

STRUCTURE DES MODULES DE PERSISTANCE ET GÉOMÉTRIE DES STRUCTURES TRIDIMENSIONNELLES DES PROTÉINES

GAGNAIRE–NORMAND YOAN

Résumé – Ce mémoire présente les travaux que j’ai réalisés au cours du stage de ma troisième année de licence de mathématiques, sous la direction de Philippe Malbos, dans le cadre du projet DeltaFold. Ce travail porte sur l’étude de l’homologie persistante et il s’articule en deux parties. La première se concentre sur l’énoncé et la démonstration du théorème de structure sur les modules de persistance. La seconde est une application de l’homologie persistante sur des structures tridimensionnelles de protéines.

1	Introduction	1
2	Homologie persistante	2
3	Théorème de structure des modules de persistance	8
4	Distances sur les modules de persistance	15
5	Applications à l’étude géométrique des structures tridimensionnelles de protéines	18

1. INTRODUCTION

L’analyse topologique des données est un domaine des mathématiques qui a émergé récemment grâce à l’augmentation des capacités de stockage et de calcul des ordinateurs. Elle repose sur de nombreux résultats issus de la topologie algébrique et de la géométrie.

L’un de ses principaux outils est l’homologie persistante. Celle-ci permet d’étudier les caractéristiques topologiques d’un espace à partir d’une filtration de celui-ci. Elle fournit notamment des informations sur l’apparition et la disparition de composantes connexes, de cycles et de cavités. La méthode est particulièrement adaptée aux espaces discrets [4].

Ce mémoire s’inscrit dans le programme de recherche DeltaFold qui vise à appliquer l’homologie persistante à l’étude des structures de protéine [2],[1]. Il introduit les notions fondamentales de l’homologie persistante et des modules de persistance, ainsi que les principaux résultats qui leur sont associés [6] [7] [5]. La structure de module de persistance permet de décrire l’homologie de l’espace selon la filtration. Nous appliquerons ensuite ces outils à l’étude géométrique des structures tridimensionnelles des protéines. En représentant une protéine par un nuage de points, puis en construisant une filtration adaptée, nous

2. Homologie persistante

chercherons à mettre en évidence certaines caractéristiques géométriques de ces structures biologiques et à distinguer les globines des non-globines.

Ce mémoire est organisé de la manière suivante. Dans une première section, nous introduisons les notions de complexe simplicial, d'homologie simpliciale et d'homologie persistante. Les deuxième et troisième sections sont consacrées aux modules de persistance, à leur théorème de structure ainsi qu'au théorème de stabilité. Enfin, dans la dernière section, nous appliquons ces résultats à l'étude géométrique des structures tridimensionnelles de protéines.

2. HOMOLOGIE PERSISTANTE

Cette section présente les principaux outils d'homologie persistante utilisés dans ce travail pour l'étude des caractéristiques géométriques des nuages de points. Nous y introduisons les notions de complexe simplicial, d'homologie, de complexe de chaîne et d'homologie persistante.

2.1. Complexe simplicial et complexe de chaîne

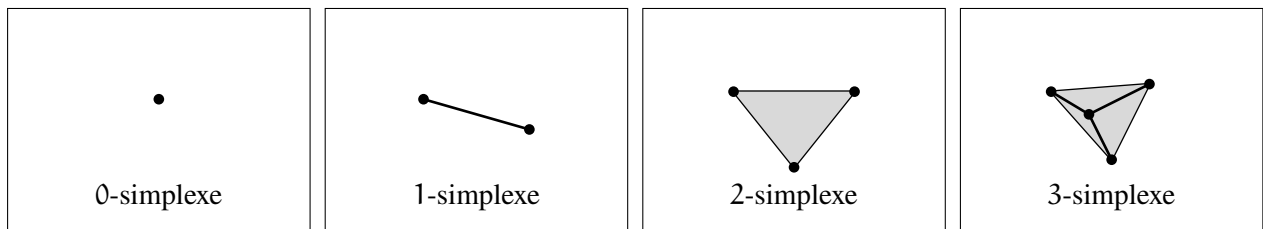
Dans cette partie, nous définissons la notion de complexe simplicial et présentons ses principales propriétés.

2.1.1. On appelle *k-simplexe* l'enveloppe convexe de $k+1$ points affinement indépendant $a_0, \dots, a_k$. On le note,

$$\sigma = \text{conv}(a_0, \dots, a_k) = \{a_0\lambda_0 + a_1\lambda_1 + \dots + a_k\lambda_k \mid \sum_{i=0}^k \lambda_i = 1, \forall i \in (0, \dots, k), \lambda_i \in \mathbb{R}\}.$$

On dit que σ est de dimension k si c'est l'enveloppe convexe de $k+1$ points. On note $\dim(\sigma) = k$.

2.1.2. Exemple. Les 0-simplexes, 1-simplexes, 2-simplexes et 3-simplexes sont respectivement les points, les segments, les triangles pleins et les tétraèdres pleins.



2.1.3. Soient $a_0, \dots, a_k \in \mathbb{R}^n$ affinement indépendant, soit σ le k -simplexe associé. Une *face* τ de σ est un simplexe obtenue comme enveloppe convexe d'un sous ensemble des points $a_0, \dots, a_k$.

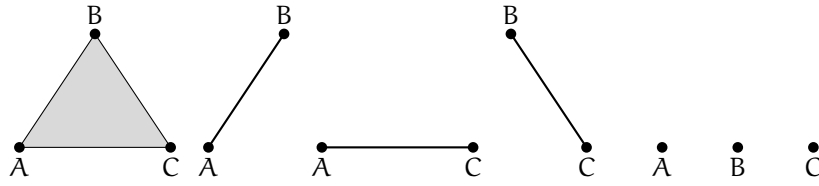
Si τ est une face de σ , on note $\tau \leq \sigma$.

2.1.4. Complexe simplicial. On appelle *complexe simplicial* K , une famille de simplexes vérifiant les propriétés suivantes :

- i) si $\sigma \in K$ et $\tau \leq \sigma$ alors $\tau \in K$,
- ii) si $\sigma, \sigma' \in K$ alors $\sigma \cap \sigma' = \emptyset$ ou $\sigma \cap \sigma' = \tau$ avec $\tau \leq \sigma$ et $\tau \leq \sigma'$.

On dit que K est de dimension d si, $d = \max_{\sigma \in K} (\dim(\sigma))$.

2.1.5. Exemple. Un 2-simplexe et ses faces



2.1.6. Dans la suite $\mathbb{K}$ désignera un corps. Pour les calculs en pratique on considère que $\mathbb{K} = \mathbb{Z}/2\mathbb{Z}$

On appelle *ensemble des k -chaînes* de K , noté $C_k(K)$, le $\mathbb{K}$ -espace vectoriel engendré par les k -simplexes de K . On notera $C_{-1}(K) = 0$ par convention.

2.1.7. Soit K un complexe simplicial et soit $\sigma \in K$ un k -simplexe. On définit le *bord* de σ , noté $\partial\sigma$, comme la $(k-1)$ -chaîne suivante

$$\partial\sigma = \sum_{i=0}^k (-1)^i \text{conv}(a_0, \dots, \check{a}_i, \dots, a_k),$$

où $\text{conv}(a_0, \dots, \check{a}_i, \dots, a_k)$ est l'enveloppe convexe des points $a_0, \dots, a_k$ privée de a_i .

La fonction *bord* $\partial_k : C_k(K) \rightarrow C_{k-1}(K)$ est définie par

$$\partial_k \left(\sum_{i=1}^n \lambda_i \sigma_i \right) = \sum_{i=1}^n \lambda_i \partial \sigma_i.$$

2.1.8. Proposition. Soit K un complexe simplicial. Alors

$$\forall k \in \mathbb{N}, \partial_k \circ \partial_{k+1} = 0. \quad (2.1.9)$$

Démonstration. Comme l'application ∂_k est linéaire, il suffit de montrer que, pour tout (k) -simplexe σ , on a

$$\partial_k \circ \partial_{k+1}(\sigma) = 0.$$

Soient $a_0, \dots, a_{k+1} \in \mathbb{R}^n$ affinement indépendant, soit σ le $k+1$ -simplexe défini sur ces points. Alors,

$$\partial_{k+1}(\sigma) = \sum_{i=0}^{k+1} (-1)^i \text{conv}(a_0, \dots, \check{a}_i, \dots, a_{k+1}).$$

2. Homologie persistante

Ainsi,

$$\begin{aligned}
 \partial_k \circ \partial_{k+1}(\sigma) &= \sum_{i=0}^{k+1} (-1)^i \partial_k(\text{conv}(a_0, \dots, \check{a}_i, \dots, a_{k+1})) \\
 &= \sum_{i=0}^{k+1} \sum_{j < i} (-1)^{i+j} \text{conv}(a_0, \dots, \check{a}_j, \dots, \check{a}_i, \dots, a_{k+1}) \\
 &\quad + \sum_{i=0}^{k+1} \sum_{j > i} (-1)^{i+j-1} \text{conv}(a_0, \dots, \check{a}_i, \dots, \check{a}_j, \dots, a_{k+1}) \\
 &= \sum_{i < j} (-1)^{i+j} \left(\text{conv}(a_0, \dots, \check{a}_i, \dots, \check{a}_j, \dots, a_{k+1}) \right. \\
 &\quad \left. - \text{conv}(a_0, \dots, \check{a}_i, \dots, \check{a}_j, \dots, a_{k+1}) \right) \\
 &= 0.
 \end{aligned}$$

□

2.2. Homologie d'un complexe simplicial

Dans cette partie, nous présentons la construction de l'homologie d'un complexe simplicial. Dans la section suivante, nous présenterons un cadre algébrique pour formuler ces concepts.

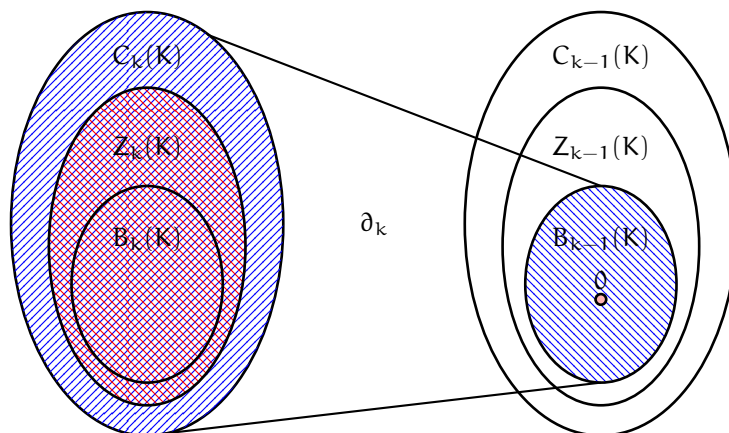
2.2.1. Soit K un complexe simplicial, on note :

- i) $Z_k(K) = \ker(\partial_k)$, le noyau de la fonction bord.
- ii) $B_k(K) = \text{Im}(\partial_{k+1})$, l'image de la fonction bord.

2.2.2. Proposition. Soit K un complexe simplicial. Alors, pour tout $k \in 0, \dots, \dim(K)$, $B_k(K)$ est un sous-espace vectoriel de $Z_k(K)$.

Démonstration. Le résultat découle de la relation (2.1.9). □

2.2.3. Exemple. Description schématique de l'application $\partial_k : C_k(K) \rightarrow C_{k-1}(K)$



2.2.4. Homologie. Soit K un complexe simplicial. On définit le k -ième groupe d'homologie de K , noté $H_k(K)$, par

$$H_k(K) = Z_k(K)/B_k(K).$$

On appelle k -ième nombre de Betti de K , noté $\beta_k(K)$, la dimension de $H_k(K)$.

2.2.5. Proposition. Soit K un complexe simplicial, alors $\beta_0(K)$ est égal au nombre de composantes connexes de K .

Démonstration. Commençons par étudier $Z_0(K)$, on a $\partial_0 : C_0(K) \rightarrow 0$, donc $Z_0(K) = C_0(K)$. Ainsi $H_0(K) = C_0(K)/B_0(K)$. Il suffit alors de prouver que deux sommets de K sont dans la même composante connexe si et seulement si elles sont congru modulo $B_0(K)$. Ainsi $H_0(K)$ sera engendrée par autant de sommets que de composantes connexes. Soient a_i et a_j deux sommets de K .

Supposons tout d'abord que a_i et a_j soient dans la même composante connexe, alors il existe une 1-chaîne τ qui relie a_i et a_j , de la forme :

$$\tau = [a_i, a_{k_1}] + \dots + [a_{k_r}, a_j],$$

où $a_{k_1}, \dots, a_{k_r}$ sont des sommets de la même composante connexe que a_i et a_j . Si l'on applique ∂_1 à τ , on obtient,

$$\partial_1(\tau) = a_i - a_{k_1} + a_{k_1} - a_{k_2} + \dots + a_{k_r} - a_j = a_i - a_j,$$

et donc

$$a_i \equiv a_j - (a_i - a_j) \equiv a_j \pmod{B_0(K)}$$

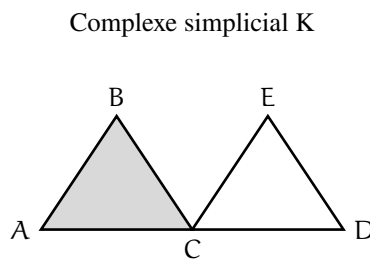
Supposons maintenant que a_i et a_j ne sont pas dans la même composante connexe. Le bord d'une arête étant formé des sommets de sa composante connexe, le bord d'une 1-chaîne est donc une somme de 0-chaînes chacune supportée dans une seule composante connexe.

Par conséquent $a_i - a_j \notin B_0(K)$, puisque seule une somme d'arêtes deux à deux reliées peut avoir comme bord $a_i - a_j$. Ainsi a_i et a_j sont dans la même composante connexe si et seulement si $a_i \equiv a_j \pmod{B_0(K)}$. Donc $H_0(K)$ est engendrée par autant de classes que de composantes connexes de K . D'où $\beta_0(K)$ est égal au nombre de composantes connexes de K . □

2.2.6. Remarques. Plus généralement $\beta_k(K)$ est le nombre k -cycles (cycle de dimension k) de K .

L'homologie mesure les cycles d'un complexe simplicial modulo les bords.

2.2.7. Exemple. Considérons les groupes d'homologie de l'exemple suivant,



2. Homologie persistante

Dans la suite $\langle A \rangle$ désignera le $\mathbb{Z}/2\mathbb{Z}$ -espace vectoriel engendré par A . Il n'y a qu'une seule composante connexe, ainsi par la proposition 2.2.5, $\dim(H_0(K)) = 1$ et on a

$$H_0(K) = \langle A \rangle$$

Étudions maintenant le 1^{er} groupe d'homologie de K . On a

$$Z_1(K) = \langle AB + BC + CA, CE + ED + DC \rangle,$$

et

$$B_1(K) = \text{Im}(\partial_2) = \langle \partial_2(ABC) \rangle = \langle AB + BC + CA \rangle.$$

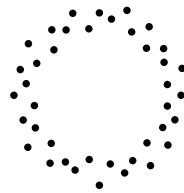
Ainsi

$$H_1(K) = \langle CE + ED + DC \rangle.$$

On a $H_1(K)$ qui est de dimension 1 et il y a bien un seul cycle.

2.3. Filtration de complexes simpliciaux

Bien que l'étude des classes d'homologie d'un complexe simplicial fournisse des informations sur ses caractéristiques géométriques, elle ne permet pas, à elle seule, de décrire convenablement la structure géométrique d'un nuage de points. Si l'on regarde le schéma de point suivant,



alors son homologie n'est constitué que du nombre de ces composantes connexes. Cependant, nous voudrions distinguer un cercle.

Pour étudier les caractéristiques géométriques de ces nuages, on introduit des filtrations de ces nuages de points et leur homologie persistante associé. Ceci permet de les décrire sous la forme d'une suite de complexes simpliciaux et d'étudier l'évolution de leurs caractéristiques géométriques.

2.3.1. On appelle *filtration d'un complexe simplicial* K une famille de complexes simpliciaux $(K_i)_{i \in \{0, \dots, n\}}$, tel que :

$$\emptyset = K_0 \subset \dots \subseteq K_i \subset K_j \subset \dots \subseteq K_n = K$$

2.3.2. Proposition. Soit $(K_i)_{i \in \{0, \dots, n\}}$ une filtration d'un complexe simplicial, alors l'inclusion $K_i \subset K_j$ induit l'application linéaire,

$$h_k^{i,j} : H_k(K_i) \rightarrow H_k(K_j).$$

Démonstration. Soient K_i et K_j des complexes simpliciaux de la filtration de K . Ainsi,

$$\forall k \in \mathbb{N}, C_k(K_i) \subseteq C_k(K_j).$$

Donc ∂_k^j est un prolongement de ∂_k^i . Ainsi $Z_k(K_i) \subseteq Z_k(K_j)$ et $B_k(K_i) \subseteq B_k(K_j)$. D'où l'existence du morphisme de $H_k(K_i)$ dans $H_k(K_j)$.

□

2.3.3. On notera $h_{k,i} = h_k^{i,i+1}$.

2.3.4. Remarque. Une classe d'homologie n'étant pas égale à zéro dans $H_k(K_i)$ peut le devenir dans $H_k(K_{i+1})$. Remarquons de plus que, si $c \in Z_k(K_i) \cap B_k(K_{i+1})$ alors $h_{k,i}(c) = 0$. Ainsi,

$$\text{Im}(h_{k,i}) = Z_k(K_i) / [Z_k(K_i) \cap B_k(K_{i+1})].$$

2.3.5. Homologie persistante. On note $H_k^{j,\ell}(K) = Z_k(K_j) / [B_k(K_\ell) \cap Z_k(K_j)]$, l'homologie persistante de la filtration $(K_i)_{i \in \mathbb{N}}$. Ce sont les classes d'homologie de K_j existant encore dans K_ℓ .

2.3.6. Intervalles de persistances. On dit qu'une classe d'homologie $c \in H_k(K_i)$ naît en K_i si $c \notin H_k^{i-1,i}(K)$ et que c meurt en K_j si elle s'annule dans $H_k(K_j)$ sans être nulle dans $H_k(K_{j-1})$, c'est-à-dire $h_k^{i,j-1}(c) \notin H_k^{i-1,j-1}(K)$ et $h_k^{i,j}(c) \in H_k^{i,j-1}(K)$. Comme illustré dans la figure 1.

Dans la pratique on représente l'évolution du nombre de générateurs du k -ième groupe d'homologie persistante avec un *barcode* constitué des intervalles $[i_c, j_c]$, dans lequel i_c est le rang de naissance de la classe c et j_c le rang de sa mort.

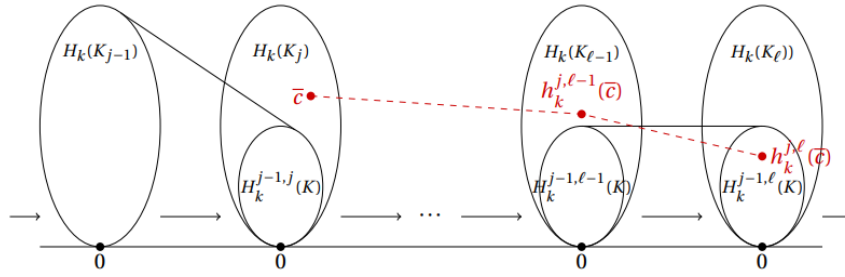
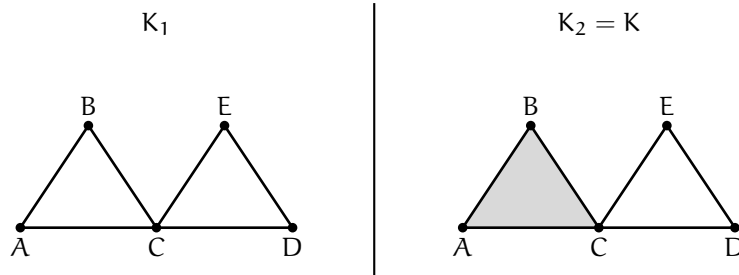


FIGURE 1 – La classe d'homologie $\bar{c}$ née dans K_j et disparaît dans K_ℓ d'indice de persistance $\ell - j$. [6]

2.3.7. Exemple. Considérons la filtration suivante.



Étudions l'homologie persistante de la filtration ci-dessus,

$$Z_1(K_1) = \langle AB + BC + CA, CD + DE + EC \rangle \quad B_1(K_2) = \langle AB + AC + CA \rangle.$$

Ainsi

$$H_1^{1,2}(K) = \langle CD + DE + EC \rangle.$$

La classe de $AB+BC+CA$ meurt en K_2 , tandis que la classe $CD+DE+EC$ reste en vie en K_2 . Le nombre de composante connexe quand à elle reste égal à 1, ainsi

$$H_0^{1,2}(K) = H_0(K) = \langle A \rangle.$$

3. Théorème de structure des modules de persistance

3. THÉORÈME DE STRUCTURE DES MODULES DE PERSISTANTE

Cette section formule algébriquement la notion d'homologie persistante sur une filtration de complexes simpliciaux. Pour cela, nous introduisons la structures algébrique de module de persistance et démontrons un théorème de structure les concernant. Nous verrons ensuite que la caractérisation de leur signature se fait à l'aide des barcodes évoqués dans la définition 2.3.7.

3.1. Module de persistance

Dans cette partie, on définit la structure de modules de persistance.

3.1.1. Module de persistance. Un *module de persistance* $\mathbb{N}$ -indexé est la donnée

- i) d'une famille $(V_n)_{n \in \mathbb{N}}$ de $\mathbb{K}$ -espaces vectoriels,
- ii) d'une famille $(\alpha_n)_{n \in \mathbb{N}}$ d'applications linéaires, telles que

$$\alpha_n : V_n \rightarrow V_{n+1}.$$

Pour chaque couple $(i, j) \in \mathbb{N}^2$ avec $i \leq j$, on note $\alpha_{i \rightarrow j}$ l'application linéaire :

$$\alpha_{i \rightarrow j} = \alpha_{j-1} \circ \alpha_{j-2} \circ \dots \circ \alpha_i.$$

Avec la convention, $\alpha_{i \rightarrow i} = \text{Id}_{V_i}$.

Un module de persistance est dit de *type fini* si chaque V_n est un espace vectoriel de dimension finie et s'il existe un rang N tel que, pour tout $n \geq N$, α_n est un isomorphisme.

3.1.2. Remarques. On peut représenter un module de persistance $\mathbb{N}$ -indexé par le diagrammes suivant :

$$V_0 \xrightarrow{\alpha_0} V_1 \xrightarrow{\alpha_1} V_2 \xrightarrow{\alpha_2} \dots \xrightarrow{\alpha_k} V_{k+1} \xrightarrow{\alpha_{k+1}} \dots$$

3.1.3. Homologie persistante de module de persistance. Soit (i, j) un couple d'entiers naturels, le *groupe d'homologie persistante* associé à un module de persistance $(V_n, \alpha_n)_{n \in \mathbb{N}}$ noté $\text{PH}_{i \rightarrow j}((V_n, \alpha_n)_{n \in \mathbb{N}})$ est définie par,

$$\text{PH}_{i \rightarrow j}((V_n, \alpha_n)_{n \in \mathbb{N}}) = \text{Im}(\alpha_{i \rightarrow j}).$$

On dit qu'un élément v dans V_i est *né* en V_i , si $v \notin \text{Im}(\alpha_{i-1})$, et qu'il est *mort* en V_j , si j est le plus petit indice tel que $\alpha_{i \rightarrow j}(v) = 0$.

On appelle *persistance* de v le couple (i, j) avec i son indice de naissance et j son indice de mort. [7]

3.1.4. Module de persistance d'une filtration d'un complexe simplicial. Soient K un complexe simplicial et (K_i) une filtration de K . Considérons les k -ièmes groupes d'homologies des complexes simpliciaux de la filtration. On obtient alors un module de persistance. En effet, considérons le diagramme suivant :

$$H_k(K_0) \xrightarrow{h_{k,0}} H_k(K_1) \xrightarrow{h_{k,1}} \dots$$

On obtient un module de persistance, précisons maintenant le groupe d'homologie persistante associé à $(H_k(K_i))_{i \in \mathbb{N}}$.

On note

$$\text{PH}_{k,i \rightarrow j}((K_i)_{i \in \mathbb{N}}) = \text{PH}_{i \rightarrow j}((H_k(K_i))_{i \in \mathbb{N}}).$$

On a alors le résultat suivant :

3.2. Théorème de structure sur les modules de type fini sur un anneau principal

3.1.5. Proposition. Soient K un complexe simplicial, (K_i) une filtration de K . Alors,
 $\text{PH}_{k,i \rightarrow j}((K_i)_{i \in \mathbb{N}}) = H_k^{i,j}(K)$.

Démonstration. Soit $(K_i)_{i \in \mathbb{N}}$ la filtration du complexe simplicial K . Alors,

$$\begin{aligned} \text{PH}_{k,i \rightarrow j}((K_i)_{i \in \mathbb{N}}) &= \text{Im}(h_{k,i \rightarrow j}) \\ &= \text{Im}(h_{k,j-1} \circ \dots \circ h_{k,i}) \\ &= Z_k(K_i) / [Z_k(K_i) \cap B_k(K_j)] \\ &= H_k^{i,j}(K). \end{aligned}$$

□

3.1.6. Soient (V_n, a_n) et (W_n, b_n) des modules de persistance $\mathbb{N}$ -indexés. Un *morphisme de modules de persistance $\mathbb{N}$ -indexés* est une famille d'applications linéaires $(\Phi_k)_{k \in \mathbb{N}}$ telle que le diagramme suivant soit commutatif pour tout $k \in \mathbb{N}$,

$$\begin{array}{ccc} V_k & \xrightarrow{a_k} & V_{k+1} \\ \Phi_k \downarrow & & \downarrow \Phi_{k+1} \\ W_k & \xrightarrow{b_k} & W_{k+1} \end{array}$$

La somme directe de (V_n, a_n) et (W_n, b_n) est le module de persistance $\mathbb{N}$ -indexé $(V_n, a_n) \oplus (W_n, b_n)$. Il est défini par

$$(V_n, a_n) \oplus (W_n, b_n) = (V_n \oplus W_n, a_n \oplus b_n),$$

où $V_n \oplus W_n$ désigne la somme directe des espaces vectoriels V_n et W_n , et où $a_n \oplus b_n$ est l'application linéaire de matrice :

$$\begin{pmatrix} a_n & 0 \\ 0 & b_n \end{pmatrix}$$

3.1.7. Module d'intervalle. Soient $0 \leq i < j \leq \infty$ des entiers. On appelle *module d'intervalle $\mathbb{N}$ -indexé*, noté $(I^{i,j}, c^{i,j})$, le module de persistance $\mathbb{N}$ -indexé définie par

$$I_k^{i,j} = \begin{cases} \mathbb{K} & \text{si } i \leq k \leq j \\ 0 & \text{sinon} \end{cases} \quad \text{et} \quad c_k^{i,j} = \begin{cases} \text{Id}_{\mathbb{K}} & \text{si } i \leq k < j \\ 0 & \text{sinon} \end{cases}$$

3.2. Théorème de structure sur les modules de type fini sur un anneau principal

La théorie des modules de persistance repose sur deux résultats fondamentaux qui joueront un rôle essentiel dans notre étude, un théorème de structure et un théorème de stabilité.

Cependant, l'établissement rigoureux du théorème de structure nécessite plusieurs résultats d'algèbre concernant les anneaux et les modules (au sens algébrique classique, et non au sens des modules de

3. Théorème de structure des modules de persistante

persistance). Ces notions seront supposées connues ; nous rappelons néanmoins le théorème de la base adaptée qui nous servira dans la démonstration du théorème de structure sur les A -modules de type fini sur un anneau principal.

3.2.1. Théorème (de la base adaptée). Soient A un anneau principal, L un A -module de type fini, libre de rang l .

Alors, tout sous- A -module M de L est de type fini, libre de rang $m \leq l$. De plus, il existe $d_1, \dots, d_m$ des éléments non nuls de A et une base $(f_1, \dots, f_l)$ de L , tel que $d_1|d_2|\dots|d_m$ et $(d_1f_1, \dots, d_mf_m)$ est une base de M .

3.2.2. Lemme. Soient A un anneau principal, $d \in A$, $p \in A$ irréductible et $h \geq 1$ un entier. Notons $E = A/(d)$ un anneau. Posons également $E_h = p^{h-1}E/p^hE$ et $\mathbb{K} = A/(p)$ (qui est un corps puisque (p) maximal car p irréductible dans un anneau principal).

Alors, si $p^h|d$, E_h est un $\mathbb{K}$ -espace vectoriel de dimension 1, sinon E_h est de dimension 0.

Démonstration. Posons $s : A \rightarrow E_h$ le morphisme de A -modules qui envoie un élément $a \in A$ sur sa classe $p^{h-1}a$. (le morphisme est obtenue par composition des morphismes canoniques $A \rightarrow E$ et $E \rightarrow E_h$). Alors s est surjective par construction. Étudions son noyau. Soit $a \in \ker(s)$, alors $ap^{h-1} \in (p^h, d)$, on distingue alors deux cas :

(i) Si $p^h|d$ alors on obtient l'équivalence suivante :

$$a \in \ker(s) \Leftrightarrow ap^{h-1} \in (p^h) \Leftrightarrow p|a.$$

Ainsi $\ker(s) = (p)$ et $E_h \simeq A/(p) = \mathbb{K}$. Le $\mathbb{K}$ -espace vectoriel E_h est donc de dimension 1.

(ii) Dans le cas contraire l'idéal (p^h, d) est engendré par un certain p^k avec $k < h$, puisque p est irréductible et A est un anneau principal. L'idéal contient donc p^{h-1} .

Ainsi $\ker(s) = A$, d'où $E_h \simeq 0$. Le $\mathbb{K}$ -espace vectoriel E_h est donc de dimension 0. [8] □

3.2.3. Théorème (de structure sur les A -modules de type fini sur un anneau principal). Soit A un anneau principal, et soit M un A -module de type fini.

Alors il existe n, r, q des entiers, avec $n = r + q$. Il existe $d_1, \dots, d_n$ des éléments de A non inversibles (éventuellement nuls), tel que $d_1, \dots, d_r$ soient non nuls, $d_1|d_2|\dots|d_r$ et

$$M \simeq A^q \oplus A/(d_1) \oplus A/(d_2) \oplus \dots \oplus A/(d_r).$$

Démonstration. Montrons d'abord l'existence des éléments $r, q, d_1, \dots, d_n$.

Soit n le nombre minimal d'éléments générateurs de M . Posons $s : A^n \rightarrow M$ un morphisme de A -modules surjectif. Le noyau $K = \ker(s)$ est un module libre de type fini de rang $r \leq n$ par le Théorème de la base adaptée.

De plus il existe $d_1, \dots, d_n$ des éléments de A et $(f_1, \dots, f_n)$ une base de $L = A^n$ (donc $\langle f_i \rangle = A$), tel que, $d_1|d_2|\dots|d_r$ et $(d_1f_1, \dots, d_rf_r)$ est une base de K , avec $d_1, \dots, d_r$ non nuls. Posons $q = n - r$, on obtient des isomorphismes de A -modules :

$$M \simeq A^n/K \simeq \bigoplus_{i=1}^n \langle f_i \rangle / \bigoplus_{i=1}^r \langle d_i f_i \rangle.$$

3.2. Théorème de structure sur les modules de type fini sur un anneau principal

En effet,

$$\langle d_i f_i \rangle \simeq d_i f_i A \simeq d_i A \simeq (d_i),$$

car f_i génère $A f_i$ et $A f_i \simeq A$. Considérons alors le morphisme $\Phi : \bigoplus_{i=1}^n f_i A \rightarrow \bigoplus_{i=1}^r A/(d_i) \oplus A^q$,

$$\Phi\left(\sum_{i=1}^n a_i f_i\right) = (\bar{a}_1, \dots, \bar{a}_r, a_{r+1}, \dots, a_n).$$

Étudions $\ker(\Phi)$. Soit $a = \sum_{i=1}^n a_i f_i \in \ker(\Phi)$, alors,

$$\forall i \leq r, a_i \in (d_i) \text{ et } \forall i > r, a_i = 0.$$

Ainsi $\sum_{i=1}^n a_i f_i = \sum_{i=1}^r b_i d_i f_i \in K$. De plus un élément de K est bien dans le $\ker$ de Φ , on a donc la relation suivante :

$$\bigoplus_{i=1}^n \langle f_i \rangle / \bigoplus_{i=1}^r \langle d_i f_i \rangle \simeq \bigoplus_{i=1}^n f_i A / \bigoplus_{i=1}^r (d_i) \simeq \text{Im}(\Phi) \simeq \bigoplus_{i=1}^r A/(d_i) \oplus A^q.$$

Montrons maintenant l'unicité des éléments $r, q, d_1, \dots, d_n$.

Soit M un A -module. Montrons que s'il existe un isomorphisme de A -modules, $M \simeq A/(d_1) \oplus \dots \oplus A/(d_r)$, où $d_1 \mid \dots \mid d_r$, alors r et les d_i à associations près peuvent être reconstruits à partir de M . Posons $d_0 = 1$. Il suffit de montrer que la liste des quotients $\frac{d_{i+1}}{d_i}$ peut être reconstruite, et pour cela, il reste à retrouver la liste des multiplicités $\alpha_p(\frac{d_{i+1}}{d_i})$ des irréductibles $p \in A$ dans $\frac{d_{i+1}}{d_i}$.

Soit $p \in A$ irréductible, de corps résiduel $k = A/(p)$. Soit $h \geq 1$ un entier, notons

$$\delta_p(h) = \dim_k(p^{h-1}M/p^hM) = \sum_{i=1}^r \dim_k(p^{h-1}A/(d_i)/p^hA/(d_i)).$$

Alors d'après le Lemme 3.2.2., on a

$$\delta_p(h) = \text{Card}(\{i; p^h \mid d_i\}).$$

Or $d_1 \mid \dots \mid d_r$. Alors

$$\delta_p(h) = r - i, \text{ si } p^h \nmid d_i \text{ et } p^h \mid d_{i+1}.$$

Ainsi

$$\text{Card}(\{\delta_p^{-1}(r - i)\}) = \text{Card}(\{h \geq 1; p^h \nmid d_i \text{ et } p^h \mid d_{i+1}\}) = \alpha_p\left(\frac{d_{i+1}}{d_i}\right).$$

Les multiplicités sont donc entièrement déterminées par les fonctions δ_p qui ne dépendent que de M . Les d_i sont donc entièrement déterminés à partir du module M . Il suffit alors de noter r le rang où $d_r \neq 0$ et $d_{r+1} = 0$ pour terminer la preuve. [8] $\square$

3. Théorème de structure des modules de persistance

3.3. Théorème de structure sur les modules de persistance

Dans cette partie on démontre le théorème de structure sur les modules de persistance $\mathbb{N}$ -indexé. Ce théorème de structure repose grandement sur le théorème de structure sur les modules de type fini.

3.3.1. Proposition. *Soit V un $\mathbb{K}$ -espace vectoriel, soit $T : V \rightarrow V$ une application linéaire. Alors V est un $\mathbb{K}[t]$ -module*

Démonstration. On considère le polynôme $P = \sum_{i=0}^n p_i t^i$. On définit la loi externe $\times : \mathbb{K}[t] \times V \rightarrow V$, par

$$P \times v = \sum_{i=0}^n p_i T^i(v_i),$$

où $T^i = T \circ T \circ \dots \circ T$ et $T^0 = \text{Id}_V$. Les propriétés suffisantes pour définir un $\mathbb{K}[t]$ -module découlent immédiatement de la linéarité des applications T^i . En effet, pour tous $P, Q \in \mathbb{K}[t]$ et pour tous $u, v \in V$, on a,

- i) $(P + Q) \times v = P \times v + Q \times v$,
- ii) $P \times (v + u) = P \times v + P \times u$,
- iii) $1 \times a = a$,
- iv) $(PQ) \times v = P \times (Q \times v)$.

□

3.3.2. Proposition. *Soient $0 \leq n < m \leq \infty$ des entiers, soit $I^{n,m}$ le module d'intervalle. Alors,*

i) *si $m < \infty$, on a :*

$$\exists j \geq i \geq 0 \in \mathbb{N}, I^{n,m} \simeq t^i \mathbb{K}[t] / (t^j),$$

ii) *si $m = \infty$, on a :*

$$\exists i \geq 0 \in \mathbb{N}, I^{n,m} \simeq t^i \mathbb{K}[t].$$

Démonstration. Commençons par montrer (i). On a $\mathbb{K}[t] = \bigoplus_{k=0}^{\infty} t^k \mathbb{K}$, ainsi

$$t^i \mathbb{K}[t] / (t^j) = \bigoplus_{k=0}^{j-i} t^{i+k} \mathbb{K}.$$

Cela correspond exactement au module d'intervalle $I^{i,j}$, puisque les composantes sont égales à $\mathbb{K}$ entre les indices i et j , et nulles sinon.

Montrons maintenant (ii). On a la même identité sauf qu'aucun des t^{k+i} n'est nul. On obtient ainsi l'équivalence avec $I^{i,\infty}$. □

3.3.3. Théorème (de structure sur les modules de persistance $\mathbb{N}$ -indexé de type fini). *Soit (V, a) un module de persistance $\mathbb{N}$ -indexé de type fini.*

Il existe alors un ensemble $\text{Bar}(V, a) \subset \tilde{\mathbb{N}}^2$, constitué de couples $0 \leq i < j \leq \infty$ et une multiplicité $\mu : \text{Bar}(V, a) \rightarrow \mathbb{N}^$ telle que*

$$(V, a) \simeq \bigoplus_{(i,j) \in \text{Bar}(V,a)} (I^{i,j}, c^{i,j})^{\mu(i,j)}.$$

De plus cette décomposition est unique à isomorphisme près.

Démonstration. Comme (V, a) est de type fini, il existe un entier n tel que, pour tout $i \geq n$, a_i est un isomorphisme. Posons $V = \bigoplus_{i=0}^n V_i$ qui est un espace vectoriel comme somme directe d'espaces vectoriels. On considère l'application linéaire $t : V \rightarrow V$ définie par

$$t((v_0, \dots, v_n)) = (0, a_0(v_0), \dots, a_{n-1}(v_{n-1})).$$

Cette application induit une structure de $\mathbb{K}[t]$ -module sur V par la Proposition 3.3.1. Comme chaque V_i est de dimension finie, V est un $\mathbb{K}[t]$ -module de type fini. De plus $\mathbb{K}[t]$ est un anneau principal puisque $\mathbb{K}$ est un corps. Ainsi, par le Théorème 3.2.3, il existe $t^{j_1}, \dots, t^{j_r} \in \mathbb{K}[t]$ tels qu'on obtienne l'isomorphisme suivant,

$$V \simeq \mathbb{K}[t]^q \oplus \mathbb{K}[t]/(t^{j_1}) \oplus \dots \oplus \mathbb{K}[t]/(t^{j_r}).$$

Où q et les $j_1, \dots, j_r$ sont des entiers.

Pour chaque facteur libre $\mathbb{K}[t]$, choisissons un générateur e_{i_k} . Si le générateur e_{i_k} apparaît au degré i_k , alors le sous-module qu'il engendre est isomorphe à $t^{i_k} \mathbb{K}[t]$. On obtient donc les isomorphismes suivant,

$$\mathbb{K}[t] \simeq e_{i_k} \mathbb{K}[t] \simeq t^{i_k} \mathbb{K}[t],$$

et

$$\mathbb{K}[t]/(t^{j_k}) \simeq e_{i_k} \mathbb{K}[t]/e_{i_k}(t^{j_k}) \simeq t^{i_k} \mathbb{K}[t]/(t^{i_k+j_k}).$$

En reportant cette relation dans la première décomposition, on obtient les isomorphismes,

$$\begin{aligned} V &\simeq \mathbb{K}[t]^q \oplus \mathbb{K}[t]/(t^{j_1}) \oplus \dots \oplus \mathbb{K}[t]/(t^{j_r}) \\ &\simeq t^{i_1} \mathbb{K}[t] \oplus \dots \oplus t^{i_q} \mathbb{K}[t] \oplus t^{i_{q+1}} \mathbb{K}[t]/(t^{i_{q+1}+j_1}) \oplus \dots \oplus t^{i_{q+r}} \mathbb{K}[t]/(t^{i_{q+r}+j_r}). \end{aligned}$$

En appliquant la proposition 3.3.2, on obtient les isomorphismes suivant :

$$V \simeq I^{i_1, \infty} \oplus \dots \oplus I^{i_q, \infty} \oplus I^{i_{q+1}, i_{q+1}+j_1} \oplus \dots \oplus I^{i_{q+r}, i_{q+r}+j_r}.$$

Il suffit de poser $\text{Bar}(V, a)$ comme l'ensemble des couples apparaissant dans la décomposition, et de définir μ en associant à chaque couple (i, j) de $\text{Bar}(V, a)$ sa multiplicité dans la décomposition. L'unicité de la décomposition à isomorphisme près vient de l'unicité du théorème de structure sur les modules de type fini sur un anneau principal. [7] $\square$

3.4. Barcodes, diagrammes de persistance

Dans cette partie, nous introduisons les barcodes, qui constituent une signature algébrique des modules de persistance. Ceux-ci ont été évoqués pour la première fois dans la définition 2.3.7.

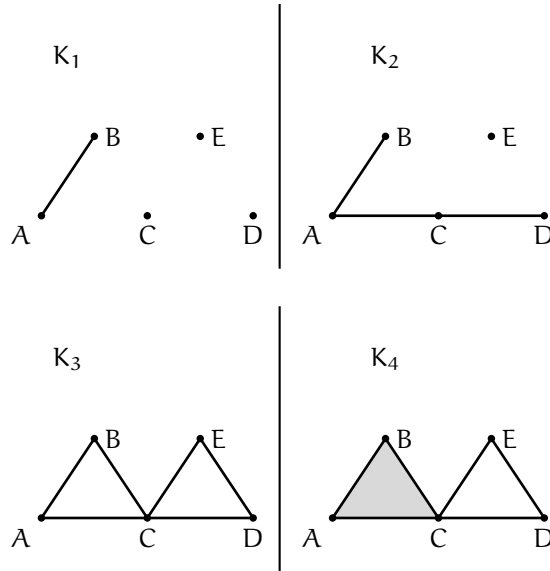
3.4.1. Barcode. Soit (V, a) un module de persistance $\mathbb{N}$ -indexé. On appelle *barcode* de (V, a) le multi-ensemble d'intervalles, où chaque intervalle $[i, j]$ est composé à partir des couples de $\text{Bar}(V, a)$. Les intervalles sont présent autant de fois que leur multiplicité dans la décomposition.

Ce barcode peut être représenté graphiquement sous forme d'une famille d'intervalles $[i, j]$, chaque intervalle étant représenté avec sa multiplicité.

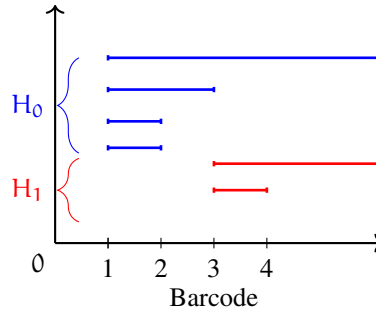
3. Théorème de structure des modules de persistante

3.4.2. Remarque. On peut représenté graphiquement la donnée des couples et leur multiplicité sous forme d'un multi-ensemble de points (i, j) , appelé diagramme de persistante.

3.4.3. Exemple. Considérons la filtration suivante.



Le barcode associé à l'homologie persistante de cette filtration est



3.5. Module de persistante $\mathbb{R}^+$ -indexé

Dans la pratique la construction d'une filtration d'un complexe simplicial ce fait avec un indicage sur $\mathbb{R}^+$. On développera cette construction dans la section 5. Dans cette partie, on définit les modules de persistante $\mathbb{R}^+$ -indexés et leurs propriétés.

3.5.1. Module de persistante. On appelle *module de persistante $\mathbb{R}^+$ -indexé*, la donnée

- i) d'une famille $V = (V_s)_{s \in \mathbb{R}^+}$, où V_s est un $\mathbb{K}$ -espace vectoriel.
- ii) d'une famille $\alpha = (\alpha_{s \leq t})_{s, t \in \mathbb{R}^+}$, où pour chaque couple de réels positifs $s \leq t$, $\alpha_{s \leq t}$ est une application linéaire de V_s dans V_t .

Ces applications linéaires doivent de plus, respecter les propriétés suivantes :

- i) $\forall t \geq 0, \alpha_{t \leq t} = \text{Id}_{V_t}$
- ii) $\forall 0 \leq r \leq s \leq t, \alpha_{s \leq t} \circ \alpha_{r \leq s} = \alpha_{r \leq t}$

3.5.2. Soit (V, α) un module de persistance $\mathbb{R}^+$ -indexé. On dit que (V, α) est *ponctuellement de type fini*, s'il respecte ces deux propriétés :

- i) Pour tout $t \geq 0, V_t$ est un $\mathbb{K}$ -espace vectoriel de dimension finie.
- ii) Il n'existe qu'un nombre fini de termes $t \leq 0$ appelés *valeurs critiques*, tel que

$$\forall \varepsilon > 0, \alpha_{t-\varepsilon \leq t+\varepsilon} \text{ ne soit pas un isomorphisme}$$

3.5.3. Module d'intervalle. Soient $0 \leq s < t \leq \infty$ des réels, soient $r, r_1, r_2 \in \mathbb{R}^+$. On définit le *module d'intervalle* $\mathbb{R}^+$ -indexé, $(I^{s,t}, c^{s,t})$, de la façon suivante,

$$I_r^{s,t} = \begin{cases} \mathbb{K} & \text{si } s \leq r \leq t \\ 0 & \text{sinon} \end{cases} \quad \text{et} \quad c_{r_1 \leq r_2}^{s,t} = \begin{cases} \text{Id}_{\mathbb{K}} & \text{si } [r_1, r_2] \subset [s, t] \\ 0 & \text{sinon} \end{cases}$$

3.5.4. Proposition. Soit (V, α) un module de persistance $\mathbb{R}^+$ -indexé. Supposons de plus que (V, α) est ponctuellement de type fini. Alors, il existe un ensemble $\text{Bar}(V, \alpha) \subset \mathbb{R}^2$ et une multiplicité $\mu : \text{Bar}(V, \alpha) \rightarrow \mathbb{N}^*$ telle que

$$(V, \alpha) \simeq \bigoplus_{(s,t) \in \text{Bar}(V, \alpha)} (I^{s,t}, c^{s,t})^{\mu(s,t)}.$$

Démonstration. Le fait que (V, α) soit ponctuellement de type fini, permet de se ramener à un module de persistance $\mathbb{N}$ -indexé. Considérons $(t_i)_{i \in \{0, \dots, n\}}$ la famille des valeurs critiques de (V, α) . Définissons alors (V', α') un module de persistance $\mathbb{N}$ -indexé tel que

$$\forall 0 \leq i \leq n, V'_i = V_{t_i} \text{ et } \alpha'_i = \alpha_{t_i \leq t_{i+1}}.$$

En appliquant le Théorème 3.3.3, le résultat découle. □

4. DISTANCES SUR LES MODULES DE PERSISTANCE

Cette section introduit une notion de pseudo-distance naturelle entre modules de persistance, permettant de les comparer entre eux. Nous verrons qu'il est possible de traduire cette distance en une distance sur les barcodes. Ce résultat fondamental est appelé théorème de stabilité. Ces distances jouent un rôle essentiel, car elles fondent les méthodes de comparaison basées sur les barcodes.

4.1. Distance d'entrelacement

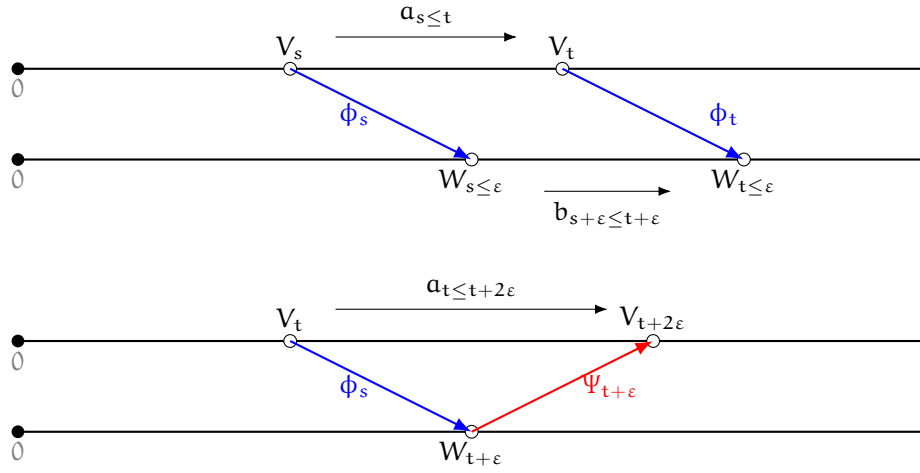
On définit dans cette partie, la pseudo-distance naturelle sur les modules de persistance.

4. Distances sur les modules de persistance

4.1.1. Entrelacement. Soient (V,a) et (W,b) deux modules de persistance $\mathbb{R}^+$ -indexé. On dit que (V,a) et (W,b) sont ε -entrelacé, s'il existe deux familles d'applications linéaires, $(\Phi_t : V_t \rightarrow W_{t+\varepsilon})_{t \in \mathbb{R}^+}$ et $(\Psi_t : W_t \rightarrow V_{t+\varepsilon})_{t \in \mathbb{R}^+}$, satisfaisant les propriétés suivantes, pour tous $s \leq t$,

- i) $\Phi_t \circ a_{s \leq t} = b_{s+\varepsilon \leq t+\varepsilon} \circ \Phi_s$,
- ii) $\Psi_t \circ b_{s \leq t} = a_{s+\varepsilon \leq t+\varepsilon} \circ \Psi_s$,
- iii) $\Psi_{t+\varepsilon} \circ \Phi_t = a_{t \leq t+2\varepsilon}$,
- iv) $\Phi_{t+\varepsilon} \circ \Psi_t = b_{t \leq t+2\varepsilon}$.

4.1.2. Remarques. Il est possible de représenter les relations de la définition ci-dessus à l'aide des diagrammes suivant,



4.1.3. Distance d'entrelacement. Soient (V,a) et (W,b) deux modules de persistance $\mathbb{R}^+$ -indexé. On définit la *distance d'entrelacement*, noté $d((V, a), (W, b))$ par

$$d((V, a), (W, b)) = \inf\{\varepsilon \geq 0 \mid (V,a) \text{ et } (W,b) \text{ sont } \varepsilon\text{-entrelacé}\}.$$

Si (V,a) et (W,b) ne sont entrelacé pour aucun $\varepsilon \geq 0$, on note alors $d((V, a), (W, b)) = \infty$.

4.1.4. Proposition. La distance d'entrelacement est un pseudo-distance. Soient (V,a) , (W,b) et (X,c) des modules de persistance, alors

$$d((V, a), (W, b)) \geq 0 \quad \text{et} \quad d((V, a), (X, c)) \leq d((V, a), (W, b)) + d((W, b), (X, c)).$$

Démonstration. La première propriété découle directement de la définition de la distance d'entrelacement. Soient $\varepsilon_1, \varepsilon_2$ tels que,

$$\varepsilon_1 > d((V, a), (W, b)), \quad \text{et} \quad \varepsilon_2 > d((W, b), (X, c)),$$

et qu'il existe un ε_1 -entrelacement entre (V,a) et (W,b) , et un ε_2 -entrelacement entre (W,b) et (X,c) . Alors par composition, il existe un $\varepsilon_1 + \varepsilon_2$ -entrelacement entre (V,a) et (X,c) . L'inégalité triangulaire suit en faisant tendre les ε_i vers leur distance respective. $\square$

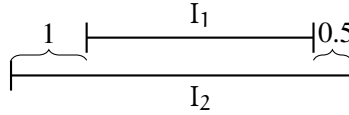
4.2. Distance Bottleneck et de Wasserstein

Dans cette partie, nous définissons une distance sur les barcodes ainsi que son lien avec la distance d'entrelacement. Ce lien est donné par le théorème de stabilité.

4.2.1. Correspondance. Soient B et B' des multi-ensembles d'intervalles. Une ε -correspondance entre B et B' est une bijection $\rho : B_0 \rightarrow B'_0$, où $B_0 \subset B$ et $B'_0 \subset B'$. Cette bijection vérifie les propriétés suivantes :

- i) Pour tout intervalle $[s, t] \in (B \setminus B_0) \cup (B' \setminus B'_0)$, on a $t - s \leq 2\varepsilon$,
- ii) Pour tout $[s, t] \in B_0$, si $\rho([s, t]) = [s', t']$, alors $|s - s'| \leq \varepsilon$ et $|t - t'| \leq \varepsilon$.

4.2.2. Exemple. Considérons deux intervalles I_1 et I_2 ,



I_1 et I_2 ont une ε -correspondance, si $\varepsilon \geq 1$.

4.2.3. Distance bottleneck. Soient B_1 et B_2 deux barcodes. On définit la distance bottleneck entre B_1 et B_2 par,

$$d_{\text{Bot}}(B_1, B_2) = \inf\{\varepsilon \geq 0 \mid \text{il existe une } \varepsilon\text{-correspondance entre } B_1 \text{ et } B_2\}.$$

4.2.4. Théorème (Théorème de stabilité). Soient (V, a) et (W, b) deux modules de persistance $\mathbb{R}^+$ -indexé ponctuellement de type fini, et B_1, B_2 leurs barcodes respectifs. Alors,

$$d((V, a), (W, b)) = d_{\text{Bot}}(B_1, B_2).$$

4.2.5. p-coût de Wasserstein. Soient B et B' deux multi-ensemble d'intervalles. Soit $\rho : B_0 \rightarrow B'_0$ une ε -correspondance entre B et B' . On appelle p -coût de Wasserstein de ρ , noté $C_p(\rho)$, la valeur suivante :

$$C_p(\rho) = \left(\sum_{([s, t], [s', t']) \in B_0 \times B'_0} \|(s, t) - (s', t')\|_\infty^p + \sum_{[s, t] \in (B \setminus B_0) \cup (B' \setminus B'_0)} \left(\frac{|s - t|}{2} \right)^p \right)^{\frac{1}{p}}.$$

4.2.6. Distance de Wasserstein. Soient B_1 et B_2 deux barcodes. On définit la p -distance de Wasserstein par,

$$d_{p\text{-Was}}(B_1, B_2) = \inf\{C_p(\rho) \mid \rho \text{ soit une } \varepsilon\text{-correspondance entre } B_1 \text{ et } B_2\}.$$

4.2.7. Remarque. La distance de Wasserstein permet de donner plus ou moins d'importance au nombre de petits intervalles. Une propriété remarquable de cette distance est que si l'on fait tendre p vers $+\infty$, on obtient la distance bottleneck.

5. APPLICATIONS À L'ÉTUDE GÉOMÉTRIQUE DES STRUCTURES TRIDIMENSIONNELLES DE PROTÉINES

Dans cette section, nous appliquons la théorie des modules de persistance et de l'homologie persistante à l'étude des structures tridimensionnelles des protéines.

Chaque protéine est d'abord représentée par un nuage de points. À partir de celui-ci, nous construisons une filtration de complexes simpliciaux grâce à la méthode de filtration par α -complexes, ce qui permet ensuite d'étudier ses caractéristiques géométriques à l'aide de l'homologie persistante.

5.1. Méthode de filtration α -complexe

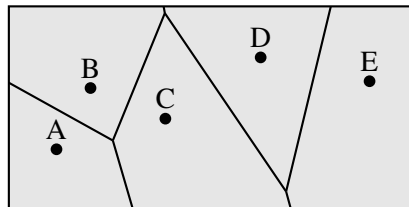
Dans cette partie, nous définissons la filtration par α -complexes afin de transformer un nuage de points en une filtration de complexes simpliciaux.

5.1.1. Diagramme de Voronoï. Soit $X \subset \mathbb{R}^n$ un nuage de points. On appelle *diagramme de Voronoï* de X , la partition de $\mathbb{R}^n$ formée par les cellules de Voronoï des points de X .

La *cellule de Voronoï* d'un point $a \in X$ est l'ensemble des points de $\mathbb{R}^n$ qui sont plus proches de a que de n'importe quel autre point de X . On note cette cellule V_a ,

$$V_a = \{x \in \mathbb{R}^n \mid d(x, a) \leq d(x, b), \forall b \in X\}$$

5.1.2. Exemple. Considérons le nuage de points $\{A, B, C, D, E\} \subset \mathbb{R}^2$, voici le diagramme de Voronoï associé :

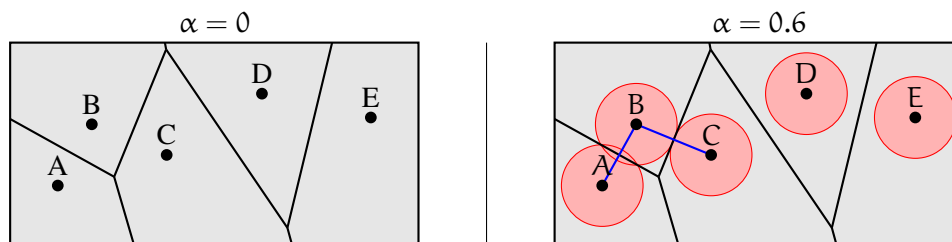


5.1.3. Filtration α -complexe. Soit $X \subset \mathbb{R}^n$ un nuage de points. La *filtration α -complexe* de X est une famille de complexes simpliciaux $(AC_\alpha(X))$ indexée par α un réel positif.

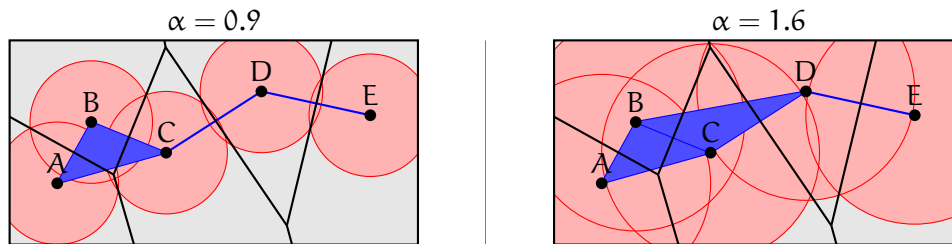
Soit $\alpha \geq 0$, on définit le complexe simplicial $AC_\alpha(X)$ de la façon suivante,

$$\sigma \in AC_\alpha(X) \iff \bigcap_{p \in \sigma} (B(p, \alpha) \cap V_p) \neq \emptyset$$

5.1.4. Exemple. Reprenons le nuage de points de l'exemple 5.1.2 et appliquons y la filtration α -complexe.



5.2. Filtration et homologie persistante d'une structure tridimensionnelle de protéine



5.2. Filtration et homologie persistante d'une structure tridimensionnelle de protéine

Dans cette partie, nous rappelons ce que sont les protéines, puis nous expliquons comment associer à une protéine un nuage de points et comment construire, à partir de celui-ci, une filtration de complexes simpliciaux.

5.2.1. Protéines. Les *protéines* sont des macromolécules constituées d'acides aminés. Elles assurent de nombreuses fonctions essentielles au sein des organismes vivants.

Un *acide aminé* est un composé organique dont la structure atomique est formée d'un atome de carbone central, appelé carbone alpha ($C\alpha$), lié à un groupe amine (NH_2) et à un groupe carboxyle ($COOH$).

5.2.2. Dans cette étude, la structure tridimensionnelle d'une protéine est représentée par un nuage de points constitué des coordonnées de ses atomes de carbone alpha.

En appliquant la filtration par α -complexes à ce nuage de points, on obtient une filtration de complexes simpliciaux. C'est sur cette filtration que nous étudions l'homologie persistante, qui fournit des informations sur certaines caractéristiques géométriques de la protéine.

L'étude de l'homologie persistante en dimension 1 permet d'observer l'apparition de cycles au cours de la filtration. Plus une classe d'homologie possède une longue durée de vie, plus la caractéristique géométrique associée est susceptible d'être significative dans la structure tridimensionnelle de la protéine. L'étude en dimension 2 permet quant à elle de mettre en évidence l'apparition de cavités.

5.2.3. Exemples. Représentons le nuage de points constitué des coordonnées des atomes de carbone-alpha et la filtration associée, de la protéine 1A1W.

L'analyse du diagramme de persistance de la protéine 1A1W permet de retrouver une poche sous la forme d'une cavité de longue durée de vie. La multitude de cycles observés est quant à elle caractéristique de la modélisation des hélices alpha.

5.3. Analyse de la structure tridimensionnelle des globines

Dans cette partie, nous considérons un jeu de données constitué de 152 protéines. Parmi celles-ci, 120 sont des globines et 32 ne sont pas des globines. L'objectif est de déterminer dans quelle mesure l'homologie persistante permet de distinguer ces deux catégories de protéines.

5. Applications à l'étude géométrique des structures tridimensionnelles de protéines

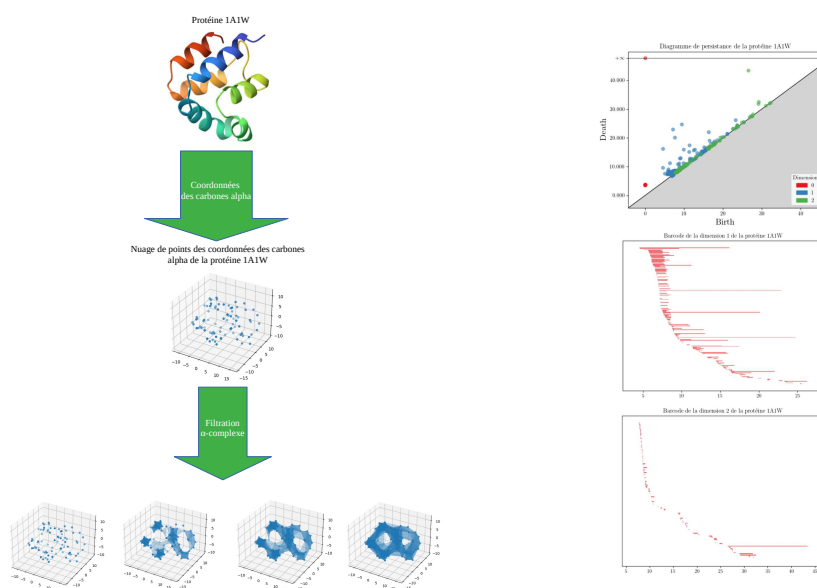


FIGURE 2 – Filtration et diagramme de persistance de la protéine 1A1W

5.3.1. Globines. Une *globine* est une protéine présente dans le sang et intervenant dans le déplacement de l'oxygène, elle possède en générale une structure composée d'une poche et d'hélices alpha. Ces hélices alpha sont une structure secondaire en forme hélicoïdale. Les globines forment une famille de protéines, c'est-à-dire un ensemble de protéines dont on suppose qu'elles descendent d'un ancêtre commun.

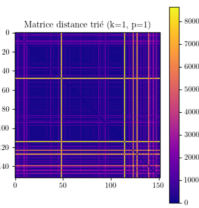
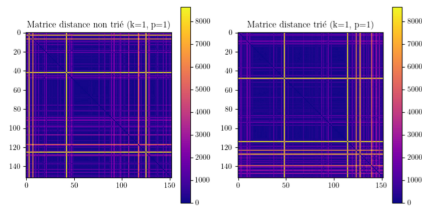
5.3.2. Les matrices suivantes représentent les distances entre les barcodes associés aux structures tridimensionnelles des protéines, calculées à l'aide de la méthode décrite au paragraphe 5.2.2. Les résultats dépendent de deux paramètres : k , le degré homologique, et p , le paramètre de la distance de Wasserstein. Les valeurs retenues dans cette étude sont celles qui ont fourni la meilleure discrimination entre globines et non-globines.

Les matrices sont présentées sous deux formes : non triées, lorsque l'ordre initial du jeu de données est conservé, et triées, lorsque les protéines sont regroupées en globines puis en non-globines.

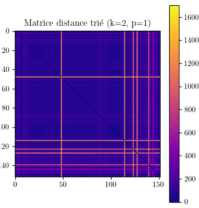
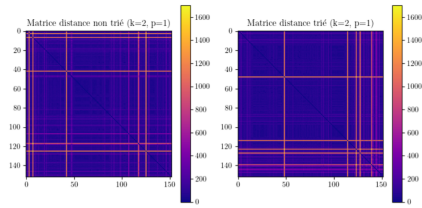
Les tableaux comparatifs donnent les distances moyennes entre deux globines, entre une globine et une non-globine, ainsi qu'entre deux non-globines.

Les résultats obtenus permettent de distinguer de manière significative les globines des non-globines. En effet, la distance entre deux structures de globines est en moyenne nettement plus faible que celle observée entre une globine et une non-globine. Deux globines apparaissent toutefois très éloignées des autres. Cette différence peut s'expliquer par leur taille exceptionnellement grande. Les protéines concernées, l'ADGB Human Androglobin et l'ADGB Mouse Androglobin, possèdent chacune plus de 1600 points dans leur nuage de points, alors que la moyenne observée sur l'ensemble des protéines étudiées est d'environ 250 points.

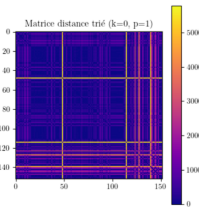
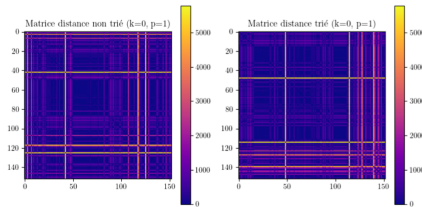
5.4. Conclusion



	k=1 p=1	
	Globines	Non-Globines
Globines	705.800	1483.089
Non-Globines	1483.089	1957.271



	k=2 p=1	
	Globines	Non-Globines
Globines	116.554	247.962
Non-Globines	247.962	335.364



	k=0 p=1	
	Globines	Non-Globines
Globines	422.649	955.136
Non-Globines	955.136	1156.058

Dans ce mémoire, nous avons présenté les principaux concepts de l'homologie persistante, en introduisant les modules de persistance ainsi que le théorème de structure qui leur est associé. Ces outils permettent de décrire les caractéristiques géométriques d'un nuage de points à travers une filtration.

Nous avons ensuite appliqué ces résultats à l'étude géométrique des structures tridimensionnelles des protéines, montrant ainsi comment la topologie algébrique peut fournir des informations sur des objets biologiques complexes.

Afin d'aller plus loin, plusieurs perspectives peuvent être envisagées. D'une part, il serait intéressant de développer des méthodes permettant de détecter efficacement les poches des globines lorsque celles-ci ne correspondent pas à des cavités de longue persistance dans les diagrammes de persistance. D'autre part, l'étude de l'homologie persistante multi-paramétrée [3] pourrait permettre de capturer des informations géométriques plus riches et ouvrir de nouvelles applications dans l'analyse des structures protéiques.

RÉFÉRENCES

RÉFÉRENCES

- [1] Léa Bou Dagher, Dominique Mader, Philippe Malbos, and Céline Brochier-Armanet. Persistent homology reveals strong phylogenetic signal in 3d protein structures. *PNAS Nexus*, 3(4) :pgae158, 04 2024.
- [2] Léa Bou Dagher, Dominique Mader, Philippe Malbos, and Céline Brochier-Armanet. Faithful interpretation of protein structures through weighted persistent homology improves evolutionary distance estimation. *Molecular Biology and Evolution*, 42(2) :msae271, 02 2025.
- [3] Gunnar Carlsson and Afra Zomorodian. The theory of multidimensional persistence. *Discrete Computational Geometry*, pages Vol. 42, pp. 71–93, 2009.
- [4] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis : Fundamental and practical aspects for data scientists. *arXiv :1710.04019*, 2021.
- [5] Grégory Ginot. Exposé Bourbaki 1225 : Un aperçu des modules de persistance et de leurs applications. *Asterisque*, January 2025.
- [6] Philippe Malbos. Raconte-moi... la persistance topologique. *La Gazette de la Société Mathématique de France*, pages Vol. 169, pp. 44–50, July 2021.
- [7] Vidit Nanda. *Computational Algebraic Topology*. Disponible à l'adresse : <https://people.maths.ox.ac.uk/nanda/cat/TDANotes.pdf>.
- [8] Matthieu Romagny. *Anneaux et modules*. Disponible à l'adresse : https://perso.univ-rennes1.fr/matthieu.romagny/M1ALGB1213/cours_ALGB.pdf.