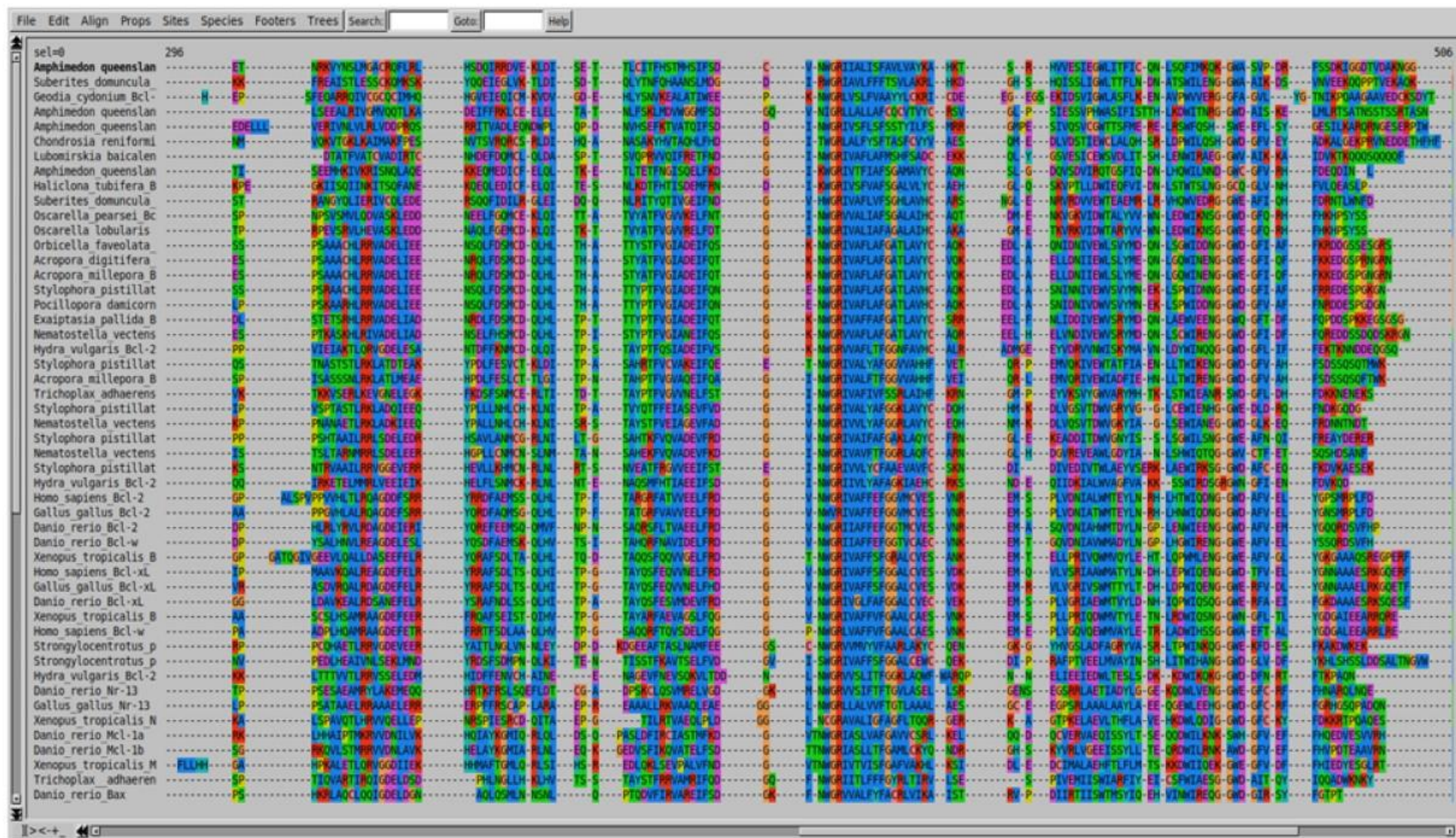
- Famille des protéines BCL2.
- Deux voies déclenchant l'apoptose:
  - - voie intrinsèque ou voie mitochondriale.
  - - voie extrinsèque ou voie des récepteurs.





- Identification difficile des homologues de bcl par BLAST ou par profil HMM.
- Ex: alignement des 111 séquences de référence de BCL2 avec le logiciel MAFFT.

- **Ex:** alignement des 111 séquences de référence de BCL2 avec le logiciel MAFFT.





# Sommaire

- Pipeline de l'analyse par homologie persistante.
- Présentation de l'analyse des jeux de données (référence et cibles).
- Homologie persistance et le principe de la filtration.
- Analyse et interprétations des résultats obtenus.
- Conclusions.

# Pipeline de l'analyse par homologie persistante.

- Prédiction des structures 3D des protéines BCL2 par alphafold.
- Identification des structures à haute confiance basées sur le pLDDT des (C $\alpha$ ) .
- Extraction des coordonnées des (C $\alpha$ ) dans des fichiers PDB .
- Calcul de barres codes et de distances de Wasserstein à l'aide de la bibliothèque Python GUDHI .

# Les jeux de données (référence et cibles).

- 111 BCL validé et caractérisé (références).

Structure	Nb.Ca.initiaux	Nb.Ca.enlevés	pLDDT.moyen.initial	pLDDT.moyen.final
1 Halisarca_caerulea_CRCL_1	216	78	73.99	85.52
2 Gallus_gallus_CRCL_5	177	37	83.42	90.42
3 Gallus_gallus_CRCL_6	216	35	83.37	88.32
4 Amphimedon_queenslandica_CRCL_7	229	69	78.93	90.33
5 Strongylocentrotus_purpuratus_CRCL_9	531	409	51.39	84.51
6 Halisarca_caerulea_CRCL_2	189	50	78.58	85.98
7 Xenopus_tropicalis_CRCL_5	283	144	68.38	91.76
8 Strongylocentrotus_purpuratus_CRCL_7	216	62	78.95	86.96
9 Haliclona_tubifera_CRCL_1	219	63	76.39	86.51
10 Stylophora_pistillata_CRCL_6	194	45	83.05	90.69
11 Hydra_vulgaris_CRCL_3	230	66	80.95	92.66
12 Homo_sapiens_CRCL_8	175	30	86.27	92.05
13 Danio_rerio_CRCL_8	176	31	87.06	93.22
14 Strongylocentrotus_purpuratus_CRCL_3	255	114	71.82	91.64
15 Oscarella_pearsei_CRCL_1	207	47	81.37	89.43
16 Lubomirskia_baicalensis_CRCL_2	244	138	60.87	82.10
17 Haliclona_tubifera_CRCL_3	239	75	76.79	85.93
18 Gallus_gallus_CRCL_7	174	0	89.68	89.68
19 Strongylocentrotus_purpuratus_CRCL_1	273	129	72.53	89.31
20 Geodia_cydonium_CRCL_2	255	100	71.86	88.99
21 Pocillopora_damicornis_CRCL_1	184	0	86.44	86.44



# Jeux de données (suite)

- Protéome d'autres organismes (cibles) :
- Homo sapiens, drosophile, poisson zèbre, c. elegans.
  - - Résolution 2,5 Å°
  - - Taille comprise 150 – 600 résidus

6K80_ca.pdb	6T2T_ca.pdb	6XYS_ca.pdb	7BJS_ca.pdb	7DAX_ca.pdb	7ME4_ca.pdb	7QUT_ca.pdb	8C1U_ca.pdb	8DKT_ca.pdb
6KCP_ca.pdb	6TBN_ca.pdb	6Y20_ca.pdb	7BTG_ca.pdb	7DAY_ca.pdb	7ME5_ca.pdb	7RL3_ca.pdb	8C69_ca.pdb	8OIJ_ca.pdb
6KD6_ca.pdb	6VJ8_ca.pdb	6Y6E_ca.pdb	7CFB_ca.pdb	7DAZ_ca.pdb	7MFW_ca.pdb	7UVE_ca.pdb	8C6A_ca.pdb	
6LPH_ca.pdb	6VJ9_ca.pdb	6YRB_ca.pdb	7CIU_ca.pdb	7DB0_ca.pdb	7NYQ_ca.pdb	7VNH_ca.pdb	8C6B_ca.pdb	
6Q9M_ca.pdb	6XG7_ca.pdb	6YRQ_ca.pdb	7CIV_ca.pdb	7DB1_ca.pdb	7OC3_ca.pdb	7VNI_ca.pdb	8C6C_ca.pdb	
6QQ4_ca.pdb	6XRO_ca.pdb	6ZMU_ca.pdb	7CIW_ca.pdb	7DB2_ca.pdb	7PO9_ca.pdb	7ZHH_ca.pdb	8C6F_ca.pdb	
6S7X_ca.pdb	6XRP_ca.pdb	7ALT_ca.pdb	7CIX_ca.pdb	7DB3_ca.pdb	7POK_ca.pdb	8A8Q_ca.pdb	8C6H_ca.pdb	
6S7Y_ca.pdb	6XTZ_ca.pdb	7AYV_ca.pdb	7CQV_ca.pdb	7DB4_ca.pdb	7PV5_ca.pdb	8AK2_ca.pdb	8D7P_ca.pdb	
6S9F_ca.pdb	6XWV_ca.pdb	7AZT_ca.pdb	7CUY_ca.pdb	7JW6_ca.pdb	7Q4I_ca.pdb	8AK3_ca.pdb	8DHU_ca.pdb	

# Objectifs

- Étudier la variabilité des caractéristiques topologiques des protéines BCL2 en utilisant l'homologie persistante.
- - Comparaison des structures prédites des BCL2 entre elles.
- - Calcul des distances homologiques.
- Comparaison des structures des protéines des protéomes des organismes d'intérêt avec les structures des protéines BCL2 de référence.



# Objectifs(suite)

- Si les distances sont inférieures aux distances calculées entre les protéines BCL2 de référence, ce sont des candidats potentiels
- Si les distances sont supérieures alors ce ne sont pas des BCL2

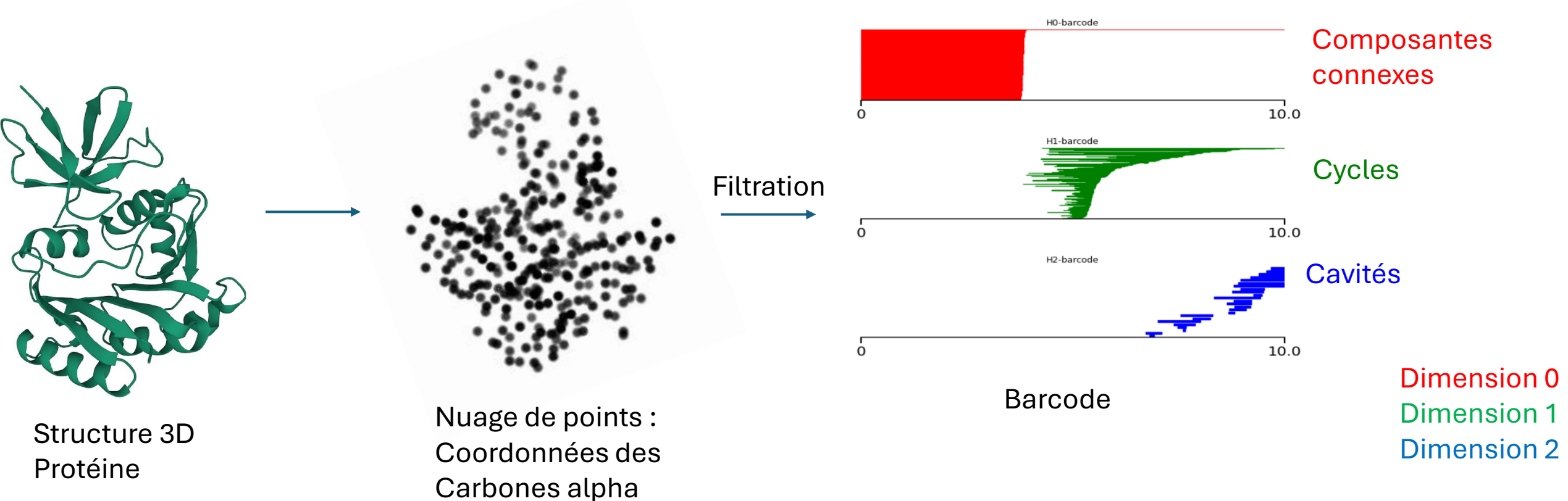
# Analyse topologique des données par Homologie Persistante :



# Analyse topologique :

## Homologie persistante :

- Homologie Persistante : Cette méthode est utilisée pour reconnaître et comparer **les formes dans les données**.
- Approche : Analyser **la naissance et la mort des caractéristiques topologiques** au fil des échelles.
- Représentation : **Les barcodes** traduisent **la persistance des caractéristiques** au fil des échelles.

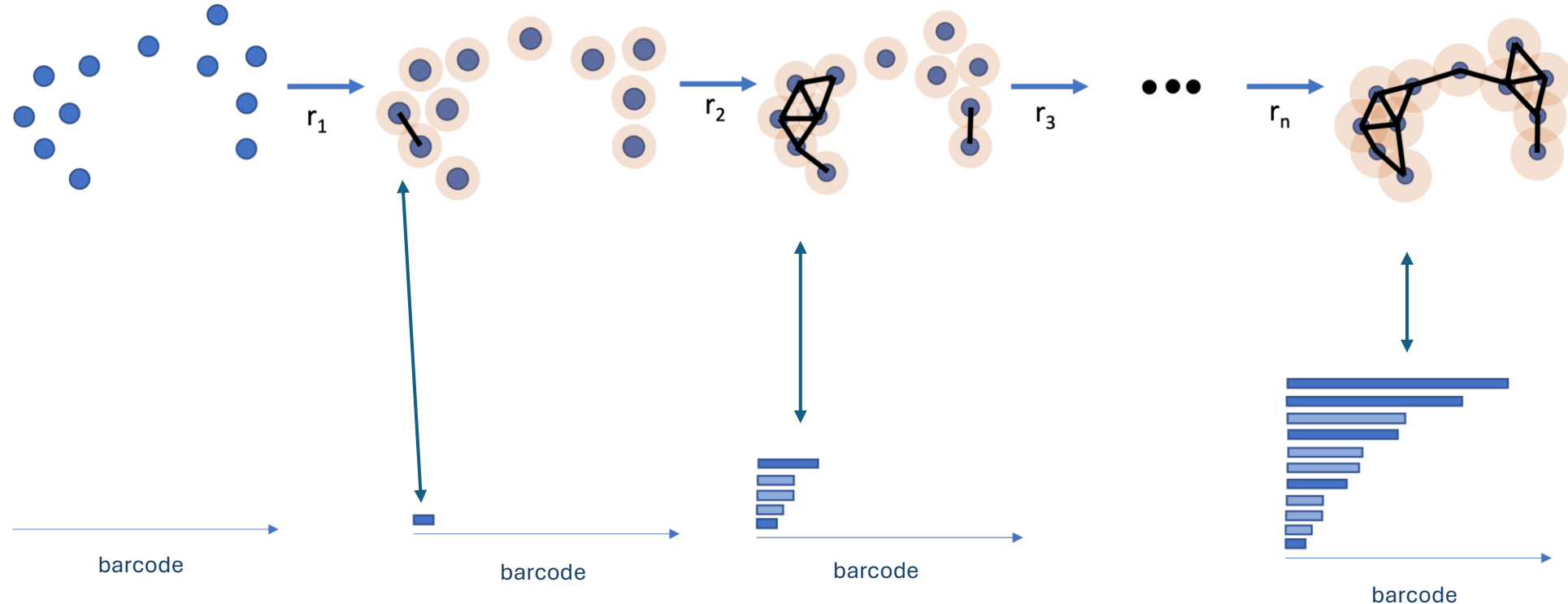


# Homologie persistante :

## Filtration : Alpha et Rips Complex

Principe : Construction progressive de simplexes à partir de données.

Processus : **Augmentation graduelle d'un paramètre ( $r$ ) pour capturer la structure topologique dans un barcode.**



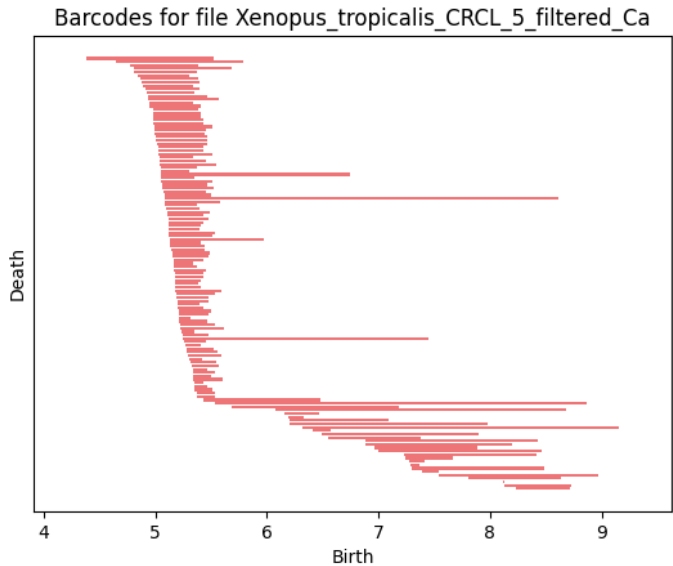


# Analyse topologique :

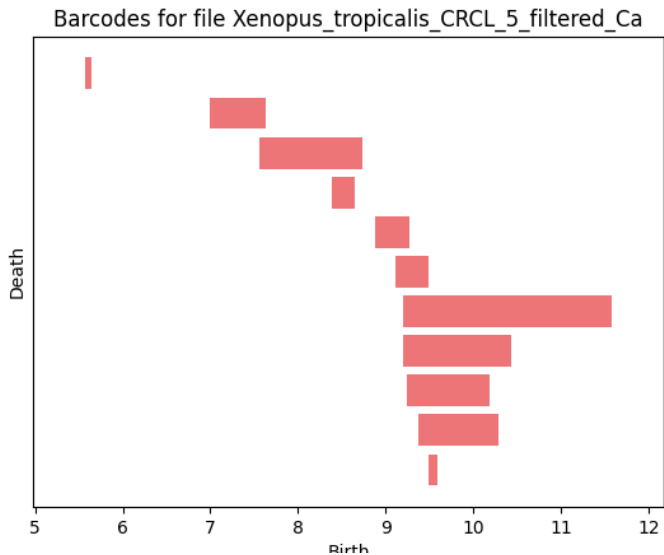
Homologie persistante :

Barcodes Rips complexe :

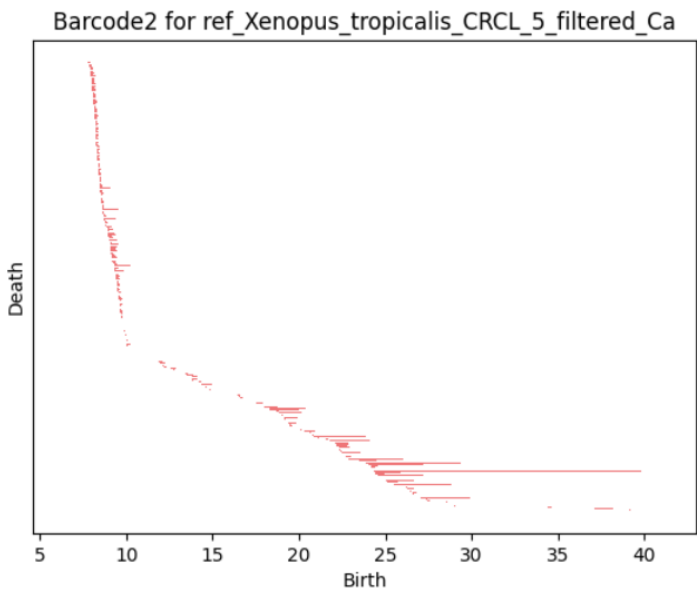
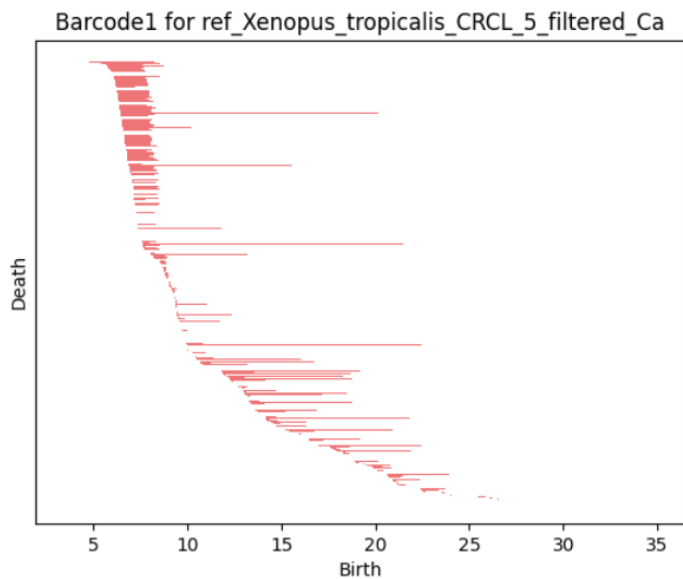
Dimension 1 :



Dimension 2 :



Barcodes Alpha complexe :

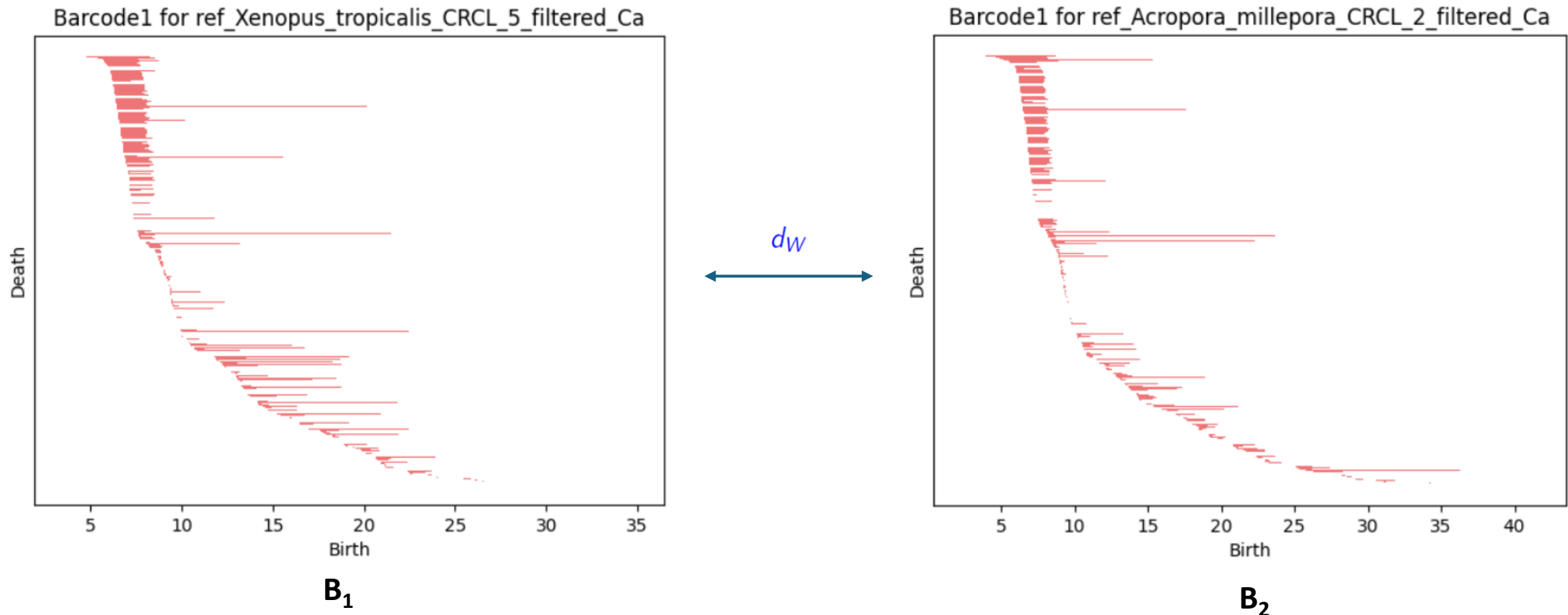


# Analyse topologique :

Homologie persistante :

Distance de Wasserstein : Comparer deux barcodes  $B_1$  et  $B_2$  en calculant la distance entre eux.

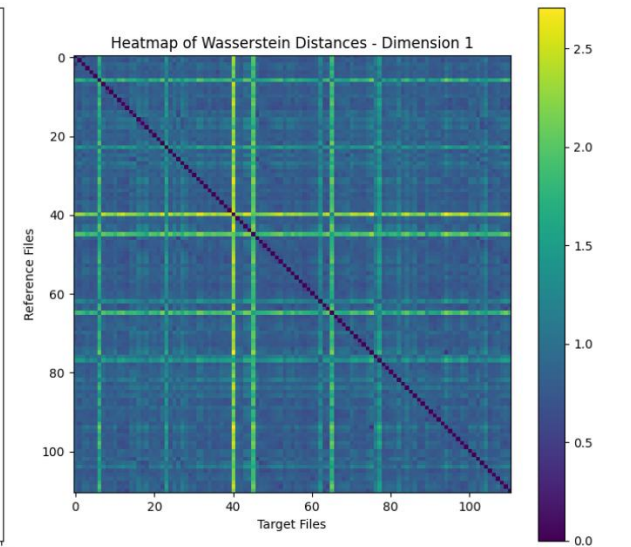
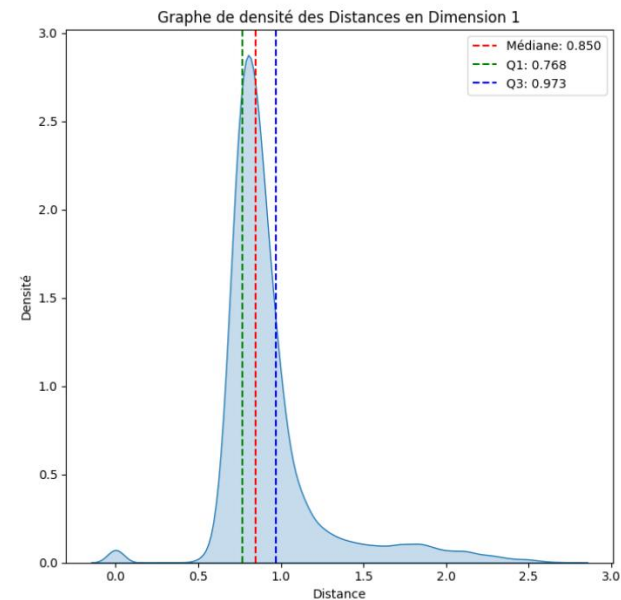
Objectif : **Quantifier le travail pour transformer  $B_1$  en  $B_2$  où 0 équivaut à aucune différence.**



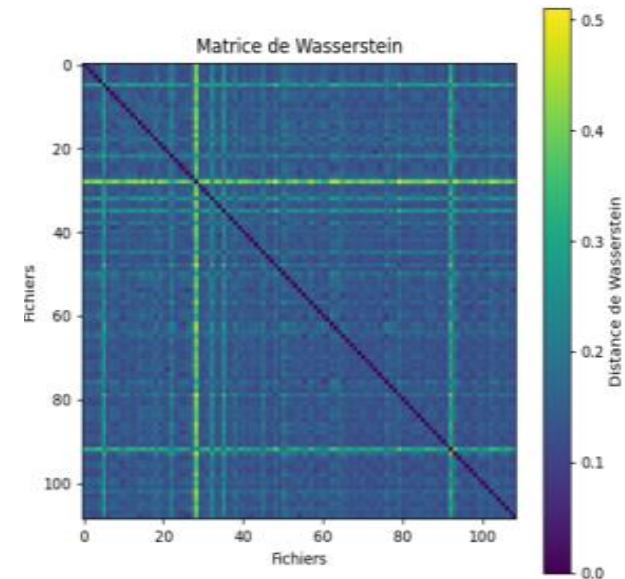
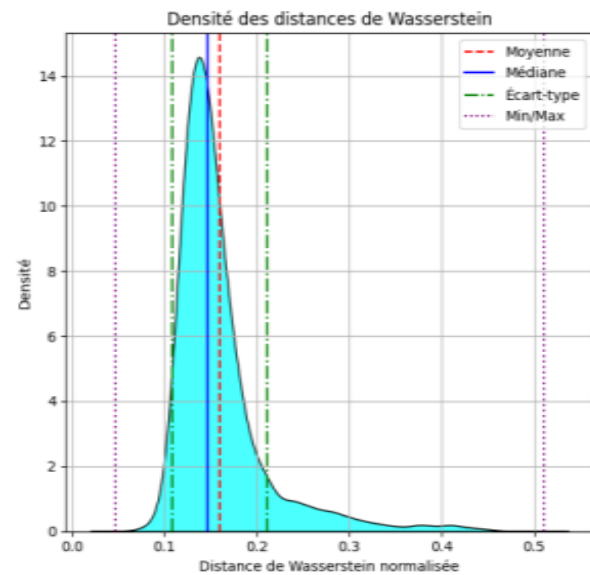


# Homologie persistante : en dimension 1

Matrice Référence (BCL2) contre Référence :  
Alpha complexe

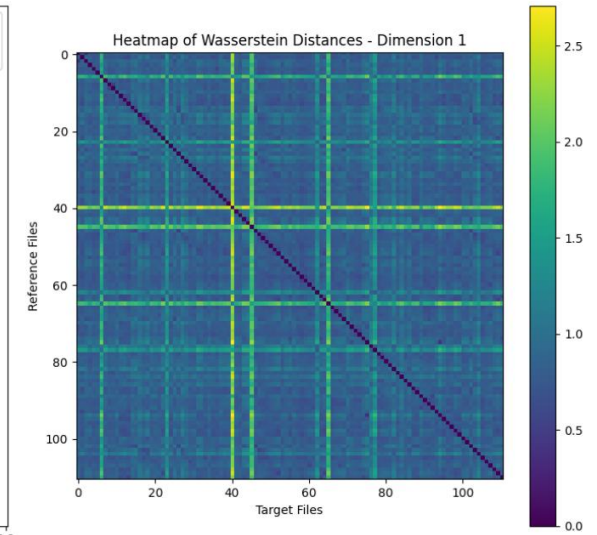
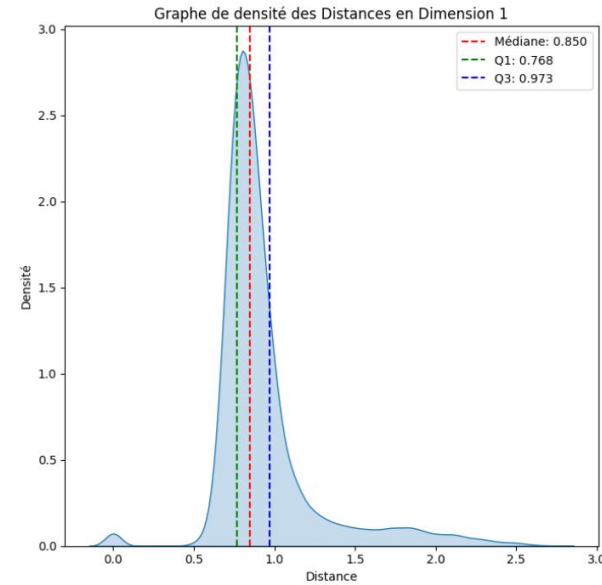


Matrice Référence (BCL2) contre Référence :  
Rips complexe

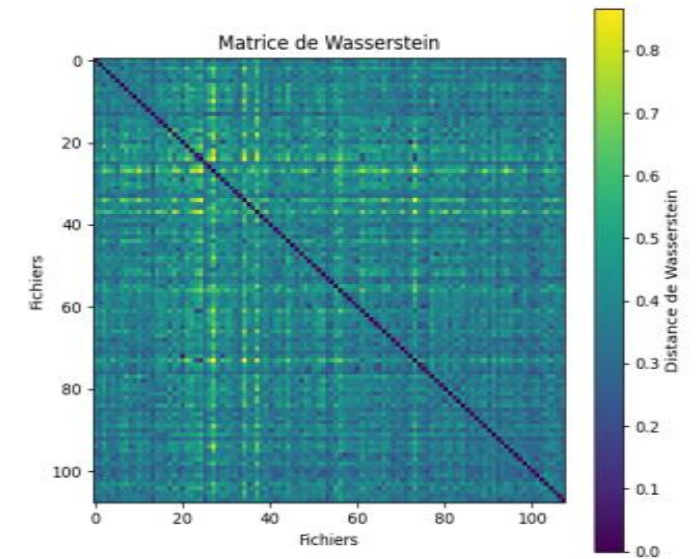
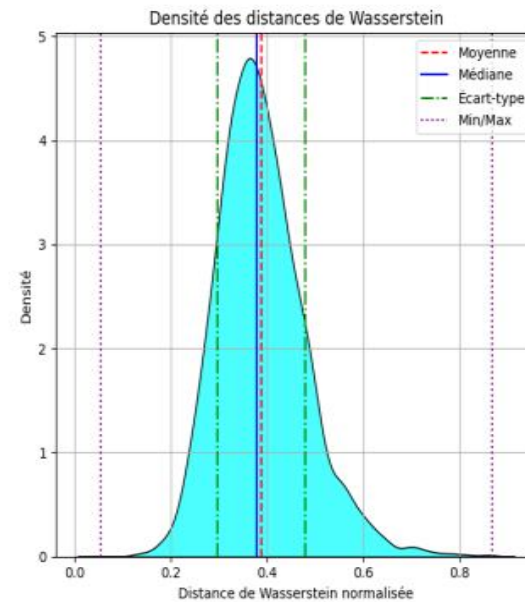


# Homologie persistante : en dimension 2

Matrice Référence (BCL2) contre Référence :  
Alpha complexe



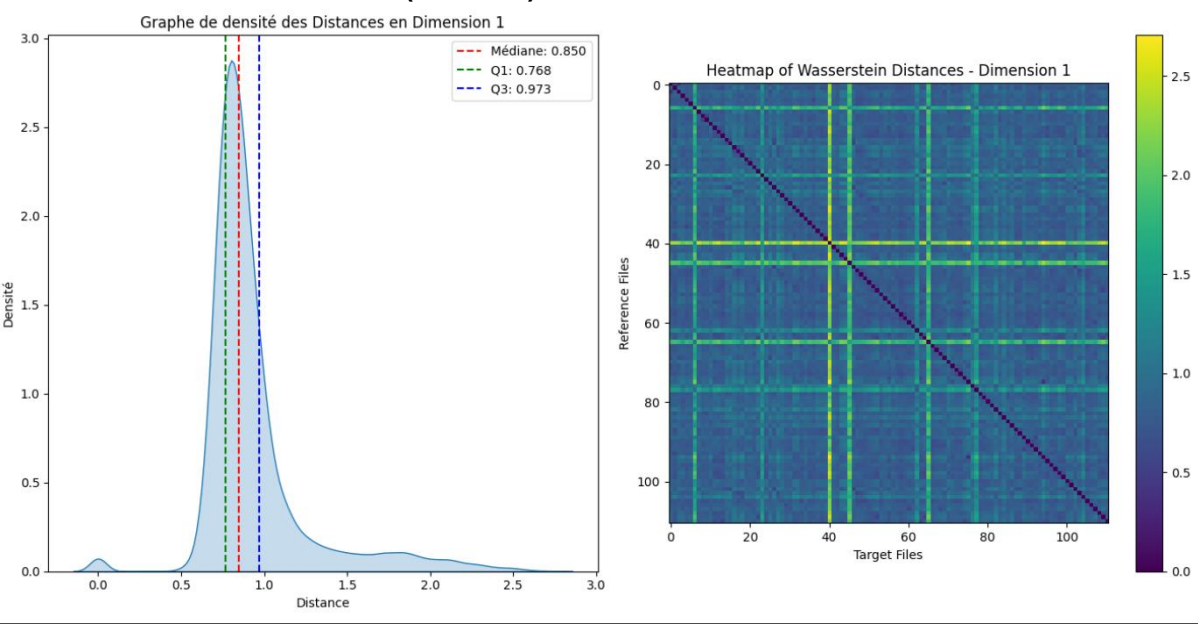
Matrice Référence (BCL2) contre Référence :  
Rips complexe



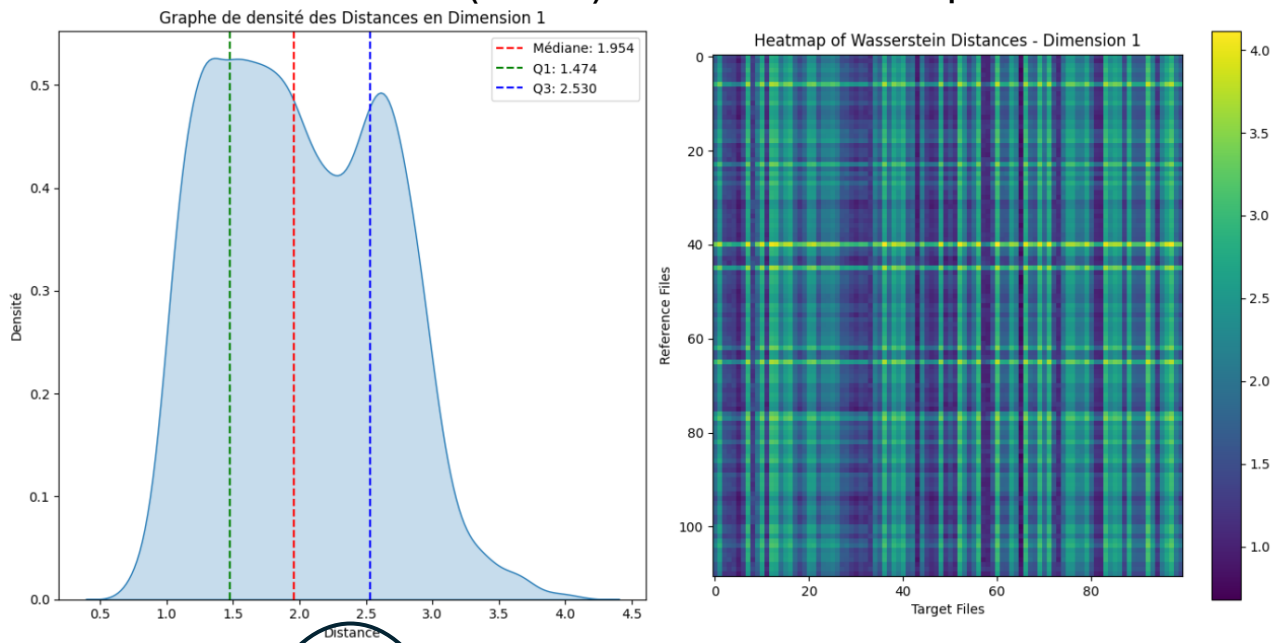


# Homologie persistante en dimension 1

Matrice Référence (BCL2) contre Référence :



Matrice Référence (BCL2) contre Homo sapiens :

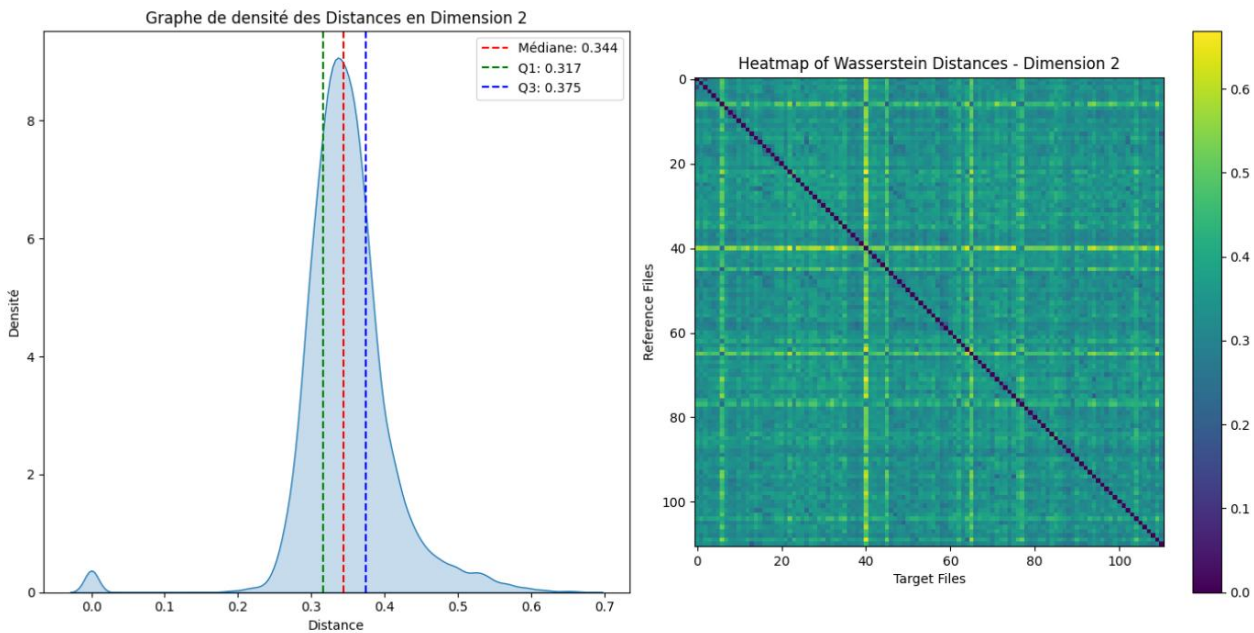


Moyenne des médianes = 1.733

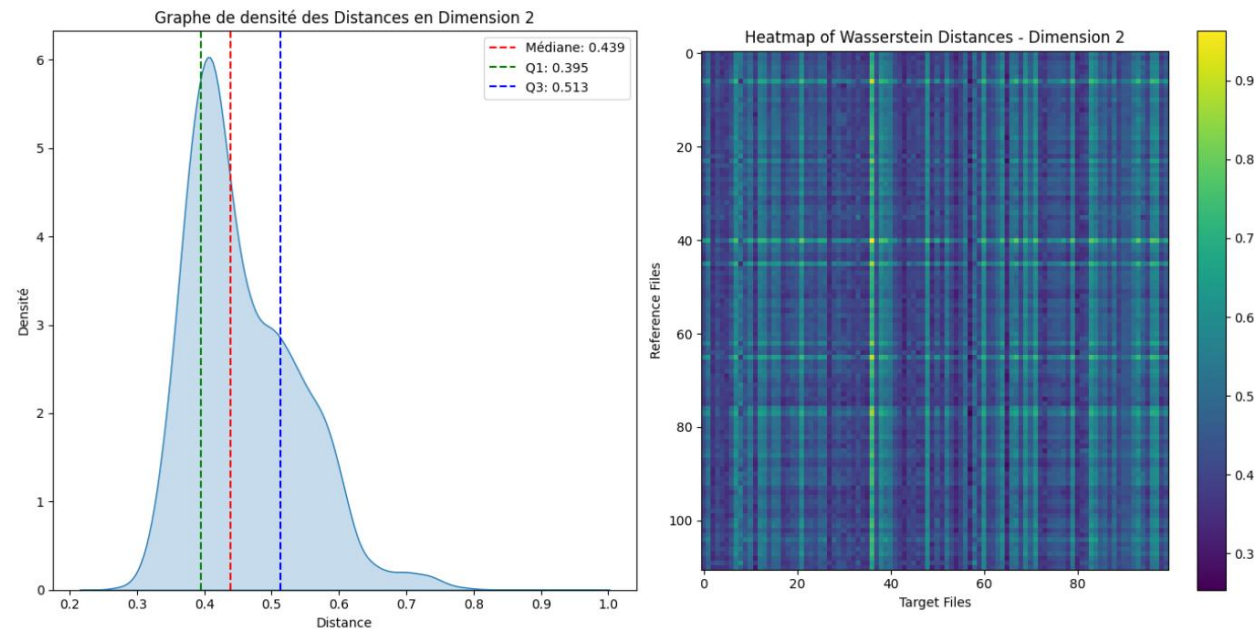
Espèces	Dimension	Médiane	Q1	Q3
<u>Référence</u>	1	<u>0.850</u>	0.768	0.973
Homo sapiens	1	1.95	1.474	2.530
Drosophila (74)	1	2.05	1.296	2.595
C. elegans (10)	1	2.46	2.283	2.651
<u>Danio rerio (52)</u>	1	1.61	1.312	2.239
All (347)	1	1.48	0.964	2.280

# Homologie persistante en dimension 2

Matrice Référence (BCL2) contre Référence :



Matrice Référence (BCL2) contre Homo sapiens :



Moyenne des médianes : 0.435

Espèces	Dimension	Médiane	Q1	Q3
Référence	2	0.344	0.317	0.375
Homo sapiens	2	0.439	0.395	0.513
Drosophile (74)	2	0.449	0.378	0.497
C. elegans (10)	2	0.487	0.458	0.526
Danio rerio (52)	2	0.443	0.393	0.499
All (347)	2	0.446	0.353	0.480

# Analyse topologique :

## Homologie persistante :

Exploration des résultats à partir des distances Référence contre All en Dimension 1 :

Reference	Target	Distance	Number of Alpha Carbons in Target	Number of Alpha Carbons in Reference
ref_Homo_sapiens_CRCL_3_filtered_Ca	target_7Y8D_ca	0.6833214444778916	141	151
ref_Stylophora_pistillata_CRCL_6_filtered_Ca	target_8GYN_ca	0.6833606476372481	149	169
ref_Stylophora_pistillata_CRCL_1_filtered_Ca	target_8GYN_ca	0.6887452332929644	163	169
ref_Danio_rerio_CRCL_8_filtered_Ca	target_8GYN_ca	0.692437400949265	145	169
ref_Strongylocentrotus_purpuratus_CRCL_8_filtered_Ca	target_7Y8D_ca	0.7010578809874943	140	151
ref_Homo_sapiens_CRCL_2_filtered_Ca	target_7Y8D_ca	0.7113653736527206	140	151
ref_Homo_sapiens_CRCL_8_filtered_Ca	target_7Y8D_ca	0.7127989536032835	145	151
ref_Homo_sapiens_CRCL_1_filtered_Ca	target_8GYN_ca	0.713506944973277	160	169
ref_Stylophora_pistillata_CRCL_2_filtered_Ca	target_8GYN_ca	0.7138200184176748	175	169
ref_Hydra_vulgaris_CRCL_3_filtered_Ca	target_8GYN_ca	0.7145156513494668	164	169
ref_Homo_sapiens_CRCL_5_filtered_Ca	target_8GYN_ca	0.7172261720853185	153	169
ref_Strongylocentrotus_purpuratus_CRCL_1_filtered_Ca	target_8GYN_ca	0.7177566041474457	144	169

Protéines BCL2 candidats :  
7Y8D et 8GYN

Résultats triés par ordre croissant de distances pour chaque paire Référence protéine-Target protéine.  
Soit les potentielles BCL2 dans le cadre de notre étude.



# Analyse topologique :

## Homologie persistante :

### Conclusion :

A partir de notre pipeline, nous pouvons détecter des BCL2 candidats, et nous proposons une exploration des résultats par une approche structurale et une fonctionnelle.

### Poursuite :

Collecter plus de données et d'insights sur la persistance des caractéristiques topologiques des protéines BCL2 et des fonctions biologiques qui leurs sont associées.

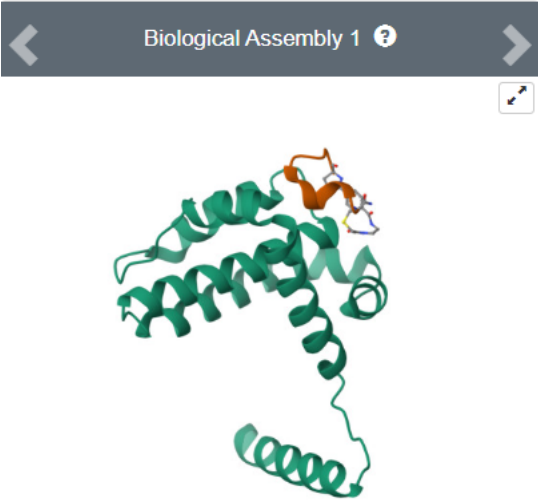
Préparation d'un dataset d'entraînement avec un label BCL2 et non-BCL2 (label BCL2 pour les BCL2-like) à chaque protéine et son barcode pour l'apprentissage supervisé d'un modèle (e.g. Random Forests) qui serait capable de détecter les protéines BCL2 et leurs homologues.

# Analyse topologique :

## Homologie persistante :

Exploration des résultats à partir de la liste des distances Référence contre All :

PDB et Uniprot (GO termes) :



### 7Y8D

Crystal structure of cp1 bound BCLxl

PDB DOI: <https://doi.org/10.2210/pdb7Y8D/pdb>

Classification: **APOPTOSIS**

Organism(s): *Homo sapiens*, synthetic construct

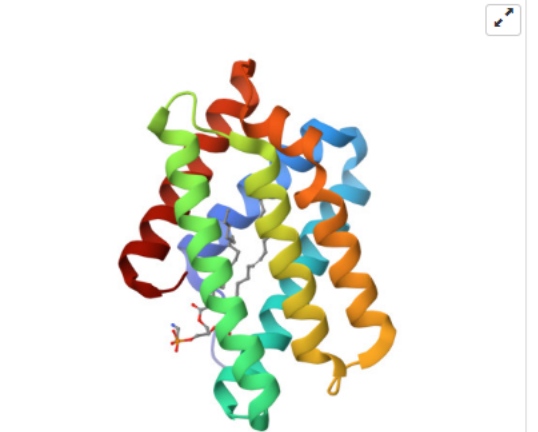
Expression System: ecological metagenomes

Mutation(s): No

Deposited: 2022-06-23 Released: 2023-11-15

Deposition Author(s): Li, F.W., Liu, C., Wu, C.L., Wu, D.L.

Funding Organization(s): National Science Foundation (NSF, China)



### 8GYN

zebrafish TIPE1 strucutre in complex with PE

PDB DOI: <https://doi.org/10.2210/pdb8GYN/pdb>

Classification: **LIPID BINDING PROTEIN**

Organism(s): *Danio rerio*

Expression System: *Escherichia coli*

Mutation(s): No

Deposited: 2022-09-23 Released: 2023-04-19

Deposition Author(s): Wang, W., Cao, S.J.

Funding Organization(s): National Natural Science Foundation of China (NSFC)

ASPECT	TERM
Cellular Component	<b>Bcl-2 family protein complex</b> <a href="#">Source:UniProtKB</a> <a href="#">1 Publication</a>
Cellular Component	<b>centrosome</b> <a href="#">Source:UniProtKB</a> <a href="#">1 Publication</a>
Cellular Component	<b>cytoplasm</b> <a href="#">Source:UniProtKB</a> <a href="#">1 Publication</a>
Cellular Component	<b>cytosol</b> <a href="#">Source:Reactome</a>
Cellular Component	<b>endoplasmic reticulum</b> <a href="#">Source:Ensembl</a>
Cellular Component	<b>mitochondrial inner membrane</b> <a href="#">Source:UniProtKB-SubCell</a>
Cellular Component	<b>mitochondrial matrix</b> <a href="#">Source:UniProtKB-SubCell</a>
Cellular Component	<b>mitochondrial outer membrane</b> <a href="#">Source:UniProtKB</a> <a href="#">1 Publication</a>
Cellular Component	<b>mitochondrion</b> <a href="#">Source:HGNC-UCL</a> <a href="#">1 Publication</a>
Cellular Component	<b>nuclear membrane</b> <a href="#">Source:UniProtKB-SubCell</a>
Cellular Component	<b>synaptic vesicle membrane</b> <a href="#">Source:UniProtKB-SubCell</a>
Molecular Function	<b>BH3 domain binding</b> <a href="#">Source:UniProtKB</a> <a href="#">1 Publication</a>

#### Keywords:

Biological process

#Apoptosis  
#Endocytosis

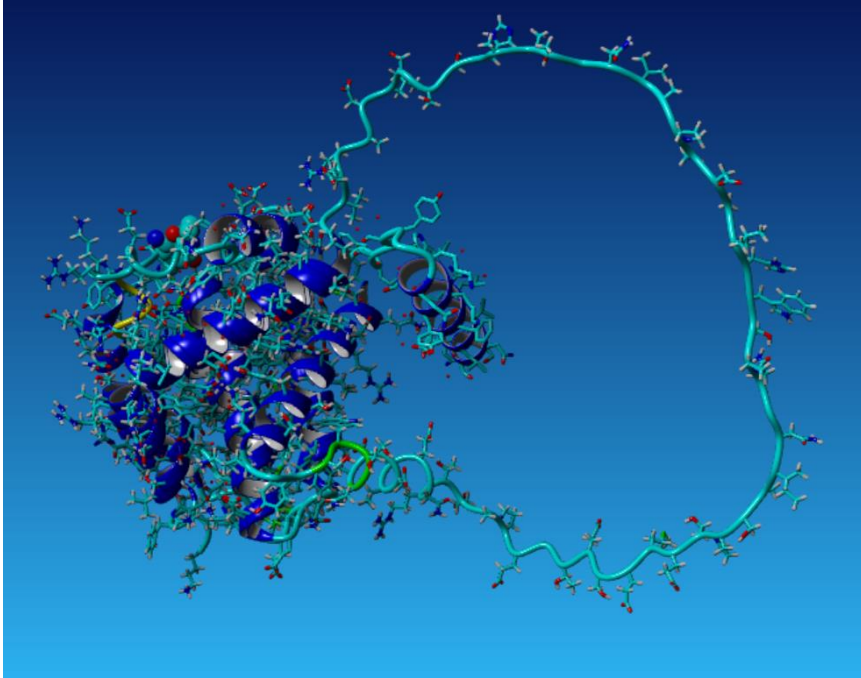
ASPECT	TERM
Cellular Component	<b>cytoplasm</b> <a href="#">Source:UniProtKB</a>
Biological Process	<b>negative regulation of TOR signaling</b> <a href="#">Source:UniProtKB</a>
Biological Process	<b>regulation of apoptotic process</b> <a href="#">Source:InterPro</a>

# Analyse topologique :

## Homologie persistante :

Etude des structures 3D :

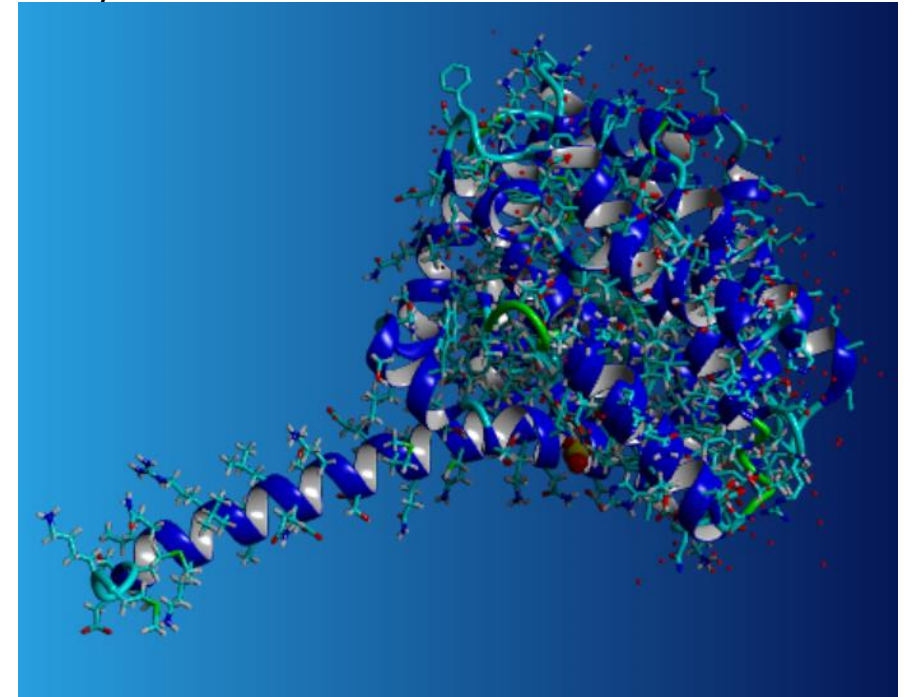
Calcul de RMSD entre BCL2 candidats et la prédiction Alphafold2 (non filtrée) :



7Y8D-Homo sapiens\_CTRL\_3 :

RMSD : 4.626 Angstrom

Ref :233 Ca – 92 Ca : 73.3% - 89.9% pLDDT



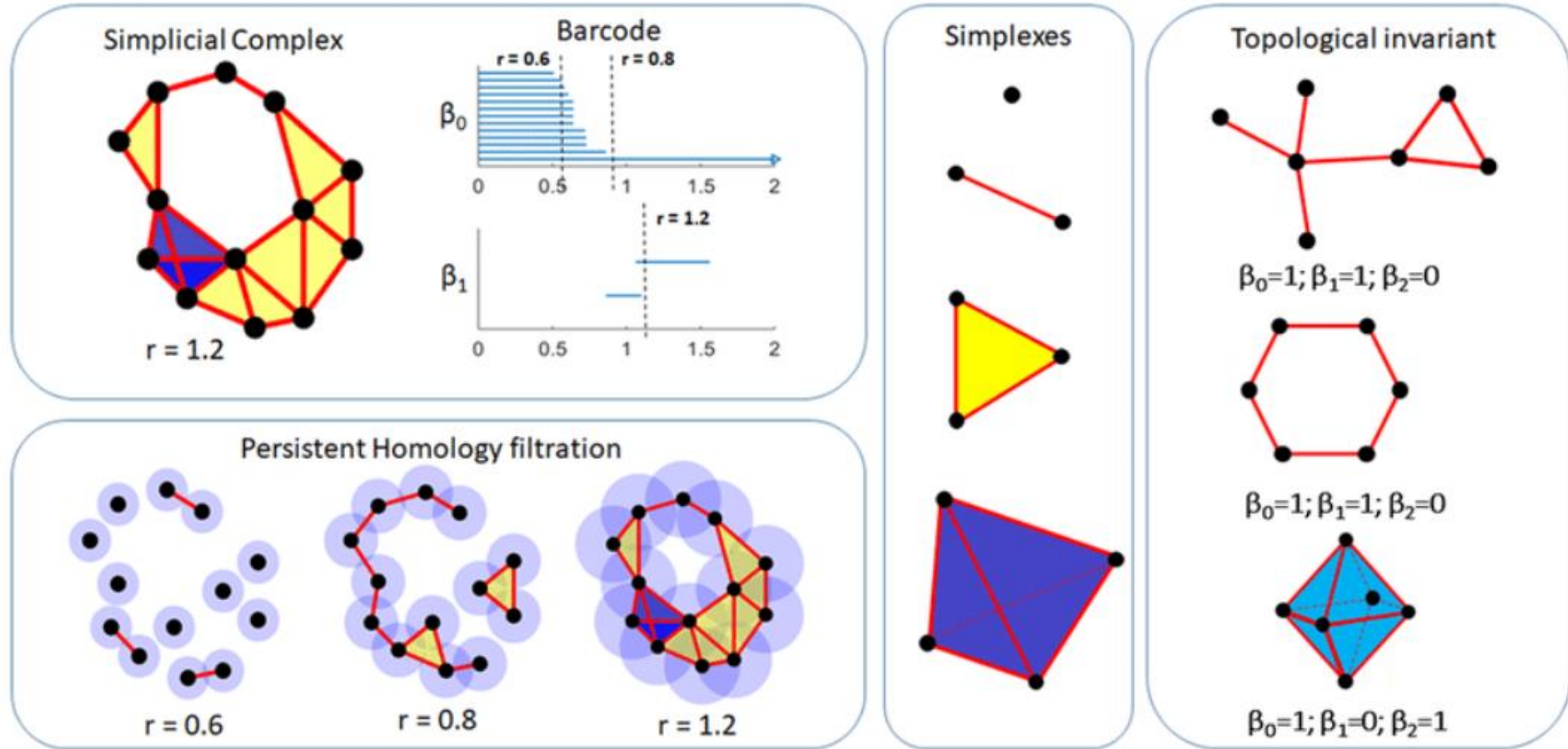
8GYN-Stylophora\_pistillata\_CTRL\_6 :

RMSD : 8.7858 Angstrom

Ref : 202 Ca – 36 Ca : 81.5% - 87.6% pLDDT



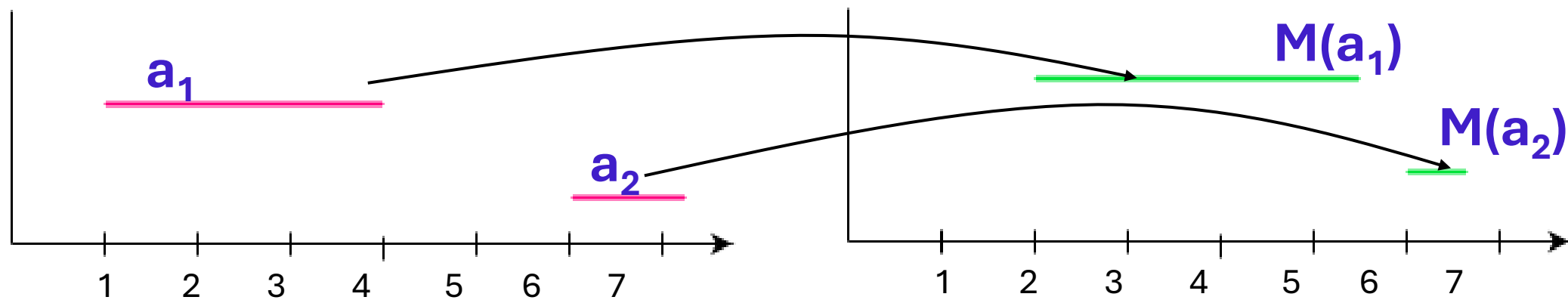
# Homologie Persistante :



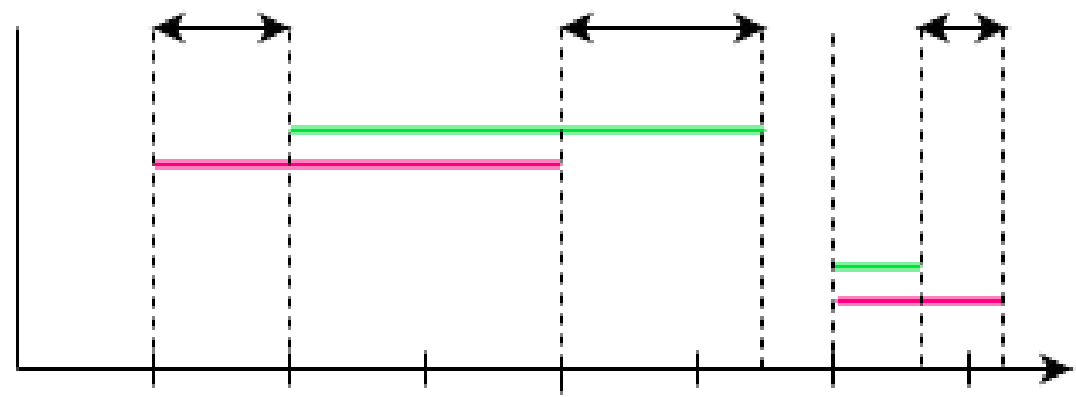
# Analyse topologique :

## Homologie persistante :

Distance de Wasserstein : Comparer deux barcodes  $B_1$  et  $B_2$  en calculant la distance entre eux.  
Objectif : Trouver le meilleur M



$$d_W(B_1, B_2) = \inf_{M \text{ matching}} \left( \sum_{a \in B_1} \|a - M(a)\|_\infty^p \right)^{\frac{1}{p}}$$



For p=1 x+y+0+z