Vieillissement du système cardio-vasculaire – Etude de l'activité des peptides d'élastine

Azzam Alwan

Thèse
de doctorat
de l'UTT

Azzam ALWAN

Vieillissement du système cardio-vasculaire – Etude de l'activité des peptides d'élastine.

Spécialité :
Optimisation et Sûreté des Systèmes

2018TROY0034

Année 2018

utt
UNIVERSITÉ DE TECHNOLOGIE
TROYES

THESE

pour l'obtention du grade de

DOCTEUR de l'UNIVERSITE DE TECHNOLOGIE DE TROYES

Spécialité : OPTIMISATION ET SURETE DES SYSTEMES

présentée et soutenue par

Azzam ALWAN

le 22 octobre 2018

Vieillissement du système cardio-vasculaire – Etude de l'activité des peptides d'élastine

JURY

Mme. S. RUAN	PROFESSEUR DES UNIVERSITES	Présidente
M. S. ALMAGRO	MAITRE DE CONFERENCES	Examinateur
M. P. BEAUSEROY	PROFESSEUR DES UNIVERSITES	Directeur de thèse
M. R. COGRANNE	MAITRE DE CONFERENCES	Rapporteur
M. E.-H. DJERMOUNE	MAITRE DE CONFERENCES- HDR	Rapporteur
M. P. TUFFERY	DIRECTEUR DE RECHERCHE INSERM	Examinateur

Personnalité invitée

M.M. DAUCHEZ	PROFESSEUR DES UNIVERSITES	Directeur de thèse
Mme. E GRALL-MAËZ	MAITRE DE CONFERENCES- HDR	

remerciements

À l'issue de la rédaction de ma thèse, je suis convaincu que la thèse est loin d'être un travail solitaire. En effet, je n'aurais jamais pu réaliser ce travail doctoral sans le soutien d'un grand nombre de personnes dont la générosité, la bonne humeur et l'intérêt manifestés à l'égard de ma recherche m'ont permis de progresser dans cette phase délicate de « l'apprenti-chercheur». Par ces quelques lignes, je tiens à remercier toutes les personnes qui ont participé de près ou de loin au bon déroulement de cette thèse, en espérant n'avoir oublié personne...

Je tiens en premier lieu à remercier les membres du jury qui me font l'honneur de juger ce travail de thèse et d'être présents parmi nous. Un grand merci au Mr. Pierre TUFFERY et au Mr. El-Hadi DJERMOUNE d'avoir accepté d'être les rapporteurs de ce travail de thèse, du temps qu'ils m'ont consacré et de leurs lectures critiques du manuscrit. Vos parcours et vos qualités scientifiques sont pour moi un exemple. Mme Su RUAN soyez assuré de ma sincère reconnaissance pour l'honneur que vous m'avez fait en acceptant de juger ce travail en tant qu'examinateur.

Je tiens maintenant à exprimer mes plus vifs remerciements et ma profonde gratitude à mes deux directeurs de thèse, Mr. Manuel DAUCHEZ et Mr. Pierre BEAUSEROY pour avoir cru en moi et m'accueilli au sein de votre laboratoire. Grâce à vos précieux conseils et votre rigueur dans le travail, vous m'avez offert la possibilité d'approfondir mes connaissances scientifiques. Ces trois ans de thèse sous votre direction resteront pour moi une très grande expérience, tant sur le plan professionnel que humain.

J'adresse aussi mes vifs remerciements et ma profonde reconnaissance à mes encadrants de thèse, Mme Edith Gräll, Mr. Rémi COGRANNE et Mr. Sébastien ALMAGRO pour m'avoir accueilli au sein de votre laboratoire. Merci pour votre confiance et vos soutiens durant les trois ans de thèse. Je n'oublierai jamais nos échanges scientifiques, les meetings et tous les moments que j'ai passé avec vous.

Edith, Manuel, Pierre, Rémi et Sébastien, votre disponibilité, vos remarques pertinentes et votre soutien tout au long de cette thèse m'ont permis d'acquérir la rigueur scientifique indispensable, la réalisation de ce travail, ainsi que votre gentillesse et patience ont instauré les meilleures conditions de travail. Merci de m'avoir toujours encouragé et d'avoir toujours été optimiste, surtout pendant les périodes

difficiles.

J'ai pu travailler pendant ma thèse avec des personnes vraiment formidables qui m'ont énormément aidé et qui ont toujours été là pour moi. Je tiens donc à les remercier pour leur soutien et leur esprit d'équipe. Merci Mr. Nicolas BELOY, Mme Stéphanie BAUD et Mr. Laurent DEBELLE pour votre accueil et gentillesse. Merci à tous pour tous ces moments passés ensemble, merci d'avoir fait de ces années passées avec vous une expérience humaine très enrichissante. Vous me manquerez tous et j'espère garder longtemps contact avec vous. Merci à toutes les personnes qui ont participé de près ou de loin à la réalisation de mon projet.

Je tiens à remercier également mes amis libanais au laboratoire et mes anciens colloques qui m'ont soutenu pendant cette aventure. Un remerciement spécial pour mon ami Karim TOUT que j'ai passé avec lui le plus de temps.

À ma famille, notre union est notre richesse-, je vous aime plus que les mots ne peuvent le dire. Tarek, mon petit frère et mon cher ami, merci pour tes encouragements, tes conseils et ton grand support et j'espère avoir été pour toi source d'inspiration comme tu l'as été pour moi et bientôt tu seras un ingénieur. Rayan, ma chère sœur et ma copine merci pour le support moral. Mouemen mon petit frère qui me taquine toujours mais qui me motive à faire toujours mieux. Ayat ma petite sœur, la gentillesse et la beauté de notre famille.

À ma mère et mon père, à mes parents sans qui l'enfant que j'étais ne serait pas devenu l'homme que je suis. C'est avec émotion qu'à mon tour je leur dévoile le fruit de mes efforts. J'espère être à la hauteur de leur fierté inconditionnelle. Votre présence vos encouragements et vos amours sont pour moi les piliers fondateurs de ce que je suis et de ce que je fais. Heureusement que vous êtes là pour me changer les idées. Ils ont tous cru en moi et ouf ! Maintenant j'y suis ! Merci pour tout le bonheur que vous m'apportez. Je vous aime fort ! Le bonheur se trouve sous vos Pieds.

À mon amour Ghadir, que je l'ai rencontré durant la dernière année de thèse et qui m'a partagé avec sa patience la période la plus difficile de mon travail. Tu étais avec ta conscience, ton moral et ton amour la meilleure chose qu'elle a passé avec moi depuis longtemps. Je t'aime beaucoup.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

Acronyme	**Signification**
MCV	Maladies cardiovasculaires.
MEC	Matrice ExtraCellulaire.
CRE	Complexe Récepteur de l'Elastine.
SVM	Support Vector Machine.
ACP	Analyse en Composantes Principales.
SVDD	Support Vector Data Description.
OC-SVM	One class SVM.
Recc	Erreur de reconstruction.
DBSCAN	Density-Based Spatial Clustering of Applications with Noise.
CAH	Classification Ascendante Hiérarchique.
CDH	classification Descendante Hiérarchique.
SOM	Self Organising Map.
RMSD	Root Mean Square Deviation.
SSAP	Sequence Structure Alignment Program.
SRVF	Square-Root Velocity Function.
ICP	Iterative Closest Point.
AUC	Area Under Curve.
COR	Receiver Operating Characteristic.

Liste des symboles

chapitre 3

Notation	**Définition**.
$k()$	Le fonction noyau.
x	Un vecteur de donnée.
y	Un vecteur de donnée.
$\mathcal{D}$	l'espace de départ de données.
$\langle , \rangle$	Le produit scalaire.
ϕ	La fonction de transformation à l'espace caractéristique.
$\mathcal{F}$	L'espace caractéristique de données.
$\mathbb{R}$	L'ensemble des nombres réels.
c	Un scalaire.
σ	L'écart-type de noyau.
f	Une fonction de densité.
$\mathcal{A}$	Une ensemble de donnée dans l'espace.
ς	Un seuil de décision.
w	Un vecteur dans l'espace caractéristique $\mathcal{F}$.
d	Un scalaire qui a deux valeur (-1 ou 1).
$\mathscr{H}$	Un hyperplan dans l'espace caractéristique $\mathcal{F}$.
h	Un vecteur dans l'espace caractéristique $\mathcal{F}$.
τ	Un scalaire.
i ,j, r, l, k	Des indices.
n	Le nombre des observations (des vecteurs).
α	Le multiplicateur de Lagrange.
K	La Matrice de Gramm (du noyau).
g	Le centre du Cercle.
R	Le rayon d'un cercle.
$diag(K)$	La somme de la diagonale de la matrice K.
C	Une variable pour contrôler la proportion de données atypiques.

d	La dimension d'un espace.
D	Une ensemble de données.
X	Une matrice des vecteurs.
$\tilde{x}$	Une donnée centrée.
$\bar{x}$	La moyenne de données (Centre).
$\mathcal{C}$	La matrice de covariance.
$\mathcal{V}$	La matrice des vecteurs propres.
λ	La matrice des valeurs propres.
q	Le nombre des vecteurs propres pris en compte.
Σ	Matrice de covariance de l'espace caractéristique.
v	La matrice des vecteurs propres dans $\mathcal{F}$.
β	Un vecteur des coefficients utilisés pour calculer les vecteurs propres dans $\mathcal{F}$.
$\tilde{\phi}$	La version centrée de la fonction de transformation à $\mathcal{F}$.
$\tilde{k}()$	La version centrée de $k()$.
$\mathcal{W}_q$	La matrice des q vecteurs propres.
d_{ij}	La distance entre i et j.
θ_i, θ_j, ρ, γ	Des scalaires.
m_i	Le centre de cluster i.
k	Le nombre de cluster.
$\mathtt{w}$	Les vecteurs références.
N	Le nombre des observations.
ε	Le rayon de voisinage autour d'une observation x.
$minPts$	Le nombre de voisins pour la méthode DBSCAN.

chapitre 4

Notation	**Définition**
d_{ij}	La distance entre l'observation i et l'observation j.
i, j, k, l, t	Des indices.
P	Un peptide.
$f(t)$	La fonction d'une courbe paramétrée.
N	Le nombre des atomes.
$\mathcal{S}_t$	La structure à l'instant t.
$\mathcal{R}$	La matrice de translation.
$\boldsymbol{\zeta}$	Le vecteur de translation.
$\mathbf{a_i}$	L'atome d'indice i.
$\mathbb{S}$	La séquence temporelle des structures inspectée.
$\mathbb{F}$	Une Conformation.
ε	Un seuil de décision.

$L(t)$	Le label de la structure $\mathcal{S}_t$.
K	Le nombre des conformations principales.
$\mathbb{F}_K$	La K-ième Conformation.
$\mathcal{M}$	La matrice des distances inter-atomes.
T	Le nombre des structures pour un peptide donné.
q	Le nombre de vecteurs propres.
σ	L'écart-type de noyau.
WC	La dispersion intraclasse.
BC	La dispersion interclasse.
$\mathbf{g}_c$	Le centre d'un cluster c.
m	Le nombre des clusters.
$\mathbf{g}$	Le centre de toutes les données.
β	La probabilité de détection.
υ	La probabilité de fausse alarme.
Recc	L'erreur de reconstruction.
λ	Les valeurs propres.
I	Le taux d'Inertie.
C_i	Le i-ième Cluster.
$A^j_{k,i}$	La distance entre la structure $\mathcal{S}_i$ du cluster C_j et le centre du groupe C_k.
ε	Le rayon de voisinage autour d'une observation x.
$minPts$	Le nombre de voisins pour la méthode DBSCAN.
x ,y, z	Les coordonnées des atomes.
A_i	Le i-ième cluster obtenu avec tous les atomes.
B_i	Le i-ième cluster obtenu avec seulement les atomes du backbone.

chapitre 5

Notation	**Définition**
s_i	Un seuil de décision.
C_i	Le i-ième cluster.
i, j, t	Des indices.
k	La taille de la fenêtre pour le mesure du comportement temporel.
x_i	Une observation.
H_0	L'hypothèse indiquant qu'il n'y a pas une rupture.
H_1	L'hypothèse indiquant qu'il y a une rupture.
$\widehat{f_0}$, $\widehat{f_1}$	Deux lois de distribution des données.
ν	La variance de donnée.
m	La moyenne de données.
D_{KL}	La distance de divergence de Kullback-Leibler.

Conf 1	La première conformation principale commune entre les peptides actifs.
f1, f2 ,..	Les conformations principales d'un peptide donné.
fi	Correspond au cluster labellisé "i" pour un peptide donné.

Chapitre 1

Introduction générale

1.1 Contexte général

Depuis l'aube de l'humanité, les êtres humains s'intéressent tout particulièrement à la compréhension du fonctionnement de leur corps et de ses constituants. Et pour cause ! Mieux connaître son corps s'est pouvoir mieux le réparer et surtout anticiper la survenue de certains problèmes. Dès l'antiquité, le système cardiovasculaire a été identifié, et le lien entre le sang et la vie a été matérialisé sous la forme d'un ensemble d'organes et de structures qui communiquent. Depuis cette époque, les progrès ont été incessants dans la compréhension du système cardiovasculaire et pourtant encore beaucoup de choses restent à découvrir. La preuve en est, hélas, le nombre de morts qui, chaque année, décèdent de maladies cardiovasculaires.

Les maladies cardiovasculaires (MCV) sont la première cause de mortalité dans le monde. En 2008, $17,3$ millions de personnes dans le monde sont mortes, soit 30% de l'ensemble des décès. En 2030, ce nombre pourrait atteindre les 23 millions. Les MCV représentent donc un enjeu majeur de santé publique. Il est donc clair qu'il est avant-tout nécessaire de comprendre les mécanismes moléculaires et cellulaires qui régissent l'activité et le vieillissement du système cardiovasculaire afin de trouver des solutions efficaces (et peu chères) pour mieux soigner.

En général, les artères sont les acteurs principaux de ces maladies. En effet, elles font parties des éléments les plus exposés que ce soit en terme de débit ou en terme de contraintes mécaniques. La moindre inflammation d'une artère peut affecter directement le fonctionnement du cœur et provoquer un infarctus. De ce fait, la structure de la paroi artérielle est extrêmement importante, car c'est un élément qui doit fonctionner pendant environ 1 siècle sous une contrainte maximale tout en restant élastique et étanche. L'élasticité artérielle est un concept que peu de gens connaissent réellement, en revanche son opposé, l'hypertension artérielle,

est connue de tous comme étant un sérieux problème de santé. Cette élasticité artérielle est essentiellement due à une molécule élastique, *l'élastine*, que l'on trouve dans beaucoup d'organes élastiques du corps humain (poumons, peau, etc). Comme pour la peau où l'on comprend aisément ce que représente visuellement une perte d'élasticité, la dégradation de l'élastine des parois artérielles pourra entrainer de graves dysfonctions menant à des pathologies sévères.

L'élastine est donc un acteur majeur des parois artérielles et de son intégrité va dépendre la fonction circulatoire. Une altération de l'élasticité génèrera une rigidité artérielle ce qui va affecter le fonctionnement du réseau circulatoire au complet et entrainer un dysfonctionnement qui sera à l'origine de pathologies cardiovasculaires. L'élastine est un polymère d'une protéine (nommée tropoélastine) et, sous sa forme polymérique, est extrêmement insoluble et résistante. La plus grande partie de notre élastine est synthétisée avant l'adolescence. Après cette période, la resynthèse d'élastine est beaucoup moins efficace, que ce soit en terme de quantité ou bien de qualité. La tropoélastine, c'est-à-dire le monomère d'élastine est produite par les cellules et excrété dans le milieu environnant au sein d'un mélange nommé Matrice ExtraCellulaire (MEC). La MEC est un assemblage complexe de molécules, de structures, produites par les cellules pour adapter son environnement (pour s'accrocher par exemple).

Avec l'âge le système cardiovasculaire subit beaucoup de sollicitations durant son activité. Par conséquent, il subit d'énormes changements au niveau structurel et cellulaire et plus précisément au niveau de la matrice extracellulaire. Un de ces changements est dû à la dégradation de l'élastine des parois vasculaires. Où cette dégradation produit des petits fragments d'élastine (i.e. des *peptides d'élastine*) et qui vont être libérés, soit localement là où la dégradation aurait eu lieu, soit de manière diffuse dans l'organisme via la circulation sanguine. Ces peptides ont des propriétés biologiques et peuvent jouer le rôle de signaux captés par d'autres organes par exemple. Récemment, les études ont montré que les structures tridimensionnelles des peptides ont des effets sur leurs propriétés physico-chimiques et par conséquent peuvent être une cause pour l'évolution des quelques maladies cardiovasculaires.

1.2 Objectif de la thèse

Dans ce contexte, le sujet de la présente thèse vise à étudier le vieillissement du système cardiovasculaire en passant par des approches statistiques afin de

comprendre l'effet de dégradation des élastines, la production des peptides et leurs relations avec les maladies cardiovasculaires. Les objectifs de la thèse peuvent être abordés par 2 approches complémentaires :

La première fondée sur des techniques d'imagerie consiste à analyser des images d'échantillons d'élastine. Son but est d'extraire des caractéristiques de la configuration spatiale des structures des lames élastiques et de développer des méthodes statistiques adaptées pour estimer le vieillissement du système cardiovasculaire. Un travail préliminaire a été réalisé sur cette approche, mais il a été interrompu à cause de la mauvaise qualité et le nombre d'images disponibles. Cependant, nous pensons que cette approche a un bel avenir avec des images d'artères de bonne qualité. Ses objectifs sont bien définis et les méthodes de traitement des images et d'apprentissage sont prêtes a être appliquées. Ces travaux pourraient aussi être menés dans des délais assez courts.

La seconde approche sur laquelle se concentre cette thèse repose sur une modélisation moléculaire. Il s'agit d'analyser les mouvements des peptides produits lors de la dégradation de ces protéines d'élastine. Les simulations des mouvements issus de la modélisation moléculaire des différents peptides produits lors de la dégradation de l'élastine sont accessibles sur la base de travaux antérieurs. Ces peptides pourraient être des «signaux moléculaires » qui agissent sur certains organes ou cellules cibles en vue de répondre à une cause possible de la dégradation de l'élastine.

Pour comprendre ces mécanismes, il est, au préalable, nécessaire d'identifier, pour les peptides d'élastine, les conformations, que l'on retrouve fréquemment. Cependant, il existe de nombreux peptides différents produits par de la dégradation de l'élastine. Parmi eux, on note la présence de peptides actifs et de peptides inactifs. Beaucoup des activités de ces peptides sont maintenant identifiées à l'aide de leur interaction avec un récepteur spécifique, nommé Complexe Récepteur de l'Elastine (CRE). Le but donc de ce travail est, en un premier lieu, d'identifier dans ce très grand nombre de molécules, des formes moléculaires redondantes afin de comprendre les « signaux moléculaires » que la dégradation de l'élastine produit et en deuxième lieu, de différencier les peptides d'élastine actifs de ceux qui ne le sont pas.

1.3 Organisation du mémoire

Pour atteindre les objectifs définis ci-dessus, la structure globale de cette thèse comprend les cinq chapitres suivants :

- Le chapitre 2 décrit en détail l'importance du système cardiovasculaire sur la

santé de l'être humain. Il met en évidence le risque de traitements inadéquats aux maladies de ce système. Il nous fait connaître les constituants cellulaires et structurels du système cardiovasculaire afin de comprendre les motivations du sujet de thèse et d'apprécier l'objectif des travaux autre part.

- Le chapitre 3 présente un état de l'art des méthodes de classification non supervisées qui sont proposées pour résoudre le problème de regroupement de données similaires ou la segmentation de données différentes. Deux aspects sont présentés dans ce chapitre. Le premier aborde le problème de la détection des données atypiques et qui peut être rapproché par les méthodes des classifications monoclasse dans notre cas. Les données atypiques sont les structures transitionnelles, et le but est de les supprimer pour éliminer leurs impacts dans le traitement des données typiques. Le deuxième aspect est la classification multi-classes non supervisée. Son objectif est de regrouper les données similaires sans avoir connaissance au préalable du nombre des groupes. Enfin, ce chapitre approfondit la description des méthodes qui s'appliquent aux données non linéairement séparables.

- Le chapitre 4, dans un premier temps, fait un état de l'art des méthodes de classification qui sont souvent utilisées pour traiter des données géométriques structurées comme les structures protéiques. Ensuite, il présente la méthode proposée pour déterminer les conformations principales au niveau de chaque peptide. Puis, ce chapitre aborde les difficultés rencontrées dans ce travail et dont l'origine est la singularité des peptides de l'élastine. Enfin, ce chapitre se conclut par une application de l'approche proposée sur nos peptides et analyse les résultats obtenus en les comparants aux approches précédentes de l'état de l'art pour ce qui concerne la détection des données atypiques et la classification de conformations principales.

- Le chapitre 5 traite le problème de l'identification d'une signature de l'activité des peptides qui constitue l'objectif final de ce travail. Dans un premier temps, les méthodes de comparaison de protéines de différentes tailles sont examinées. Ensuite, la stabilité temporelle de la conformation des peptides a été étudiée dans le but d'exploiter cette propriété pour déterminer le nombre des conformations principales de chaque peptide. Enfin, la proposition d'une méthode pour détecter l'activité de peptides inconnus en utilisant les principales conformations identifiées pour un ensemble limité de peptides actifs et inactifs conclut ce chapitre.

Enfin, le chapitre 6 synthétise les principaux défis et les contributions majeurs de cette thèse puis présente quelques perspectives pour des travaux futurs.

Chapitre 2

Contexte biologique du sujet de thèse

Sommaire

2.1 Introduction

Depuis toujours, l'être humain accorde une attention toute particulière à la compréhension du fonctionnement de son corps et à la découverte de ses constituants. Dès l'antiquité, Égyptiens et Grecs ont tenté de comprendre la relation qu'il pouvait y avoir entre le sang et la mort [13]. De ces premières recherches ou observations, un certain nombre de constatations ont pu mener à la découverte du système cardiovasculaire. Les progrès scientifiques des siècles suivants ont ensuite permis de comprendre plus finement le rôle de chacun des organes (cœur et poumon tout particulièrement) et des vaisseaux sanguins (artère, veines et capillaires) dans cette fonction vitale de l'organisme. En effet, si le symbolisme du sang dans nos cultures est si fort, c'est que le sang est absolument nécessaire à bon nombre de fonctions vitales. Parmi les fonctions principales du sang et du système cardiovasculaire, on trouve le transport des nutriments vers les organes et les cellules, l'acheminement du dioxygène, mais aussi l'élimination des déchets comme le dioxyde de carbone. Ce système circulatoire a comme moteur un muscle - le cœur – qui, par ses battements incessants, va entretenir un mouvement du fluide dans l'organisme. Ce liquide va circuler tout d'abord dans les artères, que l'on pourrait comparer à des "autoroutes sanguines" et qui vont ensuite se ramifier en capillaires pour rentrer plus profondément dans les tissus afin d'arriver au plus près des cellules. Le sang, ou plutôt les constituants qu'il véhicule, une fois consommés, retournent vers les poumons en commençant par des lymphatiques puis rejoignent le réseau veineux (Fig. 2.1). Comme expliqué auparavant et comme connu de tout un chacun, le sang

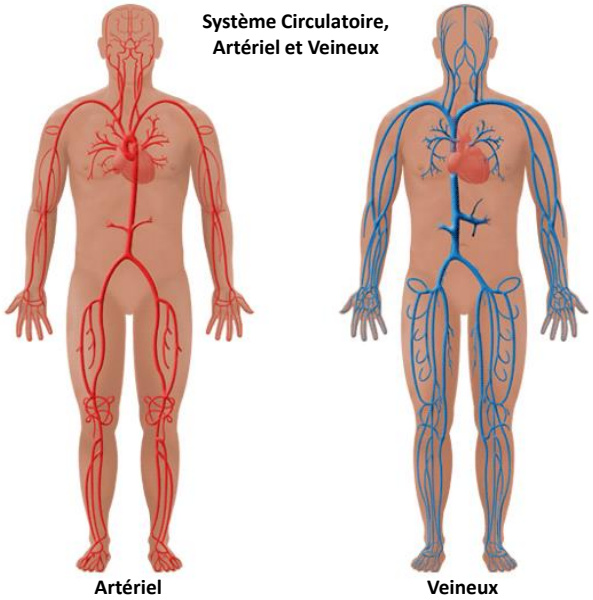


FIGURE 2.1 – Schéma général de l'organisation du système cardiovasculaire : réseau artériel (gauche, en rouge) et réseau veineux (droite, en bleu) [1].

(et donc la circulation sanguine), est vital pour l'être humain. Toute altération de ce système entraînera des problèmes dans le fonctionnement de l'organisme, voire la mort. De plus, le sang est un moyen de communication entre les organes ce qui induit que la majorité des composants que notre organisme assimile qu'ils soient bons (oxygène, eau, sucre, graisses ...) ou néfastes à certaines doses (alcool, drogues, sucre, graisses ...) vont se retrouver dans le sang et seront donc au contact direct du réseau vasculaire. Au fil des années, l'action répétée de diverses molécules circulantes, un taux élevé de sucre dans le sang par exemple, va provoquer un vieillissement accéléré des structures vasculaires en induisant des modifications chimiques provoquant un vieillissement accéléré [14].

2.2 Statistiques sur les maladies cardiovasculaires

Les Maladies Cardiovasculaires (MCV) représentent l'une des principales causes de mortalité à l'échelle mondiale [15], (figure 2.2). En 2008, environ 17 millions de personnes sont mortes d'atteintes cardiovasculaires [16]. 7% de ces décès avaient eu lieu avant 60 ans et pour la plupart, ces décès précoces aurait pu être évités. Le pourcentage de décès prématurés par maladie cardiovasculaire varie de 4% dans les pays à revenu élevé à 42% dans les pays à faible revenu entraînant des inégalités croissantes dans l'apparition, le suivi et l'issue des maladies cardiovasculaires entre les pays et les populations [2], (figure 2.3).

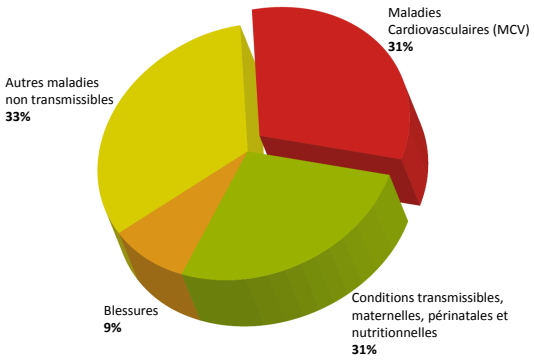


FIGURE 2.2 – Répartition des principales causes de décès, y compris les MCV [2].

Selon l'Union européenne [17], le coût de la prise en charge des maladies cardiovasculaires par nos sociétés pour l'année 2012 a été estimé à quasiment 200 milliards d'euros/ans. Ce chiffre est en augmentation continue dans tous les pays. La même année, en France, 147 000 personnes sont mortes à cause d'atteintes cardiovasculaires. Ce nombre de décès représente le deuxième plus important, juste après celui des cas de décès imputables au cancer [17].

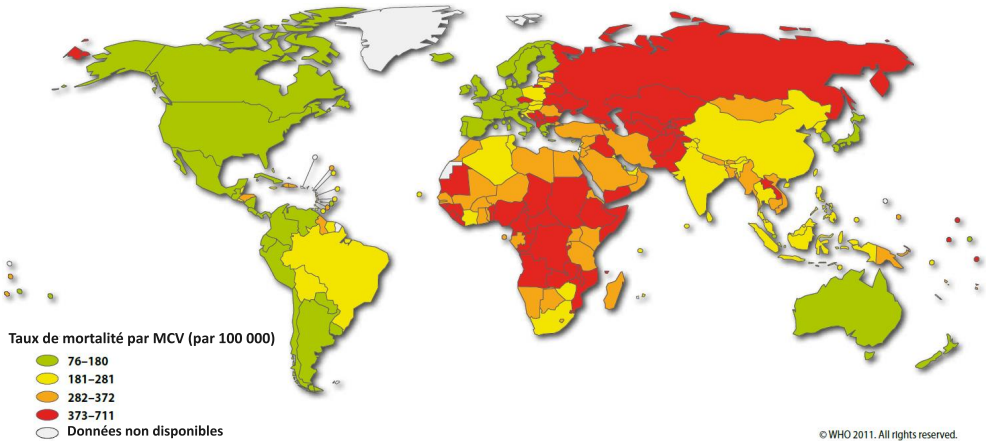


FIGURE 2.3 – Carte du monde montrant la répartition mondiale de taux de mortalité à cause de MCV [2].

Les MCV représentent donc indubitablement un problème majeur de santé publique. Par conséquent, de nombreux acteurs de la recherche fondamentale et clinique, publics et privés, étudient les mécanismes moléculaires et cellulaires impliqués dans les processus physiopathologiques, afin de mieux comprendre les mécanismes qui mènent à l'installation de pathologies souvent chroniques, mais aussi afin de proposer de nouveaux moyens de diagnostics afin de détecter le plus précocement possible la survenue de ces pathologies chroniques [18, 19, 20]. Une détection précoce permet d'influer significativement sur les perspectives de survie, et surtout permet une survie dans le meilleur état de santé possible (ce point représente un impact énorme sur le coût de prise en charge). Dans cette perspective, il est nécessaire d'identifier de nouvelles cibles pharmacologiques qui permettront au praticien de juger au mieux l'état du patient, mais aussi d'inférer sur l'évolution de cet état. Tout ceci rentre dans le cadre d'une approche, voulue par nos sociétés, d'une médecine personnalisée qui s'adapte au mieux au cas de chaque patient.

De nombreuses études montrent qu'un mauvais régime alimentaire, le manque d'activité physique, le tabagisme, la consommation excessive de l'alcool, sont parmi les principaux facteurs de MCV [21]. Parallèlement à ces facteurs considérés comme *extrinsèques*, il existe des facteurs de risque *intrinsèques* tels que le vieillissement psychologique, qui a un impact fort sur le fonctionnement de système cardiovasculaire[21].

En général, ce sont les artères qui sont les segments du réseau qui sont les plus affectés par les MCV, ce qui est tout à fait compréhensible du fait du volume sanguin qui s'y écoule et du fait de la contrainte mécanique (dont l'usure) associée au passage de l'onde de pouls [22]. Une des conséquences de ce phénomène est que

les artères se rigidifient avec l'âge. L'élasticité de l'artère lui permet "d'irriguer" correctement tout le réseau en aval et lorsque la rigidification s'opère, elle affecte le reste du circuit entraînant un dysfonctionnement de certains organes.

La Matrice ExtraCellulaire (MEC) qui entoure les cellules dans les parois artérielles est l'acteur principal de l'élasticité de l'artère, via notamment une protéine polymérique nommée élastine qui rentre dans sa composition. Par voie de conséquence, si l'élastine de la matrice extracellulaire contenue dans la paroi artérielle est détruite, ou tout simplement partiellement dégradée, l'artère perdra de son aptitude à faire circuler le sang dans un régime peu turbulent, générant par la même une augmentation des contraintes mécaniques induisant, par cercle vicieux, une contrainte mécanique supplémentaire générant ainsi de nouvelles dégradations et amplifiant peu à peu les MCV par un mécanisme qui s'autoalimente [23].

Afin de comprendre la problématique de notre travail, il nous faut tout d'abord introduire certaines notions de physiologie en présentant les différents acteurs du système cardiovasculaire et leurs caractéristiques respectives (en termes de structure et de fonction(s)).

2.3 Description des vaisseaux du corps humain

Pour résumer, il existe différents types de vaisseaux sanguins dans le corps humain : les artères, les veines et les capillaires. Ces vaisseaux sanguins servent à guider le sang au travers du corps [24]. On peut considérer les artères comme des structures que l'on qualifiera de "dures, épaisses et élastiques". Ces caractéristiques mécaniques permettent aux artères d'être durable tout en étant déformable afin d'aider à la circulation sanguine.
La plus grosse artère se trouve connectée au cœur et se nomme l'Aorte (Fig. 2.4). Les artères se ramifient en artères plus petites nommées artères secondaires puis en structure plus fines appelées artérioles. Les artérioles se ramifient aussi en structures plus fines appelées capillaires. Les capillaires sont des vaisseaux à paroi mince qui relient les artères, aux organes, puis les organes aux veines [24]. Les veines sont les vaisseaux sanguins qui transportent le sang désoxygéné, chargé en dioxyde de carbone provenant de la respiration cellulaire, l'acheminant vers les poumons pour élimination du CO_2 et récupération de l'O_2. Le système veineux possède des *valvules*, sortes de "clapets anti-retour" qui empêchent le sang de revenir en arrière. Les veines sont en position plus externes, c.-à-d. plus proches de la peau. On peut d'ailleurs observer assez facilement la coloration bleuâtre (= cyanosée) d'une personne qui a des problèmes d'alimentation en oxygène. On l'observe par exemple facilement chez

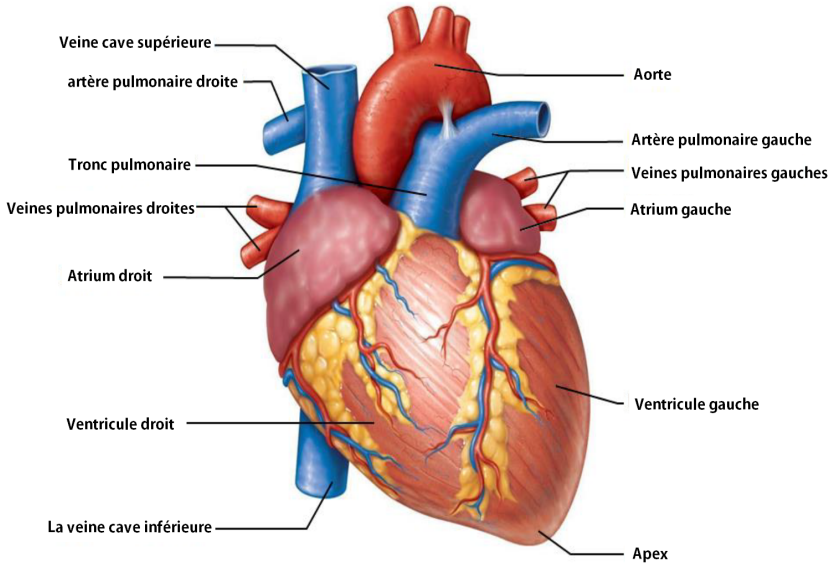


FIGURE 2.4 – Schéma anatomique du coeur humain : la couleur rouge correspond à la partie qui véhicule du sang oxygéné, en bleu, la partie qui véhicule le sang déoxygéné [3]

le nouveau-né lorsqu'il présente des problèmes d'oxygénation.

2.4 Description des structures des vaisseaux sanguins

Les différents types de vaisseaux sanguins varient légèrement dans leurs structures, mais ils partagent globalement les mêmes caractéristiques générales. Les artères et les artérioles ont des parois plus épaisses (et plus élastiques) que les veines, car elles subissent directement l'action de l'onde de pouls et participent d'ailleurs à sa transmission (figure 2.5). Chaque type de vaisseau comporte bien évidemment une lumière (l'espace vide à l'intérieur des vaisseaux appelé *lumen*), et le diamètre de cette lumière dépendra du flux qu'elle véhicule. Les artères ont des lumières de plus petits diamètres que ceux des veines. Ces deux caractéristiques, l'épaisseur des parois et le petit diamètre des artères donnent aux lumières artérielles un aspect plus arrondi en coupe transversale que les lumières veineuses [25]. Les artères et les veines ont les mêmes trois couches de tissus distinctes, appelées tuniques. De la couche la plus intérieure vers l'extérieur, ces tuniques se nomment : l'intima, le média et la tunique externe. Dans les parties suivantes, je décrirai brièvement ces structures et leurs constituants.

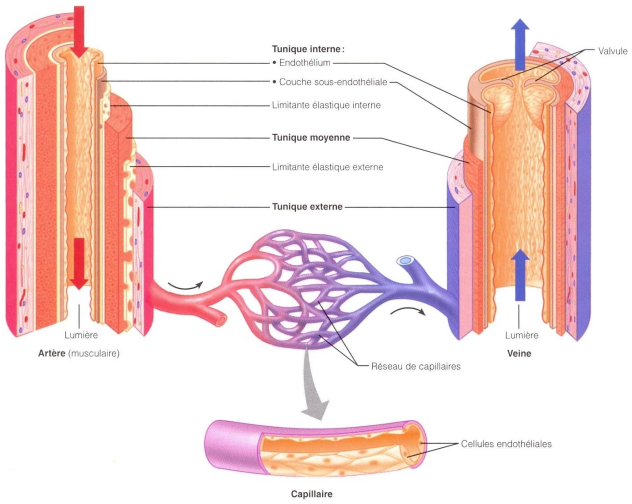


FIGURE 2.5 – Schéma qui montre le structure des artères et des veines [4].

2.4.1 La tunique interne

L'intima est la tunique interne du vaisseau sanguin. Il est composé de 3 couches. Le revêtement interne de la tunique intima formé d'un épithélium vasculaire nommé endothélium (Figure 2.6). La deuxième couche est la membrane basale, une structure matricielle qui lie efficacement l'endothélium au tissu conjonctif qui est la couche externe de l'intima. Cette dernière couche est constituée principalement de fibres élastiques qui confèrent au final une flexibilité supplémentaire à l'intima. L'intima contient également des fibres de collagènes qui permettent d'assurer la résilience générale de la structure. D'ailleurs, dans les artères les plus grandes, il existe également une couche épaisse et distincte de fibres élastiques connue sous le nom de limitant élastique interne à la frontière entre média et intima [26].

2.4.2 La tunique moyenne

Le média est la tunique intermédiaire du vaisseau sanguin. C'est la couche la plus importante (en terme de volume) de la paroi du vaisseau. Cette structure est située au milieu de la paroi vasculaire et se compose de couches de cellules musculaires lisses (Figure 2.6) soutenues par un tissu conjonctif composé principalement de fibres élastiques (donc d'élastine). De plus, une "armature" de fibres collagènes "soude" la média avec ses deux voisines [26] (intima et adventice).

2.4.3 La tunique externe

La tunique externe est un constituant important des vaisseaux sanguins. Elle est formée de tissus conjonctifs composés principalement de fibres de collagène. La tunique externe dans les veines contient également des groupes de fibres musculaires

lisses. Cette structure est normalement la partie la plus épaisse dans les veines et peut être plus épaisse que le média dans certaines artères [26].

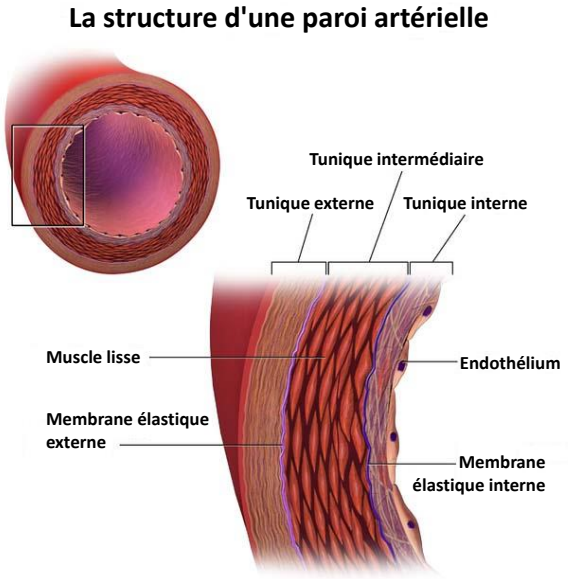


FIGURE 2.6 – Schéma représentant les constituants de la paroi artérielle [5].

2.5 Constituants cellulaires

En général, les parois des artères et des veines sont en grande partie composées de cellules vivantes et de leurs produits, c'est-à-dire de la matrice extracellulaire (MEC) qu'elles ont produite. Cette matrice se présente sous la forme d'un assemblage, covalent ou pas, de macromolécules. Elle est constituée principalement de collagènes, de protéoglycanes, d'élastine et de glycoprotéines de structure. Elle se trouve autour des cellules ou bien leur sert de support. Dans les vaisseaux sanguins, elle constitue la majorité de la matière. Cette MEC joue plusieurs rôles fonctionnels extrêmement importants : l'élasticité apportée par les fibres élastiques (élastine), et la résilience apportée par le collagène. Elle est aussi impliquée dans beaucoup d'autres phénomènes tels que la régulation de processus de signalisation cellulaire ou dans les contacts entre cellules [27, 28].

2.5.1 Le collagène

Le collagène est la principale protéine fibreuse insoluble de la MEC et du tissu conjonctif. En fait, c'est la protéine la plus abondante dans le règne animal. Il existe au moins 16 types de collagène [29]. Les molécules de collagène se regroupent pour former de longues fibrilles minces de structure similaire (Figure 2.7). Les différents collagènes sont en général très résistants et les structures qu'ils forment servent

toutes le même but : aider les tissus à résister à l'étirement [27]. Par exemple, lors d'une déchirure musculaire, un étirement menant à rupture, qui est un évènement que l'on peut qualifier de "violent", ce n'est pas le collagène qui rompt, mais les fibres musculaires qui les relient. Si un type de collagène est déficient génétiquement par exemple, cela peut mener à des maladies sévères comme l'ostéogenèse imparfaite (*c.-à-d.* maladie des os de verre) qui provoque une fragilité osseuse importante [28].

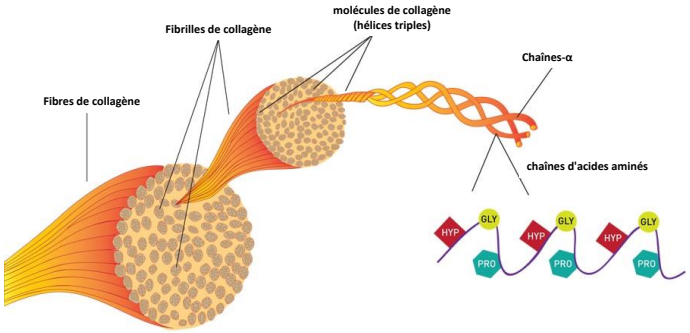


FIGURE 2.7 – Organisation schématique des collagènes [6].

2.5.2 L'élastine et les fibres élastiques

L'élastine est une protéine essentielle de la MEC des vertébrés. En effet, l'élastine est un polymère insoluble, réticulé et non branché d'une protéine, nommée tropoélastine (qui elle est soluble) d'environ 760 acides aminés dans sa forme mature ce qui correspond à une masse d'environ 60 kDa (kilodalton). La plus grande partie de l'élastine chez l'être humain est synthétisée essentiellement avant l'adolescence [28]. La production d'élastine après cette période est très lente, mais surtout mène à la formation d'une élastine moins structurée donc plus fragile et moins élastique. Elle a une demi-vie estimée à environ 74 ans [30]. L'élastine appartient à la famille des protéines élastomériques, protéines capables de se déformer réversiblement sous l'action d'une force extérieure, et ce sans rupture (dans les régimes élastiques) sous l'effet d'importantes contraintes. Son rôle est de conférer élasticité et résilience pour assurer la fonction tissulaire et maintenir son intégrité, notamment pour des tissus subissant des déformations répétitives et importantes (artères, poumons, peau, etc.) [31, 32, 33]. L'élastine constitue 90% des fibres élastiques qui se trouvent dans le média de l'artère. Par ailleurs, les fibres élastiques peuvent prendre plusieurs formes : lamelles concentriques dans les vaisseaux sanguins, réseaux en forme de nid-d'abeilles dans le cartilage élastique et les fibres plus ou moins denses et réticulés dans les poumons, la peau ou les ligaments (Figure 2.8). Les propriétés de l'élastine et du collagène se complètent avantageusement dans le sens où elles rendent le tissu résistant aux déchirures tout en restant déformable. Sans oublier qu'une certaine mémoire de forme, par exemple au niveau de la peau, est aussi assurée [28].

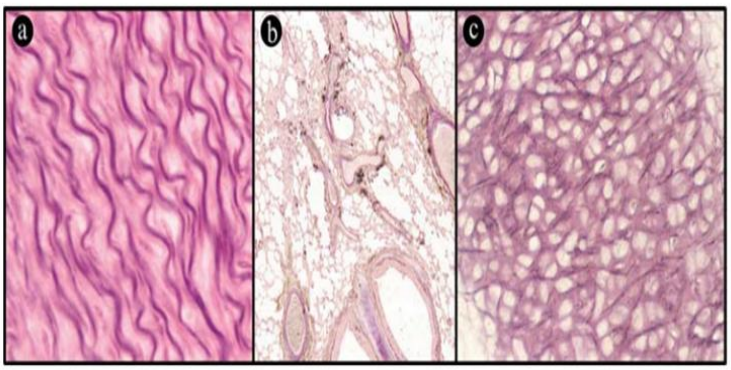


FIGURE 2.8 – Organisation différentielle des fibres élastiques (coloration de Verhoeff Van-Gieson) selon leur localisation tissulaire : (a) Lames élastiques de la paroi d'aortique, (b) Lattices pulmonaire, (c) Nid-d'abeilles dans le cartilage [7].

2.6 Dégradation de l'élastine

Au cours de l'existence, le système cardiovasculaire est soumis à de rudes contraintes dues notamment à la pression sanguine et l'onde de pouls, dépendantes de l'activité de chacun. Par conséquent, ce système circulatoire subit au cours de la vie beaucoup des changements au niveau de sa structure cellulaire, et plus précisément au niveau de la MEC des grandes artères [28].

Avec l'âge, le renouvellement/la réparation de la MEC se ralentit. Ainsi certains phénomènes d'usure commencent à impacter le fonctionnement de cette MEC, c'est ce que nous appelons entre autres le vieillissement de système cardiovasculaire. Un de ces phénomènes est la dégradation de l'élastine de la MEC qui sera quasi-irréversible du fait de sa très faible réparation (en termes de qualité et de quantité). Cette altération fait perdre les propriétés mécaniques et physiologiques de l'élastine. Lors de la dégradation de l'élastine, qui pour rappel est un polymère de tropoélastine, des petits fragments d'élastine (nommés *peptides d'élastine*) vont être libérés dans l'organisme. Ces sous-parties d'élastine, une fois sortie de leur contexte physiologique, vont avoir leurs propres propriétés biologiques et pourront générer une cascade d'événement via la signalisation cellulaire, altérant ainsi le fonctionnement de certaines cellules [34, 28].

2.7 Peptides

Les peptides sont de petits fragments de protéines (moins de 20 acides aminés). Ils forment une chaîne peptidique, c.-à-d. une succession d'acides aminés liés entre eux par des liaisons covalentes nommées liaisons peptidiques [35]. Dès qu'un acide

aminé est inclus dans la composition d'une protéine ou d'un peptide il se nomme résidu. Chaque peptide est formé d'un ensemble de résidus qui interagissent entre eux notamment via leurs chaînes latérales ce qui leur donne des fonctions précises.

2.7.1 Acide aminé

Les acides aminés sont des petits regroupements d'atomes. Ils sont composés de carbone, d'azote, d'hydrogène, et d'une chaîne latérale (R). Chaque acide aminé possède sa propre chaîne latérale. Un acide aminé peut être coupé en trois groupes : le premier groupe est un groupe portant d'une fonction aminé et formé d'un atome d'azote (N) et de deux atomes d'hydrogène (H). Le deuxième groupe qu'on appelle groupement carboxylique est formé d'un atome de carbone, deux atomes d'oxygène (O) et un H . Et enfin, le troisième groupe est la chaîne latérale qui peut avoir des compositions très diverses ; ce groupe est la seule partie qui se différencie entre les différents types d'acide aminé, figure 2.9. Notant que la chaîne formée par les deux atomes de carbone, l'atome d'azote et l'atome d'oxygène constitue la chaîne principale d'un acide aminé et elle est nommée backbone dans la suite de ce manuscrit [36].

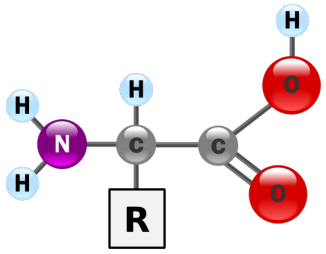


FIGURE 2.9 – Schéma représentant les constituants d'un acide aminé

2.7.2 Structure des protéines

La synthèse d'une chaîne d'acides aminés se fait toujours de la même manière. La connexion entre 2 acides aminés, nommée liaison peptidique, se fait par l'attaque du groupe carboxylique d'un résidu sur le groupe aminé de l'autre résidu, voir figure 2.10. La protéine ou le peptide se structure dans une (ou plusieurs) forme qui sera la résultante de toutes les interactions entre les acides aminés voisins spatialement, interactions qui peuvent être en *cis* (au sein d'une même protéine) ou en *trans* (avec d'autres protéines au voisinage). La forme spatiale de la protéine ou du peptide obtenu lui conférera (ou pas) des fonctions biologiques. Pour une protéine il existe 4 niveaux d'organisation spatiale [37, 36].

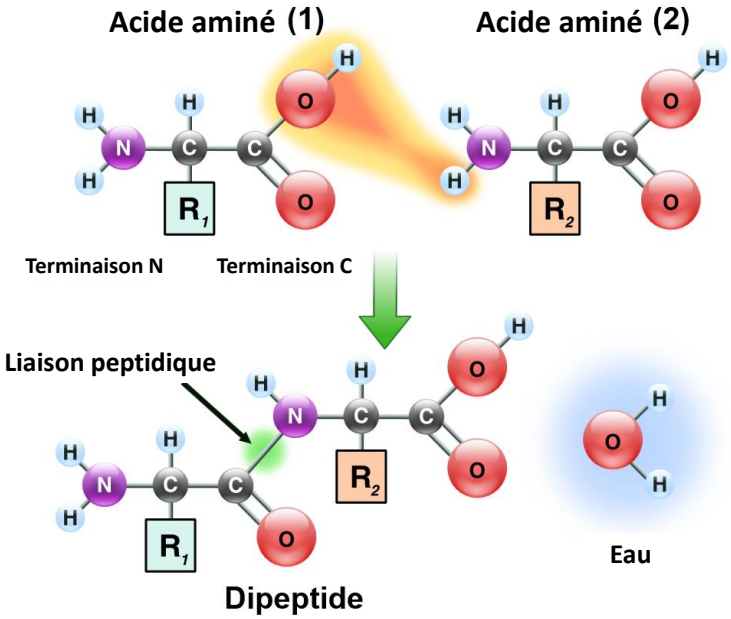


FIGURE 2.10 – Schéma illustrant la synthèse d'un peptide [8].

Structure primaire

Le niveau le plus bas de la structure de la protéine, la structure primaire, est la séquence (succession ordonnée) des acides aminés dans la chaîne polypeptidique. Elle est généralement notée comme une chaîne de caractères, énumérant les acides aminés tels que EGEFPG, avec comme lexique (E= Acide Aspartique, G = Glycine, F = Phénylalanine et P = Proline). Cet ordre des acides aminés n'est évidemment pas aléatoire et est défini par le gène qui code cette protéine [36].

Structure secondaire

Le niveau suivant de la structure de la protéine, la structure secondaire, se réfère au fait que, localement, les acides aminés s'organisent dans l'espace pour créer des structures tridimensionnelles caractéristiques (ex : hélice α, feuillets β) qui vont grandement concourir à la structuration spatiale du peptide [36].

Structure tertiaire

La structure tridimensionnelle globale d'un polypeptide est appelée sa structure tertiaire. Cette structuration correspond à la disposition respective des structures secondaires. Elle est établie via plusieurs types d'interaction chimiques entre acides aminés au sein des structures secondaires et des liaisons covalentes particulières nommées ponts disulfures [36].

Structure quaternaire

De nombreuses protéines sont constituées d'une seule chaîne polypeptidique et ne possèdent que trois niveaux de structure (ceux que nous venons de décrire). Cependant, certaines protéines sont constituées de multiples chaînes polypeptidiques, également appelées sous-unités. Lorsque ces sous-unités se rencontrent, elles donnent au complexe protéique une structure quaternaire qui va correspondre à l'organisation spatiale entre plusieurs exemplaires d'une même protéine ou entre plusieurs protéines différentes. Cet ensemble se nomme un complexe protéique [36].

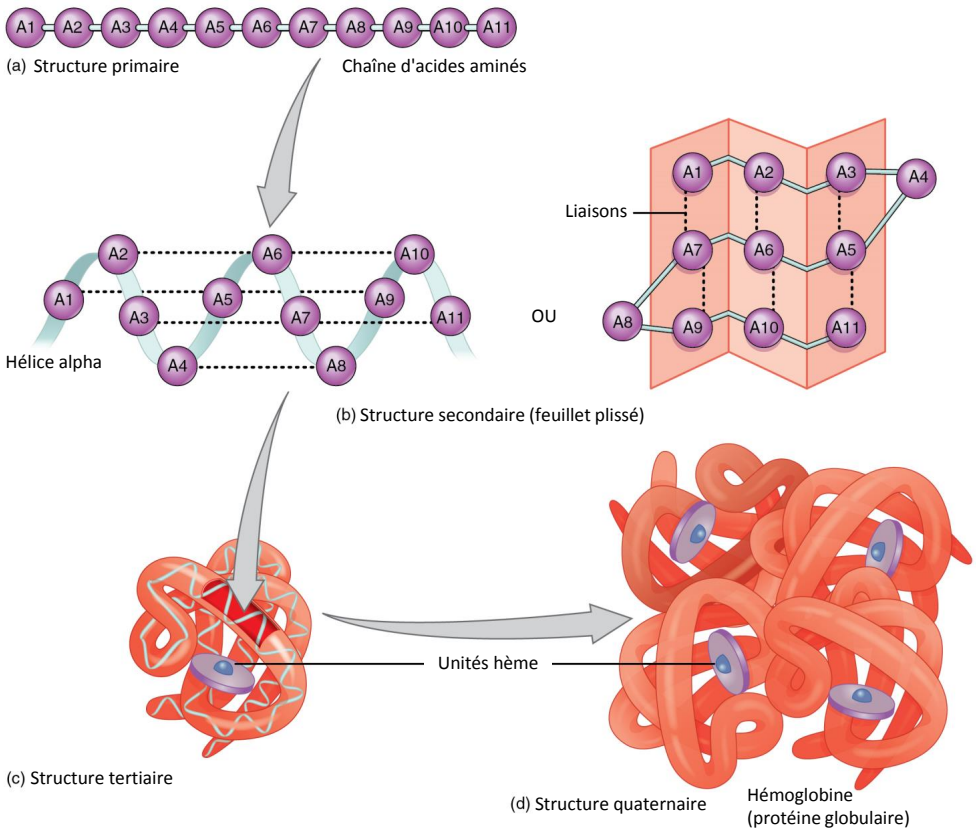


FIGURE 2.11 – Schéma représentant les 4 niveaux de structuration d'une protéine ou d'un complexe protéique [9].

2.8 Maladies cardiovasculaires et dégradation de l'élastine

L'élasticité des vaisseaux sanguins est une des caractéristiques essentielles pour maintenir le bon fonctionnement de système cardiovasculaire. Au fil du temps, malgré que l'élastine soit très résistante, elle est dégradée partiellement par des enzymes nommées élastases. Si cette dégradation est trop importante, la fonction

élastique est tellement réduite que les fibres de collagènes commencent à être sollicitées. Ces dernières étant moins élastiques, les vaisseaux sanguins deviennent plus rigides et moins distensibles [38, 39].
Une autre conséquence est la production des peptides résultant de la dégradation de l'élastine, et qui se retrouve localement concentré, mais aussi dans tout l'organisme via la circulation sanguine [40].
De nombreuses études ont montré qu'il y a une relation entre la quantité de ces peptides circulant et la progression de maladies vasculaires chez des patients atteints par l'athérosclérose, par des pathologies anévrismales ou encore présentant du diabète [30, 41, 42, 43]. Parmi les nombreux peptides produits par cette dégradation, certains d'entre eux possèdent des propriétés biologiques. Ils peuvent être considérés comme des "signaux moléculaires" circulants qui activent potentiellement la réponse de certains organes et cellules, mais ils peuvent aussi influencer l'évolution de pathologies vasculaires [44]. Des approches biologiques notamment en laboratoire indiquent que seulement certains de ces peptides, semblent transmettre un signal et sont considérés comme des "messagers". Mais ils ne sont pas que des messagers, ils peuvent être aussi des *marqueurs* qui indiquent dans quelle mesure la matrice est en train d'être dégradée. À ce jour, la plupart des rôles imputables à ces peptides sont en lien avec leur interaction avec un récepteur spécifique, nommé Complexe Récepteur de l'Élastine (CRE) [40]. D'ailleurs, l'ensemble complet des peptides et leurs rôles respectifs sont encore inconnus. Des expériences en laboratoire ont prouvé que les propriétés biologiques de certains de ces peptides sont liées à leurs structures tridimensionnelles.

2.8.1 Motivation du sujet de thèse

Plusieurs questions se posent et ont motivé ce sujet de thèse. Tout d'abord, les activités des peptides actifs et inactifs (selon leurs interactions avec les CRE) qui sont issus de la dégradation de l'élastine ne sont pas toutes connues, et nous pouvons tout à fait dire que les méthodes actuelles en biologie sont loin d'y arriver. Il serait donc extrêmement intéressant d'être en mesure de prédire les peptides potentiellement actifs à partir des modèles moléculaires qui peuvent être générés *in silico*. Actifs dans le sens bénéfique du terme en activant par exemple un mécanisme de réparation ou bien, inversement, actifs, mais néfastes par les effets qu'ils pourraient produire en bloquant certaines fonctions. De plus, malgré l'avancement de l'analyse statistique sur l'activité des peptides, cette étape reste un défi, au niveau des propriétés physico-chimiques et/ou aux niveaux structurels. En effet, l'instabilité des peptides produits particulièrement par la dégradation de l'élastine, permet d'avoir une grande quantité de conformations dans le temps, impliquant des nouvelles fonctionnalités en raison de leurs interactions avec d'autres cellules. Pour cela, il est crucial de développer

une méthode capable de prédire l'activité des peptides, qui soit simple à appliquer et qui puisse être mise en oeuvre avec les grandes bases de données existantes dans .

2.8.2 Aspect de modélisation moléculaire

Pour une meilleure compréhension du vieillissement cardiovasculaire, et comme indiqué précédemment, les structures spatiales des peptides peuvent être à l'origine de l'activité biologique des peptides. La modélisation moléculaire est une technique *in silico* qui permet de modéliser/simuler le repliement et le comportement d'objet moléculaire au cours de temps tel que des peptides et des protéines. Cette méthode utilisée au laboratoire permet donc de calculer des trajectoires de peptide au cours de temps, trajectoires qui au premier regard semblent être porteuses d'information, mais une information qui semble être noyée au milieu d'un bruit important compte tenu du fait que le peptide change de forme en permanence. Il serait donc extrêmement intéressant d'être capable d'identifier à partir de ces trajectoires, les structures qui seraient les plus répétées donc les plus favorables, car ces dernières sont potentiellement les structures qui pourraient être actives. Il serait de plus intéressant de faire cette recherche, car une méthode *in silico* d'analyse permettant ainsi de générer des centaines de peptides d'élastine possibles et de tous les tester afin de déterminer les peptides susceptibles d'avoir les structures répétitives qui pourrait avoir une fonctionnalité. Le but de l'approche employée est donc de repérer ses formes "clés" à partir d'une collection de trajectoires et de labéliser ces clés en fonction des effets biologiques avérés sur un petit nombre de peptides.

2.9 Conclusion

Pour conclure, nous avons vu les enjeux de santé publique que posent les maladies cardiovasculaires. Ces pathologies chroniques coûtent très cher à la société non seulement en termes de vies, mais aussi en terme financier. Le vieillissement de la population va de plus renforcer ce problème dans les décennies à venir. Ces pathologies sont présentes dans de nombreux cas d'atteintes des structures de l'appareil circulatoire, notamment au niveau des artères. Les parois des artères jouent un rôle crucial dans leur fonction, car leur élasticité assure une bonne circulation sanguine. Les pathologies vasculaires affectent cette fonction notamment en dégradant l'élastine des parois par divers mécanismes affectant ainsi ses propriétés mécaniques. Cette dégradation, en plus d'altérer les propriétés mécaniques, génère des peptides qui vont transmettre un signal (bénéfique et/ou néfaste) dans l'organisme, mais qui sont aussi des marqueurs du degré d'avancement de la destruction de la matrice extracellulaire.
L'approche proposée ici consiste à développer une méthode statistique qui permettrait, *in silico*, à partir de trajectoire de modèle moléculaire, de déterminer quels peptides présentent des structures spatiales clés que le biologiste pourrait ensuite associer à une fonction. En effet, il n'est pas concevable biologiquement de tester toutes les situations par des manipulations et il est donc très important de pouvoir réduire le nombre d'essais pour rendre accessible ce genre d'approche.

Chapitre 3

État de l'art des méthodes de classification

3.1 Introduction

L'apprentissage automatique est un domaine très important des sciences de l'information. Il regroupe l'ensemble des méthodes et des algorithmes qui permettent d'entraîner des machines afin de réaliser des tâches sans être explicitement programmés pour exécuter une tâche donnée [45]. L'apprentissage automatique a beaucoup d'applications et il est présent dans des domaines très variés. Il est utilisé dans les moteurs de recherche, les problèmes d'optimisation, la vision par ordinateur, etc. Ces concepts sont notamment utilisés dans les systèmes de conduite automatisés des véhicules autonomes et dans les moteurs de recommandation commerciaux [46, 47, 48].

La classification est une thématique centrale de l'apprentissage automatique. Elle consiste à entraîner des machines pour regrouper des données selon des critères particuliers. Nous pouvons distinguer principalement deux types de classification. Le premier correspond aux méthodes qui construisent un modèle de décision à partir d'un ensemble de données d'apprentissage dont les sorties sont connues ; il s'agit alors d'un problème d'apprentissage supervisé. Le deuxième type est orienté vers l'exploration de données sans que cela fasse référence à une sortie connue. Dans ce cas, il s'agit d'apprentissage non supervisé.

Dans l'hypothèse de la présence d'un lien fort entre organisation spatiale et fonction de peptide, le problème traité dans ce manuscrit consiste à regrouper les conformations principales qui reviennent fréquemment dans la dynamique du peptide afin de déterminer les caractéristiques des peptides actifs et peptides inactifs. En se basant aussi sur l'hypothèse que le peptide passe d'un état stable à

un autre avec des transitions rapides, le présent travail consiste encore à enlever les conformations correspondantes aux transitions qui ne sont pas très fréquemment présentes et qui ne ressemblent pas aux conformations répétitives. Comme le vrai nombre de conformations principales et des transitions sont inconnus, notre travail est considéré comme un problème de classification non supervisé.

La classification non supervisée peut être abordée de deux façons différentes. En premier lieu, il existe des méthodes consacrées principalement aux problèmes de classification multiclasses et qui ont pris le nom de "clustering". Dans notre cas, elles sont utilisées dans l'exploitation des conformations essentielles de chaque peptide (les conformations répétitives). En second lieu, elle peut s'appliquer à des problèmes de décision à une seule classe, telle que le problème de détection des données aberrantes qui sera aussi traité dans notre travail avec pour objectif d'ignorer l'impact des conformations transitionnelles sur les résultats des classifications.

Ce chapitre sera consacré à la présentation d'un état de l'art sur les méthodes existantes de détection des données aberrantes (structures transitionnelles) et de classification non supervisée multiclasse (conformations principales). La présentation de ces méthodes sera précédée d'une section consacrée aux noyaux qui sont très utilisés pour apprendre les règles de classification quand les données ne sont pas linéairement séparables.

3.2 Concept de noyau

Dans la plupart des applications d'apprentissage automatique, les données sont non séparables linéairement, dans le sens où nous ne pouvons pas séparer les observations associées aux données par des frontières linéaires.

En vertu des propriétés des espaces de Hilbert à noyau reproduisant [49], les méthodes à noyau permettent de réaliser des opérations sur des données déployées dans un espace caractéristique de grandes dimensions (éventuellement de dimension infinie). Le point fort des méthodes à noyau est "l'astuce de noyau", qui permet de déterminer le produit scalaire de deux observations dans l'espace caractéristique des données transformées à l'aide d'une fonction dite "noyau" à partir des deux variables dans l'espace de départ. Cette astuce permet d'effectuer des opérations dans l'espace caractéristique sans définir explicitement les coordonnées des observations dans cet espace [50]. Toute méthode basée sur des produits scalaires peut profiter de ce concept.

Le noyau est une fonction k à valeur réelle qui pour tout couple $(x, y) \in \mathcal{D}$ satisfait

$$k(x, y) = \langle \phi(x), \phi(y) \rangle \tag{3.1}$$

où ϕ est la fonction de transformation des données de l'espace départ $\mathcal{D}$ vers l'espace caractéristique $\mathcal{F}$

$$\phi : x \longmapsto \phi(x) \in \mathcal{F} \tag{3.2}$$

Les méthodes faisant usage de noyaux ont eu beaucoup de succès dans le domaine d'apprentissage [51] et pour cette raison il est très important de souligner leurs principales propriétés :

- Les observations initiales sont transformées d'un espace de départ à un espace vectoriel (espace caractéristique).
- Les produits scalaires entre les observations dans l'espace caractéristique peuvent être calculés par la fonction noyau dans l'espace de départ.
- Les coordonnées des observations dans l'espace caractéristique ne sont pas nécessaires puisque la similarité entre observations peut être calculée dans l'espace de départ.
- Les fonctions linéaires dans l'espace caractéristique correspondent à des fonctions non linéaires dans l'espace de départ. La figure 3.1 nous permet d'illustre par 1 exemple cette propriété.

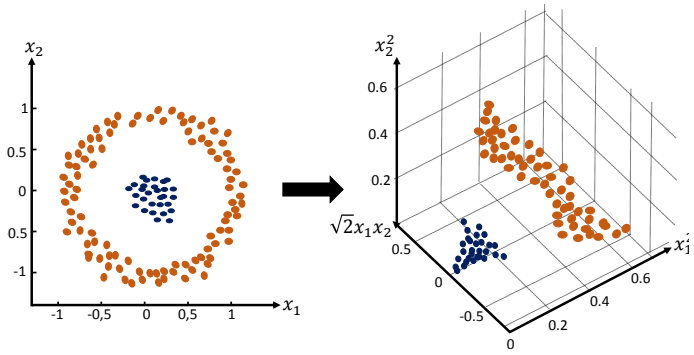


FIGURE 3.1 – Exemple de la transformation de données linéairement non séparable dans l'espace $\mathcal{D}$ (à 2 dimensions) vers l'espace caractéristique $\mathcal{F}$ (à 3 dimensions) où elles deviennent séparables.

Prenant l'exemple de la figure 3.1 pour décrire le fonctionnement de cette "astuce du noyau" : nous considérons que nous avons un espace d'entrée à deux dimensions $\mathcal{D} \subseteq \mathbb{R}^2$ et son espace caractéristique

$$\phi : x = (x_1, x_2) \longmapsto \phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1x_2) \in \mathcal{F} \subseteq \mathbb{R}^3 \tag{3.3}$$

par conséquent, nous obtenons :

$$\begin{aligned}\langle \phi(x), \phi(y) \rangle &= \langle (x_1^2, x_2^2, \sqrt{2}x_1x_2), (y_1^2, y_2^2, \sqrt{2}y_1y_2) \rangle \\ &= x_1^2y_1^2 + x_2^2y_2^2 + 2x_1x_2y_1y_2 \\ &= (x_1y_1 + x_2y_2)^2 \\ &= \langle x, y \rangle^2\end{aligned} \tag{3.4}$$

d'où la fonction de noyau que nous pouvons utiliser dans l'espace de départ

$$k(x, y) = \langle \phi(x), \phi(y) \rangle = \langle x, y \rangle^2 \tag{3.5}$$

La généralisation à des espaces caractéristiques plus riches permet d'utiliser une fonction noyau pour calculer implicitement le produit scalaire d'un espace de dimension éventuellement infini par un calcul n'impliquant qu'un nombre fini de termes [52], où ce nombre est le nombre des observations d'apprentissage qui sont dans l'espace de départ. En outre, pour qu'une fonction à 2 variables soit un noyau, il faut qu'elle vérifie les propriétés du théorème de Mercer (produit une matrice symétrique et semi-définie positive) [53]. Éventuellement, la somme et le produit de noyaux forment un noyau, ce qui donne une liberté dans la création de noyaux [54].

Par conséquent, n'importe quelle méthode de classification, comme le SVM (Support Vector Machine) ou le PCA par exemple [55], qui se base seulement sur le produit scalaire, peut profiter de ces développements qui permettent de traiter des problèmes non linéaires dans l'espace initial.
Deux exemples typiques de noyaux sont souvent utilisés :
Les noyaux polynomiaux :

$$k(x, x') = (c + \langle x, x' \rangle)^p \tag{3.6}$$

où c est la constante du polynôme de degré p.
Le noyau gaussien :

$$k(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}} \tag{3.7}$$

où σ est la largeur de bande de la fonction gaussienne. Le réglage de ces paramètres est souvent très important pour le bon fonctionnement des méthodes à noyaux. Dans ce papier [56] des exemples pratiques pour le réglage et l'évaluation de ces paramètres sont présentés.
Dans notre travail, nous nous concentrons particulièrement sur le noyau gaussien, étant donné qu'il est le plus utilisé pour les méthodes one-classe SVM [57], SVDD [58] et ACP à noyau [59], que nous allons décrire dans les parties suivantes. En plus, des expériences montrent que pour ces méthodes, le noyau gaussien est plus approprié que le noyau polynomial [60, 61, 62].

3.3 Détection des observations atypiques

La détection de données aberrantes est souvent mentionnée dans la littérature scientifique sous plusieurs noms, par exemple : la détection des valeurs aberrantes, la détection de nouveautés, la détection d'anomalies et la détection de déviation, etc. Mais toutes visent le même objectif : identifier les observations qui s'écartent d'un comportement statistique "normal" de l'ensemble de données [63]. Dans notre travail, nous cherchons à distinguer les données typiques des données atypiques dont les conformations principales des peptides sont les données typiques et les conformations transitionnelles sont les données atypiques. En général, une observation atypique est définie comme un échantillon exceptionnellement éloigné de la zone dense de donnée, dans l'espace où se trouvent les données typiques. Dans la pratique, la considération d'une observation comme une observation atypique dépend toujours des experts de domaine qui l'ont acquise.

La détection des observations atypiques a suscité de nombreux travaux de recherche liés à des domaines d'application impliquant de grandes bases de données. Ceux-ci comprennent le traitement de certains problèmes de diagnostic médical [64, 65], d'intrusions dans les systèmes d'informations [66, 67] de vidéo [68, 69], de robotique mobile [70, 71], de réseaux de capteurs [72], ou d'exploration de texte [73]. De nombreuses méthodes pour la détection des données atypiques ont été proposées dans la littérature scientifique [74]. Dans les parties suivantes, nous allons décrire brièvement les plus connu dans le contexte de ce travail.

3.3.1 La méthode de One class SVM

D'un point de vue probabiliste (les observations typiques et atypiques ont des lois de distribution différentes) [75], le problème de détection des données atypiques est lié à la distribution de données typiques qui peuvent être représentées par une fonction de densité f si l'espace d'entrée $\mathcal{D}$ est $\mathbb{R}^d$. Quand il s'agit de définir une région d'acceptation $\mathcal{A}$ pour englober ces données typiques, un bon candidat sera un sous-ensemble de $\mathcal{D}$ sur lequel la densité f prend de grandes valeurs,

$$\mathcal{A} = \{x \in \mathcal{D} \mid f(x) > \varsigma\} \tag{3.8}$$

où ς est un seuil à préciser. Définir $\mathcal{A}$ de cette manière nécessite de connaître la fonction de densité f des données qui est généralement inconnue en pratique. Une solution est d'estimer f en se basant sur un échantillon des données typiques de $\mathcal{D}$: $\{x_1, x_2, \cdots, x_n\}$. Cependant, l'estimation d'une fonction de densité peut être une tâche difficile, surtout lorsque $\mathcal{D}$ est un espace de grande dimension où la malédiction de la dimensionnalité joue pleinement son rôle et qui doit être

contournée au moyen d'hypothèses gaussiennes ou d'hypothèses de parcimonie [76]. Cela étant dit, l'estimation de f n'est qu'un problème intermédiaire pour le problème de détection des données atypiques, car nous nous intéressons uniquement à la localisation des régions de faible densité de $\mathcal{D}$ où se situent, généralement, les données atypiques. Donc même une estimation de f, qui se rapproche faiblement des régions de haute densité, peut éventuellement permettre de détecter correctement les données atypiques. Pour éviter l'estimation de f une approche plus directe est d'estimer une région d'acceptation $\mathcal{A}$ à l'aide d'une fonction donnant directement le résultat de la détection des données atypique [77].

En suivant cette logique, pour définir $\mathcal{A}$ l'équation(3.8) peut être réécrite en remplaçant f par une fonction choisie dépendant de w. En supposant que, dans le cas où les données sont linéairement non séparable dans $\mathcal{D}$, w appartient à un espace caractéristique $\mathcal{F}$ [78] avec un noyau $k : \mathcal{D}\times\mathcal{D} \to \mathbb{R}$ et une fonction de transformation ϕ à l'espace caractéristique $\mathcal{F}$, où pour chaque $x \in \mathcal{D}$, $w(x) = \langle w, \phi(x)\rangle \in \mathcal{F}$. Par conséquent, nous jugeons qu'un point x appartient ou pas à la zone d'acceptation $\mathcal{A}$ des données typiques selon la valeur de :

$$d(x) = \text{sign}\left(\langle w, \phi(x)\rangle - \tau\right) \tag{3.9}$$

Autrement dit, les données typiques et atypiques sont séparées dans l'espace caractéristique par l'hyperplan

$$\mathscr{H} = \{h \in \mathcal{F} \mid (\langle w, h\rangle - \tau) = 0\} \tag{3.10}$$

qui ressemble à la frontière de décision de la méthode de classification supervisée SVM [55] avec la différence qu'ici la plupart des données d'apprentissage sont des données typiques plutôt de deux classes différentes avec deux populations comparables. Ce type particulier de SVM s'appelle One class SVM (OCSVM) et a été introduit par Schölkopf [60]. La méthode SVM n'est pas décrite dans ce manuscrit car elle n'a pas été utilisée au cours de ces travaux.

La méthode OCSVM consiste à séparer la totalité des données de l'origine de l'espace caractéristique $\mathcal{F}$. La Fig.3.2 illustre cet hyperplan qui peut être formulé comme la solution au problème d'optimisation suivant (problème primal)

$$\begin{cases} \underset{w,\xi_i,\tau}{minimiser} & \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i - \tau \quad w \in \mathcal{F},\ \ \tau \in \mathbb{R},\ \ \xi_1,\ldots,\xi_n \in \mathbb{R}_+ \\ \text{sous la contrainte} & \langle w, \phi(x_i)\rangle \geq \tau - \xi_i \qquad x_i \in \mathcal{D},\ i = 1,\ldots,n \end{cases} \tag{3.11}$$

où w est la variable que l'on cherche à optimiser dans $\mathcal{F}$, ξ_i est une variable

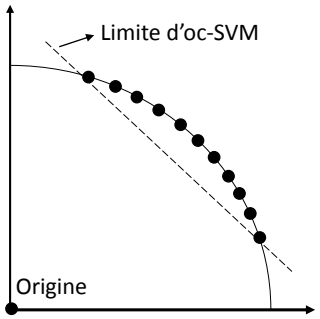


FIGURE 3.2 – (OCSVM) : recherche de l'hyperplan le plus éloigné de l'origine qui contient tous les points d'apprentissage (moins éventuellement les quelques observations atypiques).

introduite pour autoriser la solution à violer certaines contraintes, C est un paramètre qui permet de pondérer l'importance de ces violations dans le problème d'optimisation et τ définit la distance de l'origine. En introduisant le multiplicateur de Lagrange α pour formuler le problème dual [79], le problème initial peut être reformulé de la façon suivante [80] :

$$\begin{cases} \underset{\alpha}{minimiser} & \frac{1}{2}\alpha^T K\alpha & \alpha \in \mathbb{R}^n \\ \text{avec :} & e^T\alpha = 1 & \\ & 0 \leq \alpha_i \leq C & i = 1, \ldots, n \end{cases} \tag{3.12}$$

où K de dimension $n \times n$ est la matrice de Gram, α le vecteur des multiplicateurs de Lagrange et $e = (1, 1, \ldots, 1)^T$. La fonction de décision devient :

$$d(x) = \text{sign}\left(\langle w, \phi(x)\rangle - \tau\right) = \text{sign}(\sum_{i=1}^{n} \alpha_i k(x_i, x) - \tau) \tag{3.13}$$

L'observation est considérée comme une donnée atypique lorsque $d(x) = -1$.

Cette méthode utilise le concept de noyau pour détecter les données atypiques. Elle propose un hyperplan afin de séparer les données mappées de l'origine (dans l'espace caractéristique $\mathcal{F}$) avec une marge maximale et cela nécessite de résoudre un problème avec contrainte quadratique. Ainsi, cette méthode ne prend pas en compte la variance de données. D'ailleurs, il existe une méthode qui a éventuellement le même coût de calcul (de la contrainte quadratique) et qui prend cette variation en compte ; c'est l'ACP à noyau (section 3.3.3). Pour cette raison, la méthode OCSVM méthode n'a pas été employée dans ce travail.

3.3.2 La méthode de Support Vector Data Description (SVDD)

Soit $D = \{x_1, x_2, \cdots, x_n\}$ l'ensemble d'observations qui était considéré avant. L'idée du SVDD proposée par Tax et Duin [81, 58] est de chercher la plus petite hypersphère qui englobe les données typiques lors de l'apprentissage dans l'espace caractéristique $\mathcal{F}$. Elle est définie par son centre g et son rayon R. La figure 3.3 illustre cette approche. Les échantillons restants à l'extérieur de l'hypersphère sont considérés comme des données atypiques. Avec le SVDD, l'objectif est de minimiser le volume de l'hypersphère en minimisant le rayon R ou plutôt R^2. Ainsi, pour éviter une description qui ne représente pas très bien les données typiques, la présence de valeurs atypiques dans l'ensemble de données est permise en introduisant une variable $\xi_i > 0$ pour chaque échantillon d'apprentissage x_i, ce qui permet d'autoriser et de pénaliser les échantillons situés en dehors de l'hypersphère. Cela peut être formulé par le problème d'optimisation suivant :

$$\begin{cases} \underset{R,\xi}{minimiser} & R^2 + C\sum_{i=1}^{n}\xi_i & \\ \text{sous la contrainte :} & ||\phi(x_i) - g||^2 \leq R^2 + \xi_i, & g \in \mathcal{F}, i = 1, \ldots, n \\ & \xi_i \geq 0, & i = 1, \ldots, n \end{cases} \tag{3.14}$$

où C est une variable qui permet de pondérer la pénalisation associée aux échantillons hors de l'hypersphère et par conséquent de contrôler la proportion d'observations en dehors de l'hypersphère.
En introduisant le multiplicateur de Lagrange α, le problème dual du SVDD peut s'écrire de cette façon :

$$\begin{cases} \underset{\alpha}{minimiser} & \frac{1}{2}\alpha^T K\alpha - \frac{1}{2}\alpha^T \text{diag}(K) & \alpha \in \mathbb{R}^n \\ \text{avec :} & e^T\alpha = 1 & \\ & 0 \leq \alpha_i \leq C, & i = 1, \ldots, n \end{cases} \tag{3.15}$$

K est la matrice de Grmm $n \times n$ avec un terme général $K_{ij} = k(x_i, x_j)$, et $e = (1, 1, \ldots, 1)^T$.
Il est clair que le dual de SVDD est le même que celui de OCSVM à l'exception d'un terme linéaire en plus. De plus, si $k(x_i, x_i) = constant$ nous retrouvons l'OCSVM de base. Cela implique que les deux méthodes ont quasiment la même façon de classifier les données. La figure 3.3 illustre l'hyperplan de OCSVM et la sphère de SVDD et nous montre la coïncidence entre les solutions de ces deux méthodes quand le noyau utilisé satisfait $k(x_i, x_i) = constante$ et c'est le cas avec le noyau gaussien.
Une observation x_i est considérée normale si $||\phi(x) - g|| \leq R^2$ ou

$$k(x, x) - 2\sum_{i=1}^{n}\alpha_i k(x_i, x) + \sum_{i,j=1}^{n}\alpha_i\alpha_j k(x_i, x_j) \leq R^2 \tag{3.16}$$

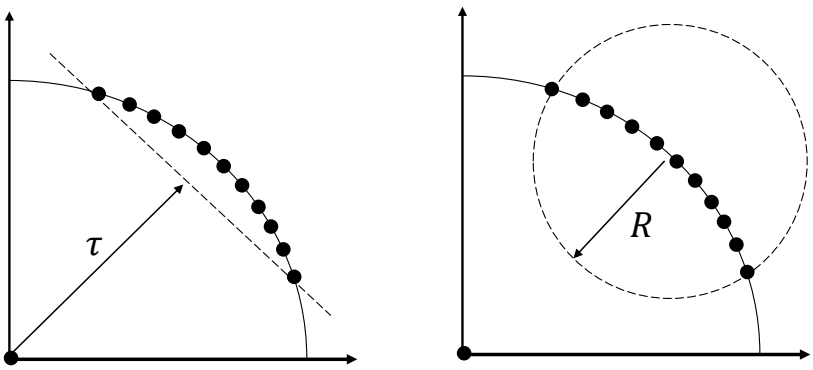


FIGURE 3.3 – Descriptions de données séparées par l'OCSVM et le SVDD.

3.3.3 La méthode ACP à noyau

ACP

La méthode ACP (Analyse en Composant Principal) a été introduite par Karl Pearson en 1901 [82]. C'est une méthode statistique utilisée pour réduire le nombre de variables et rendre les informations portées par ces variables moins redondantes. Fondamentalement, elle vise à transformer les variables corrélées en nouvelles variables non corrélées et qui maximisent la variance captée par les nouvelles variables. Ces nouvelles variables sont choisies itérativement en maximisant la variance des projections le long de l'axe retenu, sous la contrainte que ces nouvelles variables soient orthogonales entre elles. Le même résultat peut être obtenu en minimisant l'erreur de reconstruction (voir figure 3.4), c'est-à-dire la distance carrée entre les données originales et leurs projections dans le sous-espace défini par les nouvelles variables (Éq.(3.19)). La Fig. 3.4 illustre cette méthode pour un cas simple.
En pratique, considérant l'ensemble de données D constitué de n observations $x_i \in R^d$ associés à la matrice : $X = (\tilde{x}_1, \cdots, \tilde{x}_n)$ avec $\tilde{x}_i = x_i - \bar{x}$ la matrice des données centrée, où $\bar{x} = \sum_{i=1}^{n} x_i/n$. Afin d'obtenir les nouvelles variables appelées composantes principales, il faut déterminer les vecteurs propres de la matrice de variance de données $\mathcal{C}$

$$\mathcal{C} = \frac{1}{n} X X^T \tag{3.17}$$

en résolvant cette équation

$$\mathcal{C}\mathcal{V} = \lambda\mathcal{V} \tag{3.18}$$

avec $\mathcal{V}$ la matrice qui contient les vecteurs propres et λ la matrice des valeurs propres qui leur correspondent. Par la suite, le point $\tilde{x}_i$ sera projeté sur les q composantes principales de plus grande valeur propre et aura les nouvelles coordonnées suivantes : $\mathcal{V}_q\tilde{x}_i$, avec $\mathcal{V}_q$ la matrice des q composantes principales de plus grandes valeurs propre et $q \leq d$.

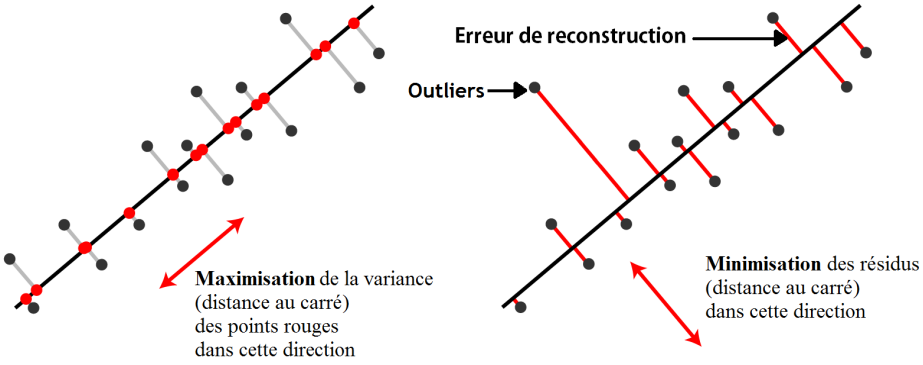


FIGURE 3.4 – Interprétation de l'ACP.

La figure 3.4 illustre aussi l'erreur de reconstruction (Éq.3.19). Dans le cas linéaire, pour une observation $\tilde{x}_i$, $\mathbf{Recc}_q(\tilde{x}_i)$ est calculé de cette manière :

$$\mathbf{Recc}_q(\tilde{x}_i) = \|\tilde{x}_i - \mathcal{V}_q \tilde{x}_i\|^2 \tag{3.19}$$

ACP à noyau

L'ACP à noyau est une extension de l'ACP classique. Elle généralise l'ACP classique pour les cas de données distribuées sur des variétés non linéaires. Récemment, cette méthode a été utilisée pour la détection des données atypiques [59] et a montré une performance très élevée par rapport aux autres méthodes [83, 62, 58]. Elle consiste à transformer les données de l'espace de départ vers un espace caractéristique $\mathcal{F}$ de plus grande dimension à l'aide d'une fonction ϕ comme nous avons vu dans l'équation (3.2). Ensuite, l'ACP est appliquée dans $\mathcal{F}$, et la détection des données atypiques est réalisée à l'aide de l'erreur de reconstruction **Recc** calculée pour chaque observation dans $\mathcal{F}$.

Afin d'appliquer l'ACP dans $\mathcal{F}$ et d'extraire les sous-espaces maximisant la variance des données, il faut trouver les vecteurs propres v associés aux plus grandes valeurs propres λ de la matrice de covariance Σ définie dans l'espace caractéristique $\mathcal{F}$ par :

$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(x_i)\, \tilde{\phi}(x_i)^T \tag{3.20}$$

avec $\tilde{\phi}(x_i)$ les données centrées autour de l'origine ϕ_0 dans $\mathcal{F}$ qui est égal à :

$$\phi_0 = \frac{1}{n} \sum_{r=1}^{n} \phi(x_r) \tag{3.21}$$

Les vecteurs propres et les valeurs propres se calculent en résolvant l'équation sui-

vante :

$$\Sigma v = \lambda v \tag{3.22}$$

À travers cette équation, les vecteurs propres v de la matrice de covariance Σ peuvent être définis comme des combinaisons linéaires des observations transformées $\tilde{\phi}(x_i)$,

$$v = \sum_{i=1}^{n} \beta_i \tilde{\phi}\left(x_i\right) \tag{3.23}$$

où les coefficients β_i sont donnés en résolvant le problème de décomposition propre suivant, profitant de l'astuce de noyau :

$$n\lambda\beta = \tilde{K}\beta \tag{3.24}$$

avec la fonction noyau $\tilde{k}(x, x)$ correspondant à la version centrée de $\tilde{\phi}\left(x\right)$ et qui a un terme général $\tilde{k}(x_i, x_j)$ égal à

$$\begin{aligned} \tilde{k}(x_i, x_j) &= \left\langle \tilde{\phi}(x_i), \tilde{\phi}(x_j) \right\rangle \\ &= k(x_i, x_j) - \frac{1}{n}\sum_{r=1}^{n} k(x_i, x_r) - \\ &\frac{1}{n}\sum_{r=1}^{n} k(x_r, x_j) + \frac{1}{n^2}\sum_{r,s=1}^{n} k(x_r, x_s) \end{aligned} \tag{3.25}$$

En fait, cette matrice de noyau centrée est utilisée dans le problème d'optimisation sans avoir besoin de calculer explicitement la matrice de covariance Σ.
L'erreur de reconstruction pour une observation $\tilde{\phi}\left(x_i\right)$ dans $\mathcal{F}$ sera donc calculée comme suit :

$$\begin{aligned} \mathbf{Recc}_q(\tilde{\phi}(x_i)) &= \|\tilde{\phi}(x_i) - \mathcal{W}_q\tilde{\phi}(x_i)\|^2 \\ &= \left\langle \tilde{\phi}(x_i), \tilde{\phi}(x_i) \right\rangle - \left\langle \mathcal{W}_q\tilde{\phi}(x_i), \mathcal{W}_q\tilde{\phi}(x_i) \right\rangle \end{aligned} \tag{3.26}$$

où $\mathcal{W}_q$ est la matrice des q vecteurs propres associés aux q plus grand valeur propre (v^l représentant le vecteur propre qui a la *l-ième* plus grande valeur propre).
Si nous prenons seulement le l-ième vecteur propre, l'équation 3.26 devient.

$$\mathbf{Recc}_l(\tilde{\phi}(x)) = \left\langle \tilde{\phi}(x), \tilde{\phi}(x) \right\rangle - \left\langle \left\langle \tilde{\phi}(x), v^l \right\rangle, \left\langle (\tilde{\phi}(x), v^l \right\rangle \right\rangle \tag{3.27}$$

avec

$$\begin{aligned} \left\langle \tilde{\phi}(x), \tilde{\phi}(x) \right\rangle &= \left\langle \left(\phi\left(x\right) - \frac{1}{n}\sum_{r=1}^{n} \phi\left(x_r\right) \right), \left(\phi\left(x\right) - \frac{1}{n}\sum_{r=1}^{n} \phi\left(x_r\right) \right) \right\rangle \\ &= k(x, x) - \frac{2}{n}\sum_{i=1}^{n} k(x, x_i) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j) \end{aligned} \tag{3.28}$$

et

$$\left\langle \tilde{\phi}(x), v^l \right\rangle = \left\langle \left(\phi(x) - \frac{1}{n}\sum_{r=1}^{n} \phi(x_r) \right), \left(\sum_{i=1}^{n} \beta_i^l \phi(x_i) - \frac{1}{n}\sum_{i,r=1}^{n} \beta_i^l \phi(x_r) \right) \right\rangle$$

$$= \sum_{i=1}^{n} \beta_i^l \left[k(x_i, x_) - \frac{1}{n}\sum_{r=1}^{n} k(x_i, x_r) - \frac{1}{n}\sum_{r=1}^{n} k(x, x_r) + \frac{1}{n^2}\sum_{r,s=1}^{n} k(x_r, x_s) \right] \tag{3.29}$$

Après évaluation de la norme de l'erreur de reconstruction pour toutes les observations de l'ensemble de données d'apprentissage, un seuil d'erreur $\tau > 0$ est fixé en fonction du nombre prédéfini de données atypiques. Une nouvelle observation est déclarée comme une observation atypique si cette erreur est plus grande que τ.

Il est généralement admis que la zone d'acceptation des données typiques obtenues avec l'ACP à noyau englobe plus étroitement les données d'apprentissage que les régions d'acceptations produites par l'OCSVM et le SVDD qui classent les observations quasiment de la même manière et particulièrement quand le noyau gaussien est utilisé. Ceci est le cas dans la figure 3.3. En effet, ces deux dernières méthodes ne prennent pas en compte la variance de données dans l'espace $\mathcal{F}$, contrairement à l'ACP à noyau. Pour cette raison, l'ACP à noyau fournit une plus petite région d'acceptation qui entoure les données typiques dans $\mathcal{F}$ plus étroitement et donc elle est moins susceptible d'accepter des données atypiques comme données typiques.

3.4 Présentation de clustering

Le clustering désigne l'ensemble des méthodes de classification multi-classes dans un contexte non supervisé. Le clustering est une tâche très importante dans l'exploration des données. Elle consiste à partitionner l'ensemble de données disponibles en sous-ensembles homogènes et bien séparés, appelés classes ou clusters. Par homogénéité, il est attendu que les objets d'un même cluster partagent des caractéristiques statistiques similaires. Par séparation des clusters, il est attendu que les données de clusters distincts soient différentes des objets qui sont dans d'autres clusters.

Les problèmes de clustering ont été étudiés depuis le 18ième siècle avec des applications en psychologie [84], biologie [85] et sécurité informatique [86]. Ce domaine de recherche a aussi pris une grande place dans des domaines tels que la reconnaissance des formes [87], le traitement des images [88] et dans la recherche

d'informations [89] de façon plus générale.

La littérature scientifique concernant le clustering est vaste avec des centaines de méthodes différentes. Dans les sections suivantes, nous décrirons les principales méthodes dans notre contexte.

3.4.1 Mise en œuvre des méthodes de clustering

La figure 3.5 illustre les 4 étapes nécessaires pour appliquer n'importe quelle méthode de clustering pour résoudre n'importe quel problème de classification non supervisée.

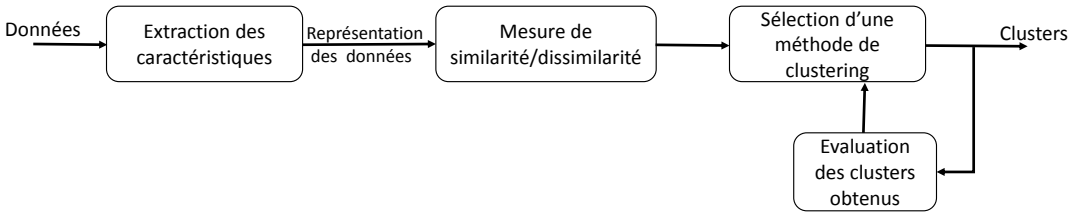


FIGURE 3.5 – Processus de clustering

Extraction des caractéristiques

C'est la tâche d'identification des caractéristiques les plus pertinentes dans les données. En outre, c'est l'étape de la génération des nouvelles caractéristiques en faisant des transformations ou des traitements spéciaux pour affiner les données. C'est une étape très importante et cruciale, car une bonne sélection et extraction des caractéristiques réduira considérablement la complexité du calcul et améliorera l'efficacité de la méthode de clustering. Dans la section 4.4 du 4ième chapitre, nous décrirons les caractéristiques pertinentes considérées dans nos travaux.

Mesure de similarité

Avant de regrouper dans les mêmes clusters les données qui partagent les mêmes caractéristiques, il est important de définir comment mesurer la similarité entre les données. Il existe de nombreuses mesures possibles de similarité dont le choix doit être bien adapté aux caractéristiques et à la population étudiée. Par exemple, la métrique la plus connue et la plus utilisée est la distance euclidienne [90].

$$dist(x_i, x_j) = \left(\sum_{r=1}^{n} (x_{ir} - x_{jr})^2 \right)^{1/2} = \|x_i - x_j\|_2 \tag{3.30}$$

C'est un cas particulier de la métrique de Minkowski avec $p = 2$

$$dist(x_i, x_j) = \left(\sum_{r=1}^{n} (x_{ir} - x_{jr})^p \right)^{1/p} = \|x_i - x_j\|_p \tag{3.31}$$

Il existe plusieurs métriques de distances qui peuvent convenir globalement à presque tous les types de caractéristiques. Nous pouvons les trouver dans [91] et les techniques pour la sélection de ces métriques ont notamment été décrites dans [90].

Sélection de la méthode de clustering

Cette troisième étape consiste à proposer et/ou choisir une méthode de clustering adaptée aux données et aux besoins de l'utilisateur. Plusieurs critères peuvent être pris en compte dans le choix de la méthode de clustering tels que : le nombre d'observations de la base de données, la complexité de la méthode de clustering, les contraintes liées au temps et aux distributions de données [90]. Plus de détails sur les méthodes de clustering seront donnés dans la section 3.4.3.

Évaluation des méthodes de clustering

Cette étape consiste à répondre aux questions : combien de groupes sont cachés derrière les données ? Les résultats obtenus par les méthodes de clustering sont-ils satisfaisants ou pas ? Afin de répondre à ces questions, de nombreuses méthodes existent pour évaluer les méthodes de clustering et pour déterminer le nombre optimal de clusters dans les données [92]. Généralement, ces méthodes d'évaluations permettent aux utilisateurs de juger la qualité des résultats par des indices connus sous le nom "critères d'évaluations de clustering". Ces indices peuvent être regroupés en trois catégories : indices internes, indices externes [93] et indices relatifs [94, 95]. Une étude bien connue a été réalisée par Milligan et Cooper comparant la majorité de ces indices [92]. Dans nos travaux, ces indices ne sont pas pris en compte, car notre objectif était d'étudier la relation entre les clusters plutôt que de connaître le nombre optimal de clusters dans les données.

3.4.2 Dilemmes de l'utilisateur

Malgré la diversité des méthodes existantes, la sélection d'une méthode de clustering appropriée à la résolution d'un problème de classification reste une

difficulté. Plusieurs méthodes ont été proposées pour essayer de comparer les différentes approches de clustering [96], mais jusqu'à présent il n'y a aucune méthode efficace nous permettant de sélectionner de façon fiable la méthode de clustering la plus adaptée pour un jeu de données. En effet, comme expliqué précédemment, l'application d'une méthode de clustering nécessite de trouver une réponse à l'ensemble des questions suivantes :

- Qu'est-ce qu'un cluster ?
- Comment sélectionner les caractéristiques et avec quelle méthode ?
- Quelle est la meilleure métrique pour mon ensemble de données ?
- Quelle méthode de clustering est la plus pertinente ?
- Quel est le nombre réel des clusters ?
- Comment évaluer la qualité des clusters obtenus ?

Les questions relatives au nombre et à la qualité des clusters sont généralement des questions ouvertes et la comparaison entre résultats différents reste un défi. Autrement dit, il n'y a aucune méthode de clustering universelle qui pourrait résoudre tous ces problèmes.

3.4.3 Méthodes de clustering

La figure 3.6 illustre une taxonomie possible des méthodes de clustering [10]. En première approche, les méthodes de clustering peuvent être classées en 4 groupes : les méthodes hiérarchiques, les méthodes par partition, les méthodes basées sur la densité de probabilité et les méthodes qui sont basées sur les grilles (réseau entre des observations) [10]. Il existe des méthodes comme les SOM (self organising maps) qui sont généralement un mélange des autres méthodes (basées sur les grilles et la densité de probabilité) et qui sont utilisées dans beaucoup d'applications [97]. Dans les parties suivantes, les méthodes de clustering, les plus connues et les plus utilisées sont présentées, ainsi que les méthodes avec lesquelles nous avons travaillé.

Méthode hiérarchique

L'objectif essentiel de la classification hiérarchique est de fournir une vue de la structure des données à l'aide d'un arbre de clusters qui représente les niveaux de regroupement entre les observations. La racine de l'arbre est associée à un cluster unique qui contient l'ensemble des données. Puis en descendant dans l'arbre, il se divise pour former des groupes spécifiques qui contiennent des observations considérées comme similaires pour terminer par des observations uniques. La figure

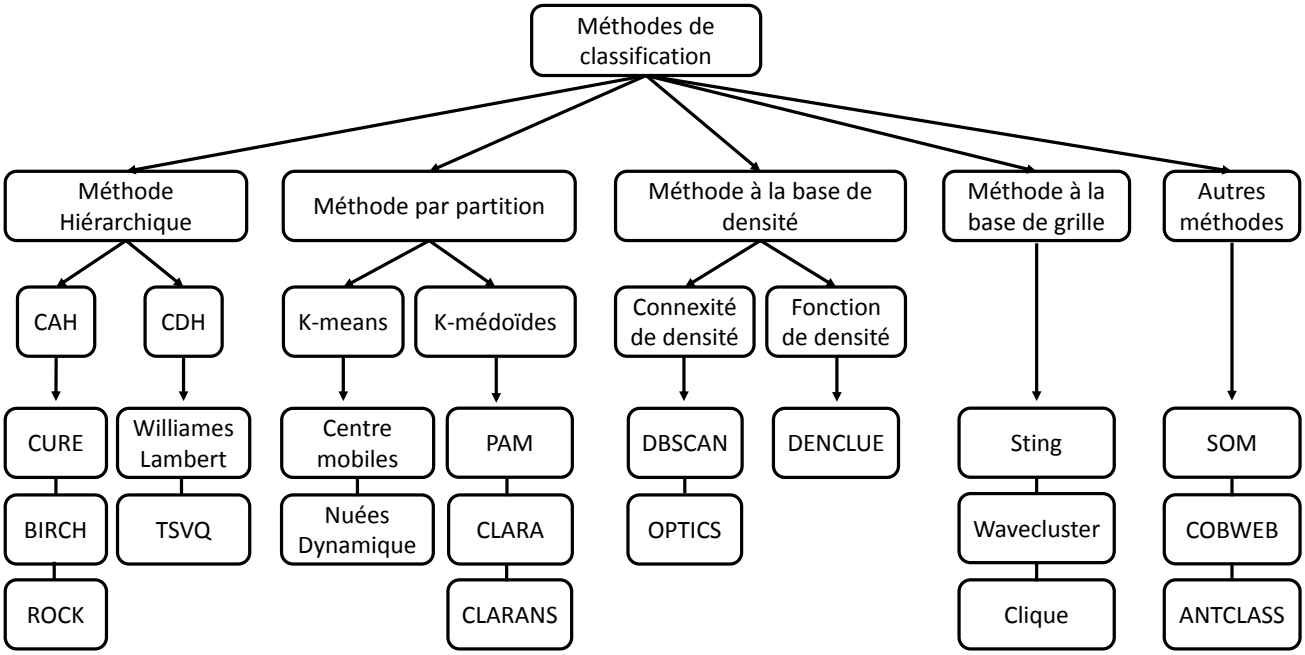


FIGURE 3.6 – Schéma regroupant les différentes méthodes de clustering [10].

3.8 illustre un exemple de ce type d'arbre nommé dendrogramme. En coupant le dendrogramme à différents niveaux, différents regroupements pourraient être obtenus avec différents nombres de groupes. Pour obtenir un dendrogramme, il existe deux approches principales ; selon qu'on part des individus pour les regrouper (Classification Ascendante Hiérarchique : CAH) ou que l'on divise les groupes en partant de l'ensemble des données pour aller jusqu'aux observations isolées (classification Descendante Hiérarchique : CDH).

La méthode de classification ascendante hiérarchique (CAH) est la plus souvent utilisée. Au départ, chaque objet est considéré comme un cluster. Ensuite, la paire des clusters les plus proches est fusionnée dans un nouveau cluster. Cette étape est répétée jusqu'à ce que toutes les observations appartiennent à un seul cluster.

Une tâche très importante dans les méthodes hiérarchiques consiste à définir et à mettre à jour la mesure de similarité/dissimilarité entre les clusters. Il existe 4 stratégies qui définissent différemment cette mesure de similarité. Une formule générale et bien connue a été proposée par Lance et Williams [98] pour définir ces stratégies (Eq.3.32) qui sont déterminées par les paramètres θ_i, θ_j, ρ, γ, d_{hk}, d_{hi}, d_{hj} et d_{ij} explicitées ci-dessous :

$$d_{hk} = \theta_i d_{hi} + \theta_j d_{hj} + \rho d_{ij} + \gamma |d_{hi} - d_{hj}| \tag{3.32}$$

où d_{xy} est la distance entre les deux clusters x et y. h, k, i et j sont les indices des clusters. Pour comprendre cette formule et le fonctionnement de la méthode de CAH, nous notons D la matrice qui contient toutes les distances d_{ij} entre le cluster i et le cluster j, avec n_i et n_j comme nombres des éléments des clusters i et j respectivement. Ensuite, nous procédons comme suit :

- associer chaque individu à sa propre classe.
- trouver la plus petite distance d_{ij} dans la matrice D
- fusionner les deux classes i et j dans une nouvelle classe k
- mettre à jour les distances entre les classes dans la matrice D en se basant sur l'équation (3.32), où h peut être n'importe quelle classe autre que k. Puis d_{hk} remplace d_{hi} et d_{hj} dans la matrice D, et n_k devient $n_i + n_j$.
- revenir à l'étape 2 jusqu'à qu'il n'y ait qu'une seule classe.

Dans la littérature scientifique, 4 mesures de distance entre les clusters sont généralement utilisées [99]. Le tableau 3.1 montre ces distances et les paramètres permettant de les obtenir à l'aide de la formule générale donnée par William [98] :

- distance minimale :
  C'est la plus petite distance qui sépare les deux ensembles des observations classées dans les deux clusters R et S.

$$d_{RS} = min\left\{d(x_{Ri}, x_{Sj})\right\},\ x_{Ri} \in R,\ x_{Sj} \in S \tag{3.33}$$

  où x_{Ri} est l'ensemble d'observations classées dans le cluster R, et x_{Sj} l'ensemble d'observations classées dans le cluster S
- distance maximale :
  C'est la plus grande distance qui sépare les deux ensembles des observations

TABLE 3.1 – Table des paramètres de la formule générale de William pour différentes mesures de similarité entre clusters [12].

Nom	θ_i	θ_j	ρ	γ
distance minimale	$1/2$	$1/2$	0	$-1/2$
distance maximale	$1/2$	$1/2$	0	$1/2$
Distance moyenne	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	0
distance de Ward	$\frac{n_i+n_h}{n_i+n_j+n_h}$	$\frac{n_j+n_h}{n_i+n_j+n_h}$	$\frac{-n_k}{n_i+n_j+n_h}$	0

classées dans les deux clusters R et S.

$$d_{RS} = max\left\{d(x_R, y_S)\right\},\ x_{Ri} \in R,\ x_{Sj} \in S \tag{3.34}$$

où x_{Ri} est l'ensemble d'observations classées dans le cluster R, et x_{Sj} l'ensemble d'observations classées dans le cluster S

- distance moyenne :
  Cette distance consiste à calculer la distance moyenne entre toutes les observations du cluster R et les observations du cluster S

$$d_{RS} = \frac{1}{n_R n_S} \sum_{i=1}^{n_R} \sum_{j=1}^{n_S} d(x_{Ri}, x_{Sj}) \tag{3.35}$$

  avec n_R et n_S le nombre d'individus des populations dans les clusters R et S respectivement.

- Distance de Ward :
  Cette distance a été proposée pour la première fois par Ward [100] et depuis lors elle a été largement appliquée dans l'approche ascendante CAH. À chaque étape de fusion, il s'agit de regrouper les deux clusters qui minimisent l'augmentation de la variance intraclasse (Inertie) (Eq.4.8).

$$d^2_{RS} = \frac{n_R n_S}{n_R + n_S} \|\bar{x}_R - \bar{x}_S\|_2^2 \tag{3.36}$$

  avec $\bar{x}_R$ et $\bar{x}_S$ le centre de gravité de R et S respectivement.

Le choix du critère de fusion est critique, car la partition obtenue en dépend. La figure 3.8 montre le dendrogramme obtenue en appliquant la méthode CAH sur l'ensemble de données de la figure 3.7 en utilisant la stratégie de la distance minimale (pour la simplicité). En général, la distance de Ward est la plus souvent utilisée, et la plus efficace par rapport aux autres distances [101] comme elle prend en compte le nomber et la variation de données durant le regroupement.

Cette méthode de clustering nécessite la mesure de similarités entre toutes les observations, elle a donc une complexité quadratique par rapport au nombre d'observations. Sa complexité dépend aussi de la métrique choisie, pour mesurer les distances entre les classes [102]. Cependant, cette méthode a le grand avantage d'illustrer les niveaux de similarités entre les observations qui est très utiles dans de cas où l'utilisateur a besoin de voir différents niveaux de clustering comme le notre (voir section 4.6).

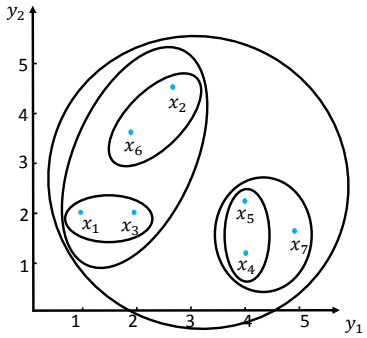


FIGURE 3.7 – Distribution d'un exemple de donnée en 2D.

Méthode hiérarchique à noyau

Comme nous avons vu dans la section consacrée aux concepts de noyau (3.2), l'application des méthodes linéaires dans l'espace caractéristique conduise à des solutions non linéaires dans l'espace d'origine grâce à l'astuce du noyau. Il suffit de définir la matrice de noyau K dans l'espace d'origine de données, puis calculer les distances (distance euclidienne) entre les observations dans l'espace caractéristique qui peuvent être définies dans l'espace d'origine comme ceci :

$$d(\phi(x_i), \phi(x_j)) = \|\phi(x_i) - \phi(x_j)\|^2 = \sqrt{K(x_i, x_i) + K(x_j, x_j) - 2.K(x_i, x_j)} \quad (3.37)$$

Ensuite, une procédure de classification hiérarchique standard s'applique aux données pour les regrouper et obtenir un arbre illustrant la relation entre les clusters résultants. Cette méthode peut être utilisée en association avec la mesure de distance de Ward car comme le montre l'équation (3.37), cette distance peut aisément se calculer dans l'espace caractéristique.

Kmeans

Les méthodes de classification par partition produisent une partition des données sans hiérarchie. En général, elle regroupent les observations de données en

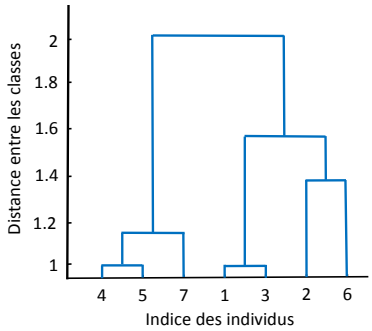


FIGURE 3.8 – Dendrogramme correspondant aux données de la figure 3.7.

optimisant une fonction "objectif" (Eq. 3.38). Kmeans est la méthode de clustering par partition la plus connue. Elle a une histoire riche et diverse étant donné qu'elle a été découverte dans différents domaines scientifiques par Steinhaus [103], Ball et Hall [104] et MacQueen [105]. Bien qu'elle ait été initialement proposée il y a plus de 50 ans, elle reste l'une des méthodes les plus utilisées pour le clustering. Sa facilité de mise en œuvre, sa simplicité et son efficacité sont les principales raisons de sa popularité [106].

Cette méthode a besoin d'un seul paramètre (le nombre de clusters) pour effectuer la classification. Elle produit le groupement des données avec le plus de similarité intraclasse [99]. Cependant, elle ne fonctionne pas qu'avec les données linéairement séparables [107]. La méthode Kmeans commence par une sélection aléatoire de k observations en tant que centres des clusters. Puis, chacune des observations restantes est assignée au plus proche cluster en se basant sur les distances entre l'observation et les centres des clusters. Après chaque classification, les centres des clusters sont recalculés. Ce processus se répète jusqu'à la convergence des centres. Le critère de convergence communément utilisé est associé à la somme de l'erreur quadratique de toutes les observations définie comme suit :

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - m_i\|^2 \tag{3.38}$$

où x est une observation et m_i est le centre du cluster C_i. L'objectif de ce critère est de produire des clusters assez compacts.

Une problématique très importante dans l'application de la méthode du Kmeans est l'identification du nombre de clusters dans les données. La plupart du temps, ce nombre n'est pas connu. Une solution pour trouver ce nombre consiste à appliquer la méthode Kmeans avec différent nombres de clusters (k). Ensuite nous choisissons la solution qui nous offre les meilleures performances (le minimum de variance intraclasse et le maximum de variance inter-classe [99]) pour le jeu de données considéré [108]. Parfois ce test ne répond pas à nos besoins. En effet, Kmeans est sensible à l'initialisation des centres qui peuvent conduire à la convergence vers un minimum local pour le critère (Éq. (3.38)). En général, cette méthode ne produit que des clusters de forme sphérique et elle n'est pas capable de gérer des clusters avec différents nombres d'individus [109]. Cependant, elle est compétitive au niveau de sa complexité qui est linéaire par rapport au nombre des observations de données traitées et reste toujours une solution acceptable pour de nombreux problèmes.

Kmeans à noyau

Comme nous avons vu précédemment, Kmeans a trois problèmes principaux qui limitent son application : 1) l'identification du nombre de clusters, 2) l'initialisation des centres des clusters et 3) il ne fonctionne pas sur des données non linéairement séparables. Une extension de la méthode Kmeans [110] avec le concept de noyau a été proposée afin de résoudre ces deux derniers problèmes.

Concernant le problème d'initialisation des centres des clusters, cette méthode propose de trouver une solution presque optimale pour (M)-clusters en partant d'une solution presque optimale pour $(M-1)$-clusters, dont l'initialisation du M-ième cluster est basé de manière appropriée sur une recherche locale dans les données. Au cours de la recherche locale, le M-ième cluster est initialisé plusieurs fois (plus précisément N fois, où N est la taille de l'ensemble de données) et la solution qui a la plus faible variance intra-classe (Éq. (3.38)) va être la solution pour (M)-cluster. Dès que la solution optimale pour le problème d'un (1)-cluster est connue, la procédure ci-dessus peut être appliquée de manière itérative pour trouver une solution presque optimale au problème de classification de l'ensemble de données.

Supposant que nous avons l'ensemble de données $D = \{x_1, \ldots, x_N\}$ avec x_i appartenant à $\mathbb{R}^d$, nous voulons les classifier à M-clusters. La méthode procède comme suit :

Étape d'initialisation :
Nous appliquons l'algorithme de Kmeans avec $k = 1$

Dans ce cas, nous aurons un seul cluster avec un centre qui égale la moyenne de toutes les observations.

Étape d'itération :
Nous appliquons kmeans avec $k+1$ clusters

- Les k premiers centres sont toujours ceux issus de l'itération précédente.
- Et le $(k+1)$-ième centre=x_i avec $i = \{1, \ldots, N\}$

La solution optimale est celle qui nous donne le moins de variance intra-classe définies dans l'équation (3.38).

Pour résoudre le problème de k-cluster, nous prenons donc les $(m_1, m_2, \ldots, m_{k-1})$ centres des $k-1$ clusters et nous effectuons ensuite N exécution de kmeans en prenant $(m_1, m_2, \ldots, m_{k-1}, x_n)$ comme centres pour $n = \{1, \ldots, N\}$. À la fin nous prenons la solution qui nous donne la plus petite valeur du critère à optimiser (Éq.

(3.38)). Ce processus est répété jusqu'à l'obtention d'un nombre de classes k égal au nombre désiré M. Cette méthode est certe complexe mais permet de s'affranchir du problème d'initialisation des centres.

Pour résoudre le troisième problème lié aux données non linéairement séparables, la méthode propose d'intégrer le concept de noyau dans la méthode kmeans. Elle consiste à effectuer le calcul de distance de l'équation (3.38) dans un espace caractéristique en utilisant l'astuce de noyau définie dans la section 3.2.
Considérons que m_k est le centre du $k-$ième cluster dans l'espace caractéristique $\mathcal{F}$. Nous ne pouvons pas calculer ce centre directement, car la fonction ϕ n'a pas d'expression explicite :

$$\begin{aligned} E &= \sum_{k=1}^{M} \sum_{i=1}^{n_k} \|\phi(x_i) - m_k\|^2 \\ \text{où } m_k &= \frac{\sum_{i \in C_k} \phi(x_i)}{n_k} \end{aligned} \tag{3.39}$$

avec M est le nombre de clusters, n_k est le nombre des observations de le $k-$ième cluster et x_i une observation du cluster C_k.

Cependant, il est possible de profiter de l'astuce du noyau pour calculer la distance $\|\phi(x_i) - m_k\|^2$ en utilisant la matrice du noyau K dans l'espace d'origine,

$$\|\phi(x_i) - m_k\|^2 = K_{ii} - \frac{2}{n_k} \sum_{j=1}^{n_k} K_{ij} + \frac{1}{n_k^2} \sum_{l,j=1}^{n_k} K_{jl} \tag{3.40}$$

De cette façon, il est possible de contourner la seconde limitation du Kmeans.

Carte auto-organisatrice

Cette méthode est inventée par le professeur Teuvo Kohonen [111] pour aider l'utilisateur à visualiser et comprendre ces données qui ont des grandes dimensions. Elle utilise le concept de réseaux neuronaux auto-organisés [112, 113] pour réduire les dimensions des données. Son utilité a été montrée dans de nombreuses applications [114, 115]. Cette méthode consiste à projeter les données de l'espace d'origine vers un espace de faible dimension, généralement 2 dimensions. Elle restitue les données sur une carte composée des neurones lié entre eux selon un principe de proximité. La figure 3.9 représente un exemple de cette carte (pour la visibilité, dans cette figure, la dimension de l'espace d'origine est égale à deux, mais souvent ce n'est pas le cas 2). Chaque neurone, selon sa position dans la carte, est lié par un arc à trois ou quatre voisins. Le but essentiel de

cette méthode est d'illustrer la similarité entre les données de l'espace d'origine en regroupant les neurones similaires dans la carte bidimensionnelle (voir figure 3.9).

Comme précédemment, considérons l'ensemble de données D constitué de n observations $x_i \in R^d$. Pour chaque neurone de la carte, nous associons aléatoirement un vecteur de référence $\mathtt{w}$ de l'espace d'origine (espace d'entrée). En notant N le nombre total des neurones de la carte, le vecteur référent du neurone l est reconnu par $\mathtt{w}^l$, avec $l \in \{1, 2, \ldots, N\}$ et $\mathtt{w}^l \in \mathbb{R}^d$. L'objectif de cette méthode est de mettre à jour les vecteurs référents $\mathtt{w}^l$ de manière à avoir une meilleure description de la distribution de données d'entrée tout en organisant les neurones de la carte de sortie. L'apprentissage de la carte se fait selon deux méthodes ; nous allons décrire ici celle qui est la plus utilisée et qui est nommée "mode séquentiel". Dans ce mode d'apprentissage, chaque itération t comprend deux étapes :

La première étape consiste à prendre aléatoirement une observation x_t de l'ensemble de données D et de trouver son plus proche neurone. Cette méthode donne le nom "neurone vainqueur" ($\mathtt{w}_t^c$) au neurone qui a le plus proche vecteur référent à x_t. Ce neurone vainqueur est défini par :

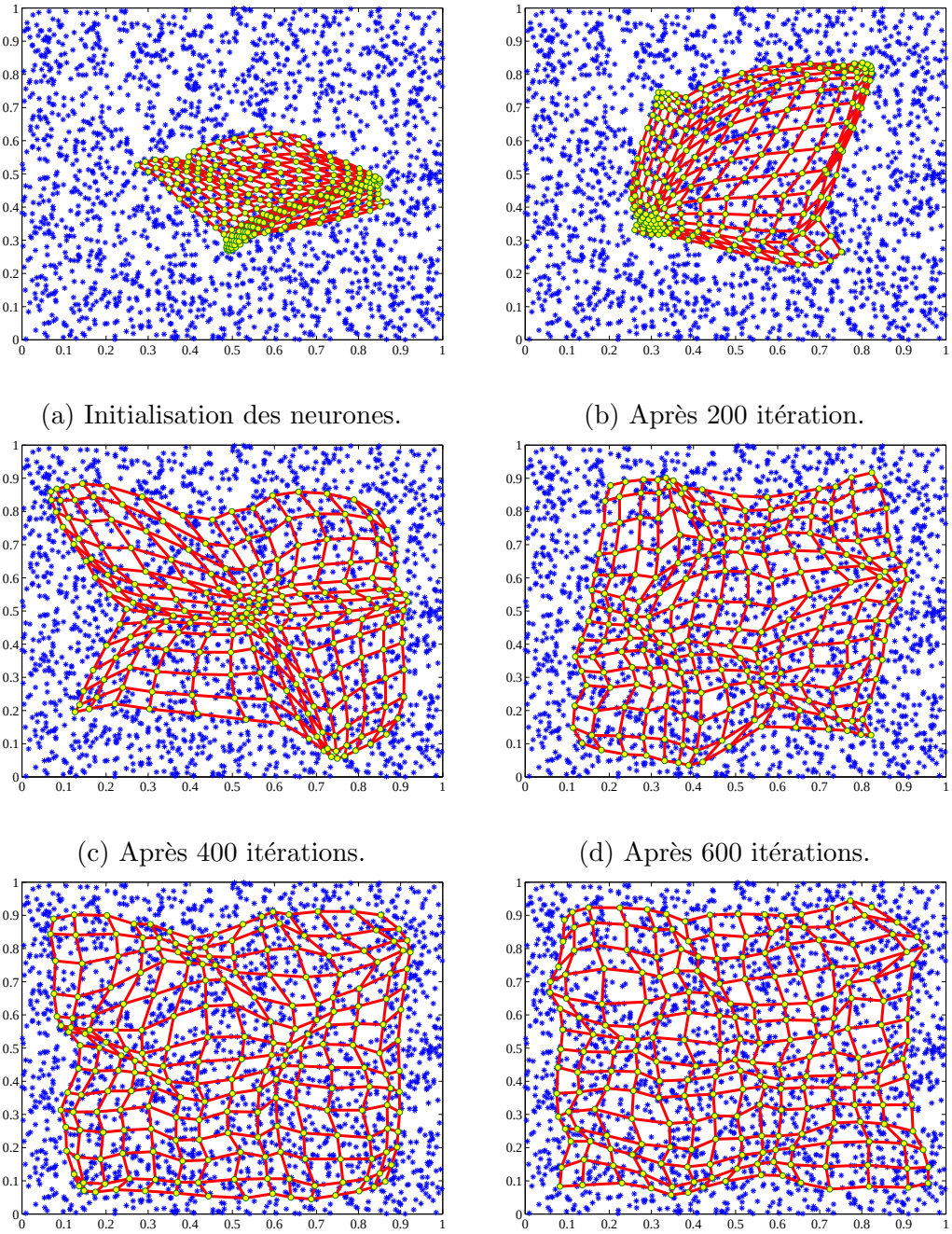
$$d(\mathtt{w}_t^c, x_t) = \min_{l \in \{1,\ldots,N\}} d(\mathtt{w}_t^l, x_t) \tag{3.41}$$

La deuxième étape consiste à mettre à jour le neurone vainqueur. Une mise à jour dans le sens où le vecteur référence du neurone vainqueur s'aligne un peu sur la direction de l'observation x_t. cette mise à jour est aussi appliquée aux neurones voisins du neurone vainqueur. Par conséquent, les vecteurs référents $\mathtt{w}$ de ces voisins sont aussi ajustés de façon relative à leurs positionnements dans la carte bidimensionnelle. Cet ajustement est défini par :

$$\begin{aligned} \mathtt{w}_{t+1}^l = \mathtt{w}_t^l + \alpha_t h_t^{cl} \left[x_t - \mathtt{w}_t^l\right] \\ \text{avec } l \in 1, \ldots, N \end{aligned} \tag{3.42}$$

où α_t est un coefficient pour ajuster la vitesse d'apprentissage, c est le neurone vainqueur de l'observation x_t, et h_t^{cl} est la fonction de voisinage qui définit la proximité entre le neurone vainqueur c et ces l neurones voisins.

La fonction h_t dépend de la position du neurone sur la carte et d'un rayon de voisinage. Dans les premières itérations, le rayon de voisinage est grand pour mettre à jour un grand nombre des neurones qui sont voisins au neurone vainqueur. Avec les itérations, ce rayon décroît progressivement pour qu'il ne contienne finalement que le neurone vainqueur.
Plusieurs fonctions de voisinage sont utilisées dans cette méthode, mais la plus



(a) Initialisation des neurones.

(b) Après 200 itération.

(c) Après 400 itérations.

(d) Après 600 itérations.

(e) Après 800 itérations.

(f) Après 1000 itérations.

FIGURE 3.9 – Illustration de la progression de la méthode de la carte auto-organisatrice sur plusieurs étapes.

populaire est la fonction gaussienne qui est définie par :

$$\begin{aligned} h^{cl}(\sigma_t) &= exp(-\frac{d_2^2(r_c, r_l)}{2\sigma_t^2}) \\ &= exp(-\frac{\|r_c - r_l\|^2}{2\sigma_t^2}) \end{aligned} \tag{3.43}$$

où r_c et r_l sont les positions de neurones c et l respectivement sur la carte, et σ_t est le rayon du voisinage à l'itération t du processus d'apprentissage [115]. La figure 3.9 représente une description de la procédure des déplacements des neurones dans une carte de la méthode SOM. Le résultat de cette méthode est la projection non linéaire des observations sur une carte. Les observations proches dans l'espace d'origine sont présentées par des neurones voisins sur la carte.

Le grand avantage de cette méthode est que le résultat produit est simple à comprendre et offre à l'utilisateur de nombreuses possibilités de visualisation pour ces données. De plus, sa complexité est linaire par rapport au nombre d'observations traitées, par conséquent c'est une bonne candidate pour les grandes bases de données. Cependant, les liaisons entre les neurones de cette méthode sont fixes et ne peuvent pas être cassés, ce qui rend difficile de présenter des données très dispersées dans l'espace.

Méthodes de clustering basés sur la densité

Généralement, les méthodes de clustering basées sur la densité [116, 117] se réfèrent à la définition originale d'un groupe (cluster) [118]. C'est-à-dire, qu'un groupe est défini comme une zone de l'espace dense en données séparées par des zones moins peuplées. Cette notion de densité est fondée sur le concept de voisinage entre les observations dans l'espace. Par ailleurs, pour ce genre de méthodes, plus le nombre de voisins dans le voisinage d'une observation est élevé, plus il est probable que ce point appartienne à une région à haute densité et qu'il fasse partie d'un cluster formé avec ses voisins.

DBSCAN (density-based spatial clustering of applications with noise) est la méthode la plus connue parmi les méthodes de clustering basées sur la densité. Elle a été proposée pour la première fois dans [119]. Comme la plupart des méthodes basées sur la densité, cette méthode s'appuie sur deux paramètres pour classifier les données :

- ε : le rayon de voisinage autour d'une observation x.
- $minPts$: le nombre minimum des observations que nous voulons avoir dans un voisinage pour définir un cluster.

En utilisant ces deux paramètres, les observations peuvent être classées en trois types :

- Observation de base : une observation x_i est considérée une observation de

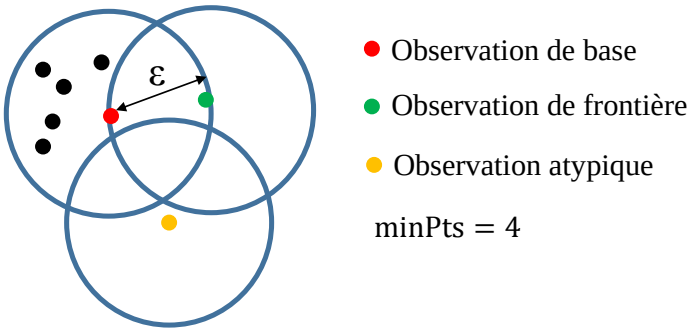


FIGURE 3.10 – Illustration représentant le principe de la classification par DBSCAN.

base si dans une boule de centre x_i et de rayon ε, cette observation x_i a au moins $minPts$ observations comme voisin.

- Observation de frontière : une observation x_i est considérée une observation de frontière, si cette observation x_i est parmi les voisins d'une observation de base, mais dans son cercle de ε-voisinage n'a pas $minPts$ de voisins.
- Observation atypique : une observation x_i est une observation atypique s'il ne s'agit ni d'une observation de base ni d'une observation de frontière.

Par conséquent, les clusters sont formés par les observations de bases et les observations de frontière. La figure 3.10 illustre les différentes catégories des observations, et décrit la façon de regrouper les observations par rapport aux deux paramètres ε et $minPts$.

L'avantage essentiel de cette méthode est qu'elle permet de considérer des clusters de formes très variées. Ainsi elle est capable de gérer les données atypiques qui peuvent exister dans les données d'apprentissage. Cependant, elle est très sensible aux valeurs fixées pour les deux paramètres (ε et $minPts$) et elle est incapable de grouper des clusters dont les densités sont très différentes. Dans la section 4.6 cette méthode a été appliquée sur nos données, mais elle a donné des résultats peu satisfaisants.

3.5 Conclusion

Dans ce chapitre, nous avons évoqué le problème de classification non supervisée. Deux types des méthodes ont été présentés. Le premier est destiné à la séparation d'une classe du reste. Son but est de détecter les données atypiques pour une distribution de données uniques. Le deuxième est consacré aux méthodes de classification ou de partitionnement multiclasse qui sont aussi nommées méthodes de clustering.

Plusieurs méthodes sont présentées pour traiter le problème général de la détection de données atypiques. Elles diffèrent dans leurs manières de détection, de traitement de données et leurs complexités. Les avantages et les inconvénients sont discutés. La méthode de l'ACP à noyau permet notamment de contourner les inconvénients des autres méthodes et plus précisément concernant la prise en compte de la variance de données. Pour cela, cette méthode a été appliquée sur nos données.

De l'autre côté, plusieurs méthodes de clustering sont présentées. Nous avons décrit et montré leurs points forts et faibles. Les méthodes Hiérarchique Ascendante (CAH) sont souvent utilisées en cas des données de petite taille, car la complexité est très élevée, par contre, ces méthodes sont le meilleur choix quand l'utilisateur cherche à illustrer les structures de ses données à plusieurs niveaux de similarité.

Si des problèmes de temps d'exécution se posent, alors ce sont les méthodes de type Kmeans qui sont utilisées. Enfin, si l'objectif est de regrouper des clusters avec des formes quelconques, les méthodes basées sur la densité peuvent s'avérer plus pertinentes. Alors, le choix d'une méthode adaptée à nos objectifs dépend fortement de l'application et des ressources disponibles. En générale, une analyse attentive des données aide à bien choisir la meilleure méthode soit pour le clustering ou la détection de données atypique. Il n'existe pas un algorithme qui peut répondre à toutes les demandes.

Malgré la diversité de méthodes existantes, le problème de classification non supervisée reste toujours ouvert à de nouvelles propositions. Dans le chapitre suivant, une première application de ces méthodes sera présentée afin de résoudre le problème de classification des structures peptides tridimensionnels.

Chapitre 4

Analyse des conformations au niveau d'un seul peptide

Sommaire

4.1 Introduction

Constituées de 20 acides aminés majeurs [35], les protéines sont les éléments essentiels à la vie. La séquence protéique est typiquement notée comme une chaîne

de lettres, énumérant les acides aminés (tels que EGEFPG...). Ces acides aminés sont appelés les résidus dans ce qui suit. La structure de cette séquence définit la forme tridimensionnelle de la protéine qui est, comme évoqué dans la chapitre 2, appelée structure tertiaire. La conformation est la représentation spatiale de la structure indépendamment de sa position dans l'espace et de son orientation. La fonction des protéines est liée à leurs interactions spécifiques avec des récepteurs. Un récepteur peut accepter plusieurs ligands possibles qui peuvent présenter des conformations spécifiques lors de l'interaction. La nature dynamique de l'interaction entraîne des conformations majeures qui cependant peuvent fluctuer pour permettre une meilleure interaction.

Sur les dernières décennies, de nombreuses protéines et peptides ont été découverts dans le domaine de la biologie à l'échelle atomique. Celles-ci ont augmenté les demandes en méthodes mathématiques automatiques capables de faciliter le travail des biologistes avec les protéines ; plus précisément, elles aident pour classifier les structures protéiques et établir le lien avec leurs fonctionnalités. Dans ce contexte, de nombreuses méthodes ont été proposées pour identifier les conformations répétitives des protéines afin de comprendre leurs fonctionnalités. Ces études pourraient être divisées en deux principales catégories.

La première catégorie repose sur la séquence de la protéine [120, 121, 122]. La plupart des méthodes de cette catégorie extraient les propriétés biologiques des séquences des protéines (comme leurs points isoélectriques, leurs poids moléculaires, leurs compositions atomiques et la longueur des chaînes des acides aminées) et les utilisent ensuite dans l'apprentissage automatique pour déterminer certaines fonctions clés. Les méthodes de cette catégorie ne sont pas suffisantes pour déduire les fonctions principales de la protéine, car la fonctionnalité des protéines est liée à la fois à leur structure et à leur thermodynamique [123]. Ceci explique en particulier pourquoi deux protéines avec des séquences différentes peuvent, ou non, avoir une fonction similaire sachant qu'elles peuvent présenter la même conformation en statistique, mais aussi pendant leurs dynamiques moléculaires. Par extension, les mêmes difficultés sont rencontrées avec les peptides, fragments de protéines.

La seconde catégorie est plus étroitement liée à notre travail. Elle repose sur la structure géométrique tridimensionnelle des protéines, que nous décrivons dans la partie suivante [124, 11, 125, 126, 127, 128, 129, 130].

4.2 Méthodes de classification des protéines

La majorité des méthodes de cette catégorie visent à résoudre le problème de la comparaison de plusieurs structures protéiques en améliorant les méthodes d'alignement de structures, comme DALI et SSAP [11, 126]. L'utilisation de l'expression alignement signifie la superposition des structures géométriques des protéines sous la contrainte de maximiser la correspondance des résidus entres les deux protéines superposées.
En premier lieu, la problématique principale de la classification des protéines se base sur la mesure de similarité structurale entre les protéines comparées. Les grandes différences entre les méthodes de classification des protéines se manifestent dans cette mesure. Récemment, beaucoup de recherches ont tenté de capturer cette ressemblance entre protéines de manière plus performante ; mais en général, comparer un grand nombre de ces structures génère des problèmes complexes [131] et particulièrement dans les cas des protéines comme de l'élastine dont la structure est instable. La majorité des méthodes proposées sont conçues, soit en simplifiant le problème et aboutit à une approximation de la solution optimale comme le montre Godzik dans son article [132] (parfois il n'y a pas unicité de la solution optimale), soit le problème est bien défini et les solutions sont accessibles, mais les trouver nécessite des temps de calcul non raisonnables.

4.2.1 Mesure de la similarité structurale

Cependant, indépendamment de la méthode utilisée pour aligner ou superposer deux structures protéiques, la mesure de similarité la plus utilisée est l'écart quadratique moyen RMSD (Root Mean Square Deviation). Il s'agit simplement de la racine carrée de la distance au carré moyenne des atomes équivalents entre les protéines superposées. Elle est définie par [133]. En considérant que la comparaison se fait entre deux protéines P_1 et P_2, le RMSD est calculé comme suit :

$$RMSD = \sqrt{\frac{\sum_{i,j=1}^{n} d_{ij}^2}{n}} \tag{4.1}$$

où d_{ij} est la distance entre les deux atomes équivalents i et j dans P_1 et P_2 respectivement, et n les nombres d'atomes équivalents.
Des structures de protéines similaires ont tendance à donner une petite valeur de RMSD [133]. Néanmoins, ce critère peut être plus élevé pour des structures qui ont beaucoup plus de résidus (acides aminé). Ceci met en évidence le problème principal avec l'utilisation du RMSD comme mesure de similarité : elle dépend de la taille des protéines alignées. Il est donc important de tenir compte à la fois du RMSD et du nombre de paires de résidus équivalentes lors de l'évaluation de la significa-

tion de la similarité. Mais, malgré ces limites, le RMSD reste une mesure largement utilisée et précieuse. Concernant les outils de recherche de similarité sur des domaines protéiques, plusieurs méthodes ont proposé des solutions statistiques pour caractériser la structure de ces protéines, dont un ensemble est brièvement présenté ci-dessous.

4.2.2 Méthode DALI

L'idée essentielle de la méthode DALI est de représenter chaque structure protéique par une matrice (nommée matrice des distances) qui contient les distances entre les carbones C du backbone de la structure [11].
Holm et Sander ont développé l'algorithme DALI, qui consiste à fragmenter les protéines en hexapeptides (6 acides aminées). Ils ont ensuite utilisé leurs matrices des distances pour les comparer. Ils décrivent leur méthode de comparaison comme un processus de coulissement d'une matrice sur une autre. Puis, des petits patchs similaires de taille (6) sont identifiés dans les deux grandes matrices des distances pour prouver la présence de fragments potentiellement équivalents dans les deux structures comparées. La figure 4.1 illustre de manière simplifiée la méthode. Nous pouvons la décrire en trois étapes :

- diviser les matrices des distances en sous-matrices de taille fixe $(6 * 6)$.
- rechercher à travers les deux grandes matrices (de deux protéines) pour trouver des motifs similaires.
- assembler les paires de sous-matrices qui se chevauchent pour optimiser une fonction de similarité qui nous donne la meilleure combinaison entre le nombre d'atomes alignés et la valeur de RMSD global.

Cette méthode a l'avantage majeur de comparer des protéines de différentes tailles. Malgré sa précision et sa sensibilité dans plusieurs applications, cette méthode est encore très coûteuse, car elle explore toutes les possibilités de comparaisons entre les deux matrices des distances.

Comme nous travaillons au niveau d'un seul peptide et que nous traitons des structures ayant la même taille, cette méthode n'est pas adaptée, car nous ne pouvons pas profiter de son avantage principal (comparaison des protéines de différentes tailles). De plus, nous devons également garder à l'esprit que les peptides d'élastine traités dans ces travaux sont très spécifiques (82% des séquences de tropoélastines sont composées de cinq acides aminés, avec de nombreux motifs répétitifs) et participent à l'élasticité des tissus, loin des domaines globulaires ou des structures transmembranaires des protéines qui ont des structures secondaires régulières (comme

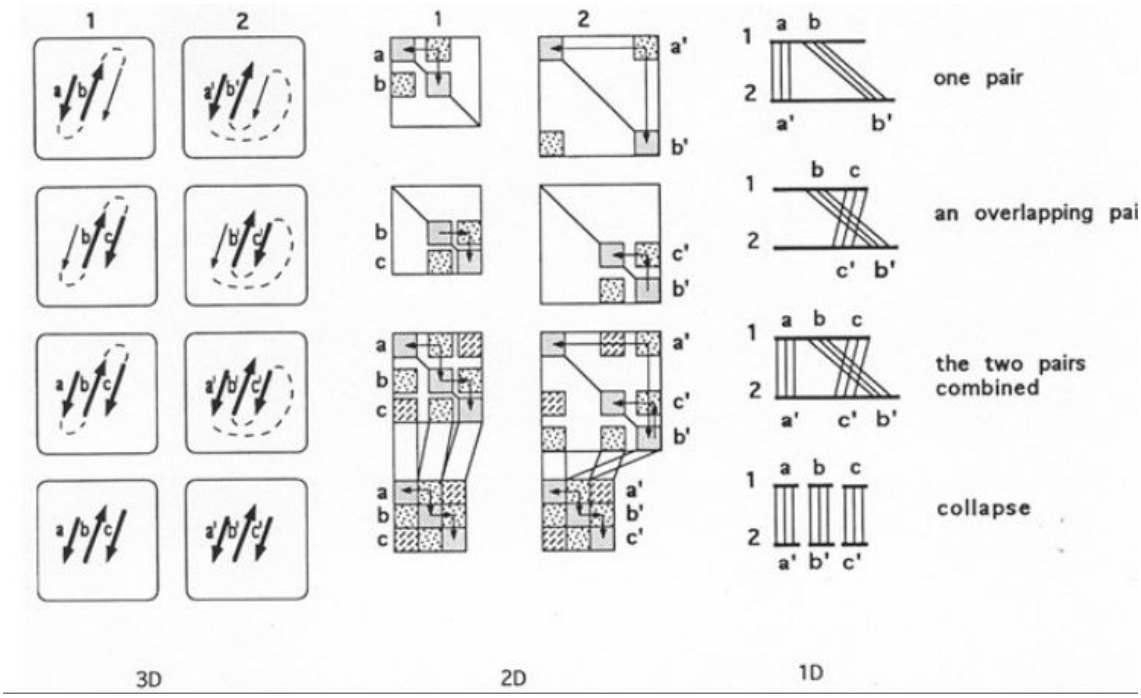


FIGURE 4.1 – Illustration de la procédure de fonctionnement de la méthode DALI de Holm et Sander. Étape 1 : les deux grandes matrices des distances de deux protéines sont divisées en des sous-matrices de taille fixe (6). Étape 2 : nous cherchons les sous-matrices qui ont un modèle similaire dans les deux protéines. Étape 3 : les fragments hexapeptides de ces deux sous-matrices sont concaténés et leurs valeurs de RMSD sont vérifiées. Étape 4, une optimisation de Monte-Carlo est utilisée pour guider le processus vers un alignement complet [11].

hélices, brins ou feuilles et virages) et qui sont souvent analysés par DALI. Pour le moment, la plupart des applications qui traitent de la comparaison des protéines n'ont jamais appliqué leurs méthodes aux peptides d'élastine.

4.2.3 Méthode reposant sur les courbes paramétrées

Un autre genre de méthodes appliqué sur les protéines considère leurs structures comme des courbes continues paramétrées au lieu d'une suite de points discrets en 3D [124, 125]. L'idée de cette approche est d'interpoler entre les coordonnées des atomes de la protéine pour créer une courbe qui représente cette structure. Puis, après cette interpolation, les formes des courbes sont analysées en utilisant une nouvelle fonction, appelée "Square-Root Velocity Function" (SRVF) définie par :

$$q(t) \equiv \frac{\dot{f}(t)}{\sqrt{|\dot{f}(t)|}} \tag{4.2}$$

où $f(t)$ est la courbe paramétrée $f : [0,1] \rightarrow \mathbb{R}^3$, $t \in [0,1]$ et $|.|$ est la norme du vecteur. Pour comparer les structures des protéines, l'auteur de cette méthode [124, 125] propose une nouvelle mesure de similarité entre les courbes basée sur cette fonction .

Cependant, malgré l'efficacité de cette méthode, il existe un risque de perte d'information liée à la chaîne latérale des acides aminés en raison de la fonction d'interpolation utilisée entre les points discrets pour obtenir les courbes. Dans notre travail, nous avons observé que la chaîne latérale a des effets sur les résultats de classification des structures peptidiques (problème discuté dans le chapitre 4). Par conséquent, nous ne pouvons pas appliquer cette méthode sur nos données.

4.2.4 Méthodes de superposition de corps rigides

Il est possible de traiter les structures protéiques comme des objets rigides et de considérer la superposition comme la meilleure façon de les adapter les unes aux autres (malgré la différence énorme parfois entre les structures superposées). De nombreuses discussions et études ont considéré cet aspect de comparaison et la première a été réalisée en 1970 [134, 135, 136, 137, 138]. Ces méthodes se traduisent par trois grandes étapes :

- translater toutes les structures à une position commune dans l'espace, ce qui généralement se fait par la translation de leurs centres de masse (cela est calculé par la moyenne de positions de tous les atomes de la structure) à l'origine ;
- trouver des positions initiales convenables pour démarrer la superposition ;
- tourner une protéine, par rapport à l'autre, autour de trois grands axes pour rechercher le meilleur ajustement.

La difficulté majeure avec cette méthode réside dans l'identification des positions équivalentes putatives pour commencer la superposition. Selon plusieurs études [138, 139, 126], ces méthodes peuvent être utilisées pour des familles de protéines proches (c'est-à-dire qui ont des propriétés physico-chimique similaires). Par contre, pour des protéines plus éloignées, ceci n'est pas précis comme il n'y a pas une solution unique. De plus, cette méthode reste très complexe. Par conséquent, dans notre cas, pour des milliers de structures, cette méthode est trop lourde et peu efficace puisqu'il est quasiment impossible de trouver deux structures qui se superposent exactement à cause de l'instabilité des peptides traités. De plus, les corps rigides ne prennent pas en compte la dynamique et la flexibilité que nous cherchons à exploiter.

Une méthode récente basée sur l'ACP et sur la méthode du point le plus proche itératif (ICP) a montré son efficacité par rapport aux autres [138]. Elle a pour but de superposer les structures protéiques sous la contrainte des atomes manquants dans les deux structures superposées (c'est-à-dire qui ont des tailles différentes). Cette méthode a attiré mon attention afin de comparer les conformations principales de chaque peptide et d'identifier les différences entre les peptides actifs et non actifs (section 5.4). Cependant, nous ne l'utilisons pas dans cette partie où nous comparons les structures qui appartiennent aux mêmes peptides. Ceci est dû aux problèmes communs à toutes les méthodes de superposition : le temps de simulation avec des grandes structures et la position initiale pour commencer la superposition.

4.3 Formulation du problème

Cette section expose formellement le problème de l'identification des conformations rencontrées dans les structures peptidiques avant de présenter la méthode de résolution proposée.
Comme mentionnée précédemment, un peptide est un ensemble d'atomes dans l'espace. La position tridimensionnelle des N atomes qui composent un peptide est appelée structure du peptide et notée $\mathcal{S}$. Une telle structure est un ensemble de N vecteurs en 3D : $\mathcal{S} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N)^T$ avec $\mathbf{a}_n \in \mathbb{R}^3$ la position de $n-$ième atome dans l'espace. Dans cette partie, nous travaillons avec une suite temporelle de structures peptidiques qui appartiennent au même peptide, et qui se déplacent beaucoup au cours du temps, puisque comme pour toute molécule, l'atome se déplace relativement aux autres, suite aux influences thermodynamiques. Notons $\mathbb{S} = (\mathcal{S}_1, \ldots, \mathcal{S}_T)$ la séquence temporelle inspectée des T structures d'un peptide.

Comme décrit dans la motivation du sujet de thèse, dans un premier temps, le but de notre travail est de proposer une méthode statistique pour trouver sans intervention manuelle, les conformations les plus fréquentes dans une suite temporelle de structures peptidiques $\mathbb{S}$. Formellement, deux structures $\mathcal{S}_t$ et $\mathcal{S}_u$ ont exactement la même conformation si et seulement s'il existe une rotation, caractérisée par la matrice $\mathcal{R} \in \mathbb{R}^{3*3}$ et une translation, caractérisée par le vecteur $\boldsymbol{\zeta} \in \mathbb{R}^3$, tel que

$$\mathcal{S}_u = \mathcal{S}_t\mathcal{R} + \mathbf{1}_N\boldsymbol{\zeta}^T \tag{4.3}$$

avec $\mathbf{1}_N$ un vecteur de N ligne contenant seulement des 1. Par conséquent, la conformation $\mathbb{F}$ peut être considérée comme l'ensemble de toutes les structures qui ont la même conformation que $\mathcal{S}_u$:

$$\mathbb{F} = \left\{\mathcal{S}_u | \mathcal{S}_u = \mathcal{S}_t\mathcal{R} + \mathbf{1}_N\boldsymbol{\zeta}^T\right\} \tag{4.4}$$

la structure à partir de laquelle la conformation est définie, ou n'importe quelle structure dans l'ensemble $\mathbb{F}$. $\mathbb{F}$ est une classe d'équivalence.

Les conformations principales d'une suite temporelle peptidique $\mathbb{S}$ sont celles dans lesquelles tombe le plus grand nombre de structures de la suite temporelle $\mathbb{S}$. Cependant, comme il est connu que les déplacements de chaque atome sont aléatoires, il est évident deux structures n'ont jamais la même conformation, telle que définie par l'équation 4.4. Puisqu'il s'agit de trouver les structures qui sont associées aux fonctionnalités des peptides, nous pouvons définir une conformation dans un sens plus large comme suit :

$$\mathbb{F} = \left\{ \mathcal{S}_u | \text{distance}(\mathcal{S}_u, \mathcal{S}_t \mathcal{R} + \mathbf{1}_N \boldsymbol{\zeta}^T) < \varepsilon \right\} \tag{4.5}$$

où la distance doit être définie plus tard et ε définit la limite supérieure de la distance à la structure $\mathcal{S}_u$ (centre) qui nous autorise pour considérer une structure $\mathcal{S}_t$ comme ayant une conformation équivalente ; nous pouvons donc considérer ε comme le périmètre de la fonction des peptides.

En se basant sur la définition « pratique » (Éq.(4.5)) de la conformation, il est évident que n'importe quelle méthode de classification devrait affecter toutes les observations, ou les structures de la suite temporelle inspectée $\mathbb{S}$, à la conformation équivalente qui lui rassemble le plus, si elle existe. Dans l'hypothèse que le peptide passe d'un état stable (conformation principale) à un autre avec des transitions rapides qui ne ressemblent à aucune conformation principale, ces structures peuvent être considérées comme des conformations atypiques. Donc, elles peuvent être appelées les conformations transitoires.

Notons $L(t)$ l'étiquette de la structure $\mathcal{S}_t$ et $\{\mathbb{F}_1, \mathbb{F}_2, \ldots, \mathbb{F}_K\}$ l'ensemble des K conformations principales. La règle d'attribution aux conformations principales décrite ci-dessus peut être définie formellement comme suit :

$$\begin{cases} L(t) = k \;\; \text{si} \;\; \mathcal{S}_t \in \mathbb{F}_k \\ L(t) = 0 \;\; \text{si} \;\; \forall k \in 1, \ldots, K \,,\, \mathcal{S}_t \notin \mathbb{F}_k \\ L(t) = 0 \;\; \text{si} \;\; \mathcal{S}_t \in \mathbb{F}_k \; et \; \mathcal{S}_t \in \mathbb{F}_l \; ou \; plus, l \neq k \end{cases} \tag{4.6}$$

où 0 sont les étiquettes des structures rejetées (les données atypiques) ayant des conformations uniques ou qui se situent entre plusieurs conformations principales $\mathbb{F}_k$.

Par ailleurs, les équations (4.3) - (4.6) formalisent les problèmes associés à cette partie de travail, qui visent à définir les K conformations principales à partir d'une suite de structures peptidiques $\mathbb{S}$. Dans un premier temps, nous allons identifier les difficultés majeures de ce travail.

Première difficulté

La première difficulté est due à la superposition entre les structures peptidiques (voir équations (4.3)-(4.5)), qui est nécessaire pour définir les conformations de façon indépendante de la rotation et de la translation des structures peptidiques. Cependant, la superposition est très complexe et ne fournit pas toujours des résultats précis pour répondre à notre objectif en raison des problèmes cités dans la section 4.2.4 (problème de la position initiale pour commencer la superposition et le problème de la non-unicité des solutions). Par conséquent, la méthode proposée doit éviter le calcul des rotations et translations qui permettent de superposer aux mieux les structures. En parallèle, elle doit prendre en compte la rotation et la translation proprement parlées.

Deuxième difficulté

Comme annoncé, l'élimination des structures transitionnelles (données atypiques) dans la suite temporelle des structures $\mathbb{S}$ est très importante pour mieux faire ressortir les structures stables ou fréquentes lors des traitements. Par conséquent, la deuxième difficulté est liée au nombre inconnu de données atypiques dans la suite temporelle des structures $\mathbb{S}$. De plus la définition des structures atypiques (Éq. 4.5) elle-même est imprécise (la distance qui définit une conformation n'est pas connue). Donc, l'approche proposée doit pouvoir éviter les effets des données atypiques sans savoir comment définir la similarité entre les conformations. Pour aller plus loin, même après alignement ou superposition des structures, il faut définir une distance (Éq. 4.5) entre deux structures qui soit liée à la fonction des peptides.

Troisième difficulté

La troisième difficulté est le nombre inconnu de conformations principales dans la suite temporelle des structures $\mathbb{S}$. En effet, la majorité des méthodes de clustering nécessitent de définir préalablement le nombre de clusters. De plus, nous souhaitons également étudier les similarités entre les différentes conformations principales. Donc, la méthode proposée doit être flexible et permettre à l'utilisateur de visualiser les différents clusters trouvés et lui permettre de changer facilement le nombre de clusters si nécessaire. L'objectif est de permettre à l'utilisateur d'ajuster les résultats. Finalement, il faut que la méthode proposée nous permette de représenter les conformations principales et de diviser une conformation en des sous-conformations à la suite de décisions de l'utilisateur en lien avec les questions qu'il étudie.

4.4 Méthode proposée

Cette section présente la méthodologie proposée pour le regroupement des peptides pour atteindre les objectifs qui viennent d'être énoncés. En premier lieu, concernant la difficulté liée aux effets de rotation et de translation des structures dans l'espace, nous avons proposé de représenter chaque structure peptidique comme une matrice de distance $\mathcal{M}$, qui caractérise les distances entre tous les atomes de la structure. Cela signifie qu'une structure est traitée comme une matrice de distance plutôt qu'une séquence de points discrets en 3D. Ce mode de représentation présente l'avantage d'être invariant par rotation et par translation.
Une structure peptidique composée de N atomes est donc représentée par "La matrice de distance "$\mathcal{M}_{\mathcal{S}_i}$ définie comme suit :

$$\mathcal{M}_{\mathcal{S}_i}(k,l) = \|\mathbf{a}_k - \mathbf{a}_l\|_2 \tag{4.7}$$

où $\mathbf{a}_k$ est le k-ième atome dans la structure $\mathcal{S}_i$ et $\|.\|_2$ est la distance euclidienne.

La matrice de différence de deux matrices des distances de deux structures peptidiques est nulle si et seulement si les deux structures ont exactement la même conformation et que leurs atomes sont numérotés dans le même ordre [11]. La matrice $\mathcal{M}_{\mathcal{S}_i}$ est considérée comme une observation dans un espace appelé "espace d'origine". La dimension de cet espace change en fonction du nombre d'atomes qui forment les peptides.

Donc, pour le peptide de la présente étude, la dimension de son espace est égale à $N(N-1)/2$. En raison de la symétrie de $\mathcal{M}$, il n'est pas nécessaire de l'utiliser totalement. Nous pouvons prendre la partie qui est soit au-dessous ou au-dessus de la diagonale. Les axes de l'espace correspondant aux distances retirées de $\mathcal{M}$ sont dans l'ordre $((2,1),(3,1),...,(N,1),(3,2),...,(N,2),...,(N,N-1))$. Chaque structure est ensuite représentée par un vecteur de dimensions $(N(N-1)/2)$. Enfin, un ensemble de T structures forme une matrice de données de dimension $T * (N(N-1)/2)$. La figure 4.2 représente la transformation proposée qui est appliquée sur la structure 3D, dont les observations qui seront traitées dans notre approche.

Ensuite, nous cherchons à éliminer les structures atypiques qui correspondent à des singularités ou à des transitions. L'ACP à noyau est l'une des méthodes les plus performantes, pour détecter ces observations atypiques et nous avons choisi de la mettre en œuvre sur ces données. Le noyau gaussien (Èq.(3.7)) a été choisi pour nos données, car c'est un noyau considéré comme polyvalent [60, 61, 62]. De surcroît, l'ACP à noyau donne à l'utilisateur la possibilité d'ajuster la précision de la détection afin de régler le niveau de sévérité de la sélection qui est opérée. Notons que cette méthode n'avait jamais été appliquée auparavant à ce type de données.

L'étape finale consiste à utiliser une méthode de clustering sur les données restantes pour détecter les conformations principales et étudier les relations entre elles.

Une méthode de classification hiérarchique a été choisie, car elle ne requiert pas la connaissance préalable du nombre de clusters. En outre, elle est capable de mettre en évidence les relations entre les différentes conformations ; elle est assez intuitive et laisse beaucoup de liberté pour ajuster le nombre de clusters en fonction des données et des besoins. La figure 4.3 représente la progression de la méthode appliqué dans notre approche.

4.5 Application de la méthode proposée

Dans l'approche proposée, la détection des données atypiques par ACP à noyau nécessite de choisir 3 paramètres en fonction de nos données. Ces paramètres sont l'écart type de noyau gaussien σ (Éq. 3.7), q le nombre des vecteurs propres pour définir le sous-espace propre et le seuil de détection à appliquer à l'erreur de reconstruction. Pour élaborer une stratégie de choix des paramètres, nous avons généré des données qui imitent le comportement hypothétique des structures peptidiques dans l'espace. Ces données sont ensuite utilisées pour observer l'effet des paramètres sur la détection des données atypiques et choisir celles qui seront appliquées sur les données réelles. Étant des données simulées, nous connaissons le résultat attendu et nous pouvons aussi évaluer les stratégies de choix.

4.5.1 Données issues de simulation

Cet ensemble de données est nommé *DB2*. Les figures 4.4 et 4.5 illustrent les conformations pour cette base de données. Elle est formée de 4 conformations différentes représentant les 4 formes principales que nous cherchons à regrouper et de formes transitoires entre les 4 conformations principales qui simulent l'effet de transitions chez ces peptides. Les 4 conformations principales ont les propriétés suivantes :

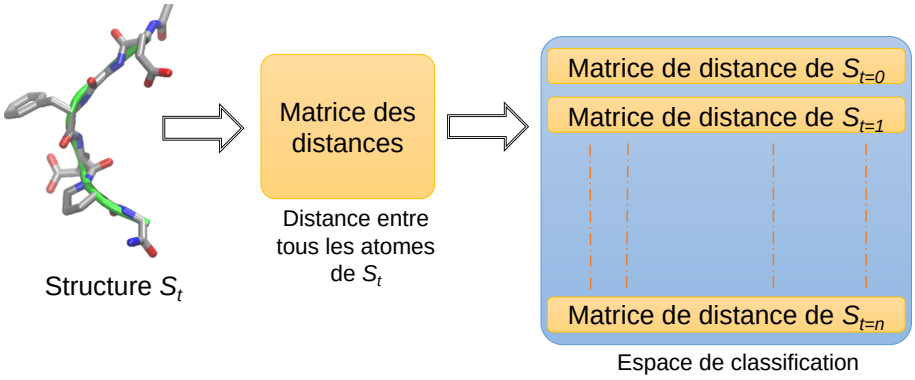


FIGURE 4.2 – Illustration de la transformation proposée pour caractériser une structure peptidique et la formation de l'ensemble de données.

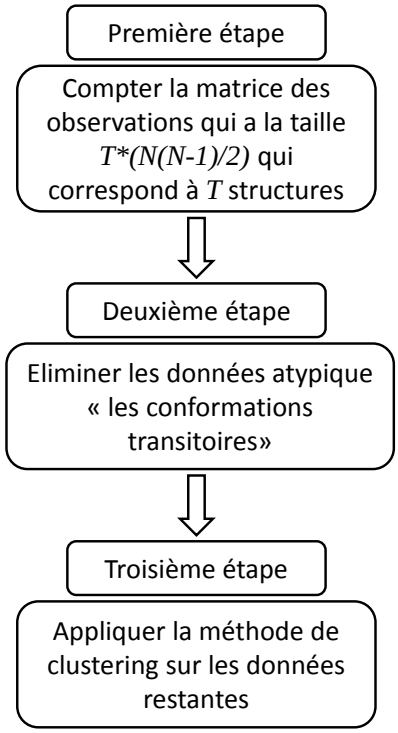


FIGURE 4.3 – Organigramme représentant la progression des méthodes appliquées dans notre approche.

- chaque conformation est composée de 9 atomes. La distance entre chaque paire d'atomes est de $0,2$ nm au bruit près (la même distance qui est entre les atomes

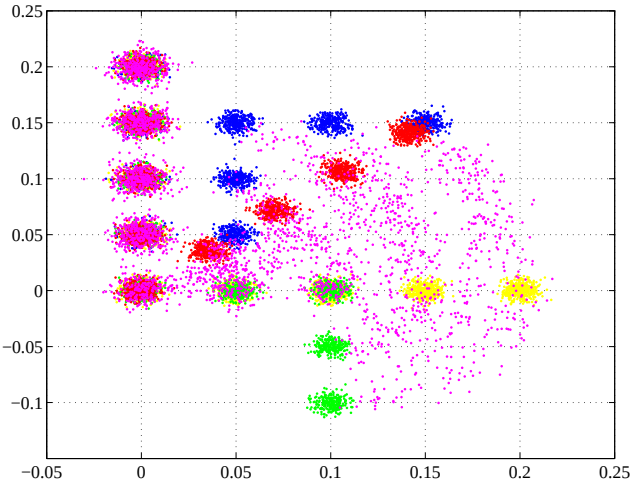


FIGURE 4.4 – Illustration des données générées pour évaluer notre approche. Les données sont formées de 4 formes principales qui sont colorées en bleu, rouge, vert et jaune. Les formes de transition apparaissent en violet.

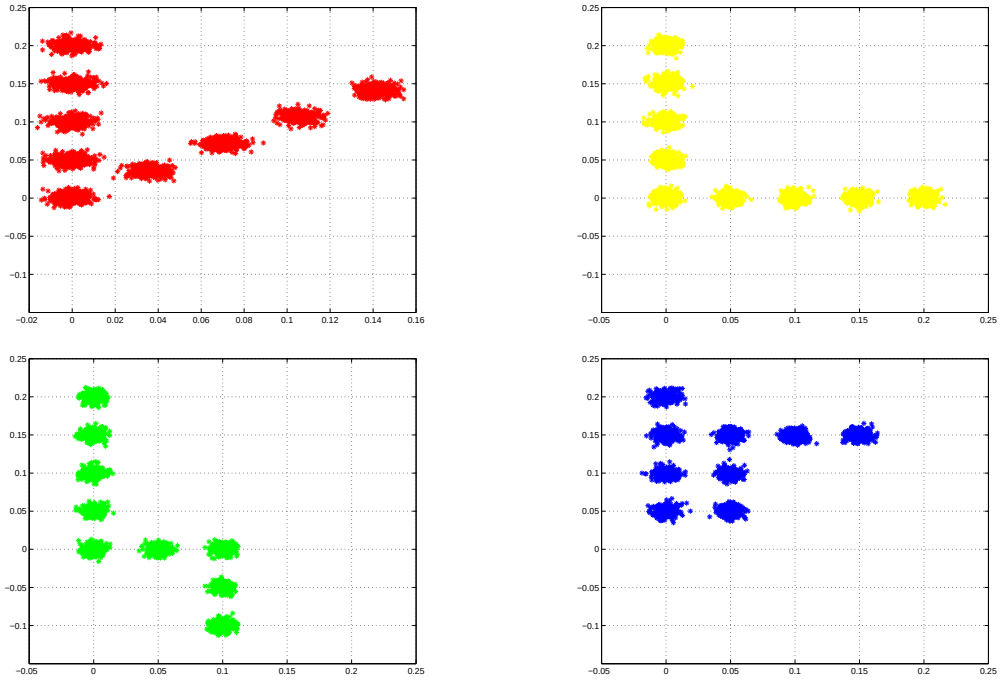


FIGURE 4.5 – Illustration des 4 conformations créées pour simuler le comportement des peptides

de structures des peptides).

- la position de chaque atome suit une loi normale centrée sur la position de référence de l'atome.
- les conformations sont planes (x, y).
- au total, nous avons généré 1300 individus pour les 4 formes principales et 350 formes de transitions entre eux que nous souhaitons détecter.

Réglage des paramètres

Dans cette section, la sélection des paramètres de la méthode proposée est discutée. Comme mentionné dans les sections précédentes, deux paramètres principaux doivent être optimisés : le nombre de vecteurs propres q qui sera utilisé dans pour définir l'espace propre et l'écart-type du noyau gaussien σ. De fait, dans notre cas, nous avons un problème de classification non supervisée, en raison du nombre inconnu d'observations atypiques et de clusters dans les données. Par conséquent, pour évaluer l'approche proposée et optimiser les paramètres, il faut viser des méthodes indépendantes du nombre de clusters et de la proportion de données atypiques. Une liste de ces méthodes a été présentée dans [74]. L'une d'entre elles propose d'appliquer la méthode de classification sur une donnée simulée proche des données réelles, mais dans ce cas le nombre de données atypiques et

de clusters reste connu. Une deuxième approche suggère d'étudier la dispersion intraclasse (équation (4.8)) et la dispersion interclasse (équation (4.9)) des différents résultats de classification, puis prendre les paramètres qui optimise ces deux critères :

$$WC = \sum_{c=1}^{m} \frac{1}{n_c} \sum_{i=1}^{n_c} d(\mathbf{x}_i, \mathbf{g}_c)^2 \tag{4.8}$$

$$BC = \sum_{c=1}^{m} d(\mathbf{g}_c - \mathbf{g})^2 \tag{4.9}$$

avec $\mathbf{g}_c$ est le centre du cluster c, $d(x_i, x_j)$ est la distance euclidienne entre les deux observations x_i et x_j, m le nombre des clusters, et $\mathbf{g}$ le centre global des données.

Comme il n'y a pas d'information concernant l'ensemble des données des peptides, il était nécessaire d'appliquer notre méthode choisie (section 4.4) sur les données simulées et d'étudier l'effet des paramètres σ et q sur la détection des données atypiques afin d'évaluer la méthode ACP à noyau. Le critère d'évaluation que nous avons retenu est la courbe COR [140]. Une courbe COR trace la probabilité de détection (β) en fonction de la probabilité de fausse alarme (υ). Plus précisément le critère est la surface de la zone qui est sous la courbe COR qui est utilisée dans de nombreuses applications [141]. Il est nommé AUC (Area Under Curve) :

$$AUC = \int_0^1 \beta(\upsilon)\, d\upsilon \tag{4.10}$$

où $\beta(\upsilon)$ est la probabilité de détection pour une probabilité de fausse alarme υ. AUC est utilisé pour comparer différentes courbes COR. Plus ce critère est grand, plus la détection est meilleure. Une détection parfaite est obtenue pour $AUC = 1$. La probabilité de détection est estimée à partir de structures qui sont correctement classées comme des données atypiques (vrais positifs). La probabilité d'une fausse alarme est estimée à l'aide des observations classées incorrectement dans le groupe des données atypiques (faux positifs).

Pour construire la courbe COR, il faut tout d'abord que l'indice de "reconstruction error" **Recc**, cité dans la section 3.3.3, soit calculé pour chaque observation. Deuxièmement, toutes les observations doivent être classées par ordre croissant selon l'indice de détection **Recc**. Enfin, en comptant le nombre de structures typique et le nombre de structures transitionnelles ayant une valeur **Recc** supérieure à un seuil donné, nous pouvons estimer la probabilité de fausse alarme et la probabilité de détection associée à chaque seuil de détection.

Les effets de l'écart type du noyau (σ) et le nombre de vecteurs propres (q) sur la détection des données atypiques sont illustrés dans la figure 4.6 et la figure 4.7

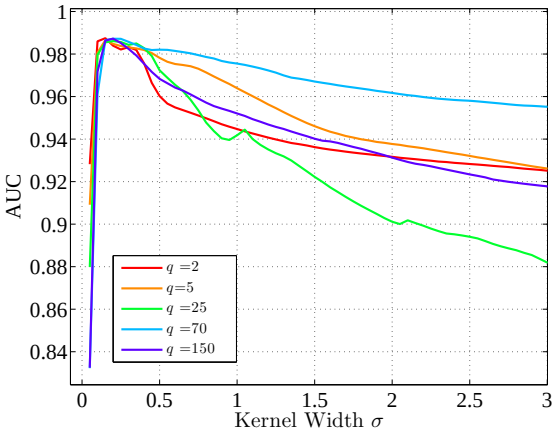


FIGURE 4.6 – Évolution du critère AUC en fonction de l'écart type de noyau σ en utilisant les données simulées (*DB2*). La comparaison est faite pour différentes valeurs de $q = \{2, 5, 25, 70, 150\}$.

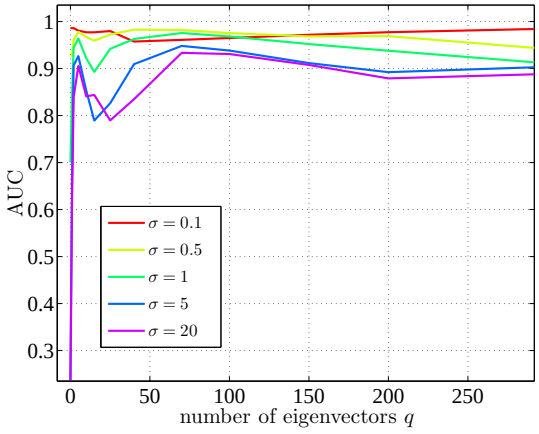


FIGURE 4.7 – Évolution du critère AUC en fonction de nombre de vecteurs propres q en utilisant les données simulées (*DB2*). La comparaison est faite pour différentes valeurs de $\sigma = \{0.1, 0.5, 1, 5, 20\}$.

respectivement. Nous pouvons voir dans la figure 4.6 que pour une petite valeur de σ, la méthode de l'ACP à noyau a une bonne performance pour presque toutes les valeurs de q utilisées. Mais, lorsque la valeur de σ augmente, la valeur d'AUC diminue. Nous concluons qu'une grande valeur de σ n'est pas recommandée. Mais malgré cette observation, la valeur optimale de σ reste toujours inconnue.

De plus, en observant les courbes jaune et rouge de la figure 4.7, qui correspondent aux petites valeurs de σ, leurs courbes apparaissent plutôt stables et les valeurs d'AUC varient peu par rapport à q. Au contraire, lorsque σ est loin de sa valeur raisonnable (courbes bleue et rose) ayant été prévue comme faible, la détection des données atypiques est moins précise. Par conséquent et à partir de ces deux figures, nous pouvons dire qu'il n'est pas nécessaire d'assigner une très grande valeur pour q pour que l'ACP à noyau ait une bonne performance.

En outre, il ne faut pas oublier que la valeur optimale de q dépend aussi du taux d'inertie [142] des données, qui représente la quantité d'informations restante dans les données après exécution de l'ACP à noyau. Ce taux d'inertie est défini par :

$$I = \frac{\sum_{j=1}^{q} \lambda_j}{\sum_{i=1}^{h} \lambda_i} \tag{4.11}$$

où $lambda_i$ est une valeur propre et h est le nombre total de vecteurs propres.

La figure 4.8 présente le nombre de vecteurs propres q qu'il faut garder pour conserver un taux d'inertie donné dans l'espace propre en fonction de σ (quatre valeurs de taux d'inertie différentes ont été considérées). Pour évaluer et observer le comportement de l'ACP à noyau sur la détection des données atypiques pour différents taux d'inertie, les coordonnées des points de chaque courbe de la figure 4.8 ont été considérées comme des paramètres pour la méthode de l'ACP à noyau afin de mesurer les probabilités de détection et de fausse alarme. Les deux figures 4.9 et 4.10 illustrent ces deux probabilités pour différents taux d'inertie. Il est clair que la zone entourée par un cercle noir dans ces deux figures (où β a les valeurs les plus élevées et υ les valeurs les plus basses) donne de meilleures performances que les autres régions. En comparant cela aux courbes de la figure 4.8, nous voyons que cette zone correspond bien aux points d'inflexion.

En conclusion, ces figures aident à estimer les valeurs des deux paramètres optimaux. σ et q ont une marge de valeurs optimales illustrée dans la figure 4.8 par la zone entourée par un cercle noir. Si σ a une petite valeur, la fonction de noyau gaussien $k(x_i, x_j) \approx 0$ pour toutes les observations i et j avec $i \neq j$. Par conséquent, nous éloignons les données, en ce sens chaque point devient un vecteur propre, ce qui entraîne une mauvaise représentation de la variété des données.

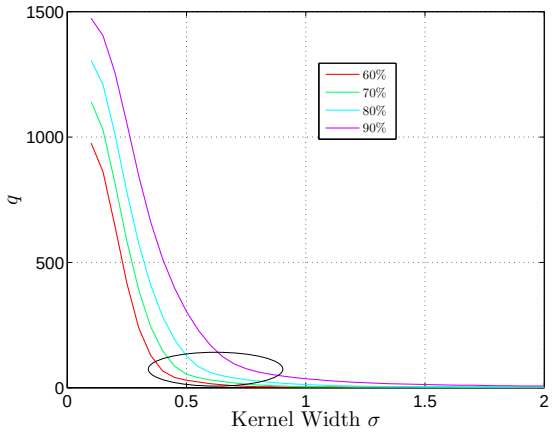


FIGURE 4.8 – Illustration de l'évolution du nombre de vecteurs propres (q) nécessaire pour présenter un taux d'inertie donné, en fonction des valeurs de l'écart type de noyau σ. Ces courbes sont obtenues sur les données de simulations (*DB2*).

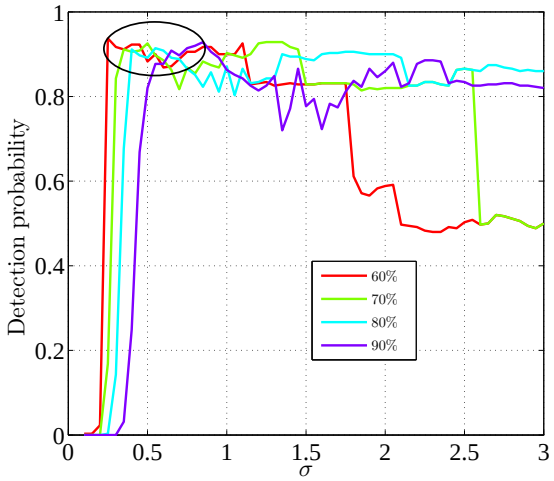


FIGURE 4.9 – Probabilité de détection pour différentes valeurs du taux d'inertie. Nous prenons les coordonnées des points de chaque courbe de la figure 4.8 et nous les considérons comme des paramètres pour la méthode de l'ACP à noyau. Chaque pas de l'abscisse correspond à une valeur de σ avec la valeur de q qui le correspond dans les courbes de la figure 4.8.

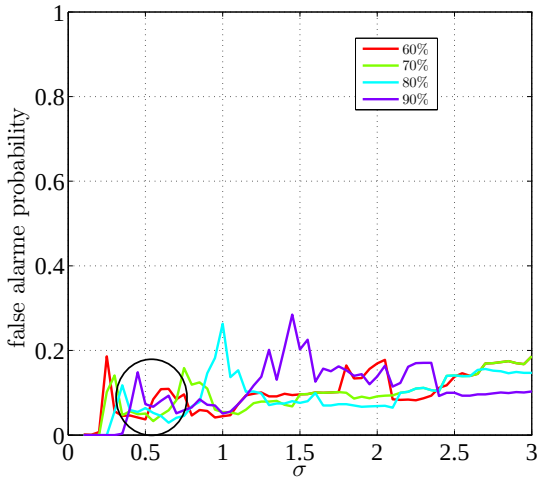


FIGURE 4.10 – Probabilité de fausse alarme pour différentes valeurs du taux d'inertie. Nous prenons les coordonnées des points de chaque courbe de la figure 4.8 et nous les considérons comme des paramètres pour la méthode de l'ACP à noyau. Chaque pas de l'abscisse correspond à une valeur de σ avec la valeur de q qui le correspond dans les courbes de la figure 4.8.

Par contre, si la valeur de σ est grande, l'erreur de reconstruction dans l'espace caractéristique approche l'erreur de reconstruction obtenue par l'ACP classique et donc ACP à noyau perd toute sa signification, comme il est expliqué dans [59]. De plus, l'augmentation de la valeur de q n'a aucune influence sur la détection des données atypiques. Il est donc préférable de choisir une valeur plutôt faible pour q correspondant à une variété de dimension faible.

En se basant sur ces interprétations, nous avons pris les valeurs de σ et de q qui donnent la plus grande valeur de probabilité de détection et la plus petite valeur de probabilité de fausse alarme, puis nous les avons appliqué sur les formes qui sont dans la figure 4.4. La figure 4.11 illustre les conformations restantes après la détection des données atypiques. Ces résultats sont obtenus avec une valeur de $\sigma = 0.4$ et une valeur de $q = 3$. En outre, ils ont été obtenus avec une probabilité de détection égale à 0.94 et une probabilité de fausse alarme égale à 0.034.

4.5.2 Application sur les données réelles

Base de données

Après que les expériences et les tests avaient été réalisés sur les données simulées, nous pouvons appliquer notre méthode de détection des données atypiques et de clustering aux données de notre projet. Ces données sont composées de suites de séquences de points liés en 3D. Elles représentent les configurations spatiales des

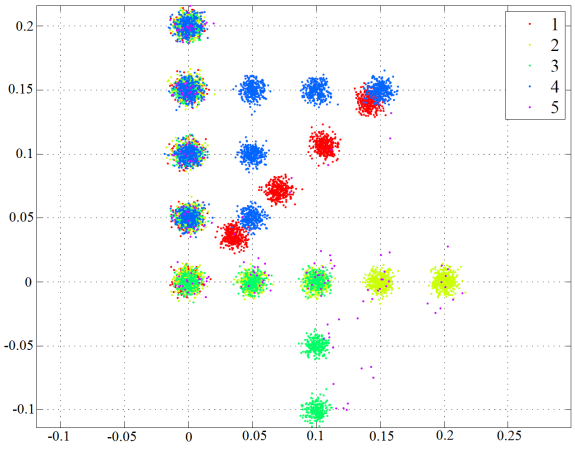


FIGURE 4.11 – Illustration des formes obtenues après l'application de l'ACP à noyau.

atomes composant les différents peptides. Elles simulent les trajectoires dynamiques des peptides au cours du temps. Ces trajectoires de modélisation moléculaire des différents peptides produits lors de la dégradation de l'élastine sont accessibles sur des bases de données précédemment construites. Notre base de données contient les trajectoires de 12 peptides ; chaque trajectoire contient 40 000 positions dans l'espace et est composé de N atomes (N variant de 79 à 89 selon le peptide). Les simulations de dynamique moléculaire (pour les 12 peptides) ont été réalisées avec le logiciel Gromacs et le champ de force OPLS-AA dans l'ensemble NPT (nombre de particules, pression et température constants ; pression = 1 atm, température = 300K). Les peptides ont été simules en présence de solvant explicite (modèle TIP3P utilisé pour décrire les molécules d'eau) dans des boites cubiques de dimension $40 * 40 * 40\mathring{A}^3$. Chaque simulation a été conduite pendant 200 ns avec un pas d'intégration de 2 fs (utilisation de l'algorithme SHAKE qui fixe la longueur des liaisons dans lesquelles un atome d'hydrogène est impliqué) et une sauvegarde des coordonnées toutes les 5 ps. Les conformations initiales des peptides étaient des chaines allongées (angles dièdres phi, psi = 180°, sauf pour les résidus proline phi = $-75°$, $psi = 180°$) et les extrémités ont été neutralisées.

Évaluation de la méthode

Après avoir évalué la méthode proposée sur les données simulée (*DB2*) et identifié la procédure pour déterminer les paramètres de l'ACP à noyau, nous appliquons cette approche aux données des peptides.

Comme noté précédemment, nous sommes dans le cas d'apprentissage non supervisé puisque nous ne connaissons ni le nombre de données atypiques ni le nombre

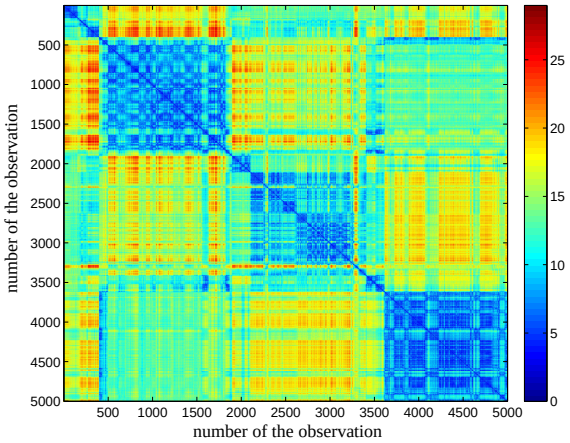


FIGURE 4.12 – Illustration de la matrice représentant la distance euclidienne entre des paires d'observations, pour présenter la similarité entre elles.

de clusters. Cependant, en examinant les données disponibles de l'un des peptides de la base de données, nous avons identifié une séquence de structures de 5 000 observations successives qui peuvent être classées en trois groupes ; comme le confirme la figure 4.12 qui représente la matrice des similarités entre ces 5 000 observations. Chaque ligne et colonne de la figure 4.12 représentent une observation. Le pixel qui a l'indice (i, j) illustre la distance euclidienne entre les observations i et j (équation (4.7)) dans l'espace des distances inter-atomes. La figure 4.12 montre qu'il existe trois types de structures avec de faibles variations dans chacune. La barre de couleur à côté de la figure représente l'échelle de la distance entre les observations. Lorsque la couleur passe du bleu au rouge, cela indique que la distance a augmenté et qu'il y a une différence entre les observations. Celles-ci appartiennent au peptide EGFEPG qui est composé de 87 atomes.

En nous basant sur l'hypothèse que ces 5 000 observations peuvent être classées en trois groupes, nous avons appliqué la méthode de classification proposée pour observer l'effet des paramètres (σ et q) et pour valider notre approche. Cependant, dans ce cas, nous n'avons pas pu utiliser la courbe COR pour évaluer la performance du système, car contrairement aux simulations, les données atypiques sont inconnus. Par conséquent, il est proposé d'utiliser une méthode qui consiste à exploiter la connaissance du nombre de clusters dans les 5 000 observations et d'utiliser les positions des centres de ces clusters (obtenus après la classification en hiérarchie) pour évaluer notre méthode selon les paramètres choisis.

Le processus mis en œuvre pour évaluer l'effet des paramètres (σ et q) sur les résultats de classification est le suivant :

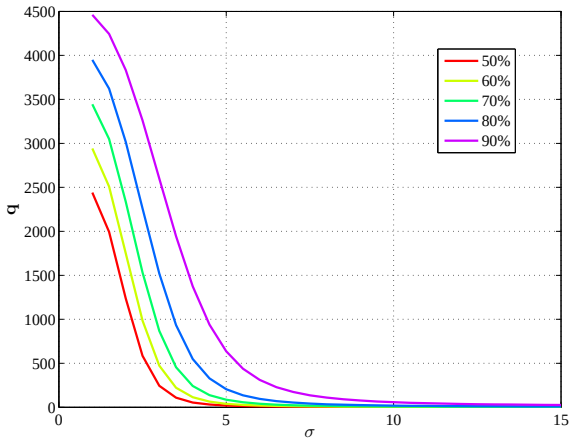


FIGURE 4.13 – Illustration de l'évolution de nombre de vecteurs propres (q) nécessaires pour capter un taux d'inertie donnée en fonction des valeurs de l'écart type σ du noyau. Cette évolution est obtenue avec les données du peptide EGFEPG.

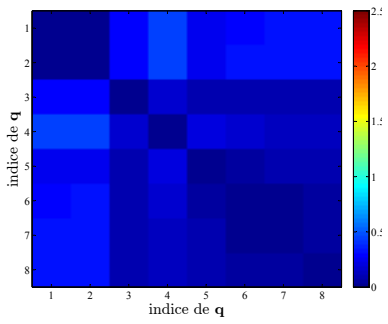
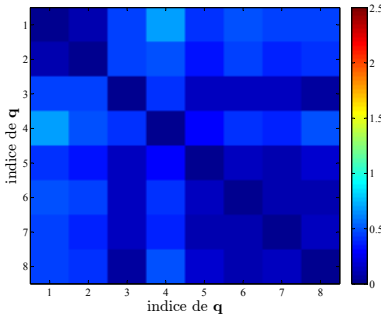
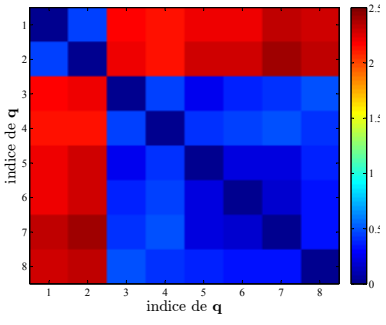
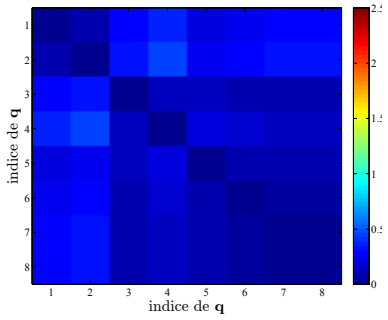
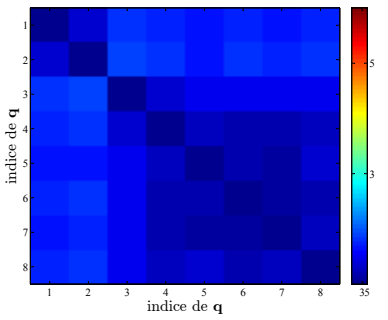
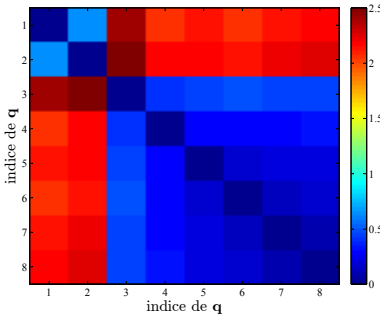
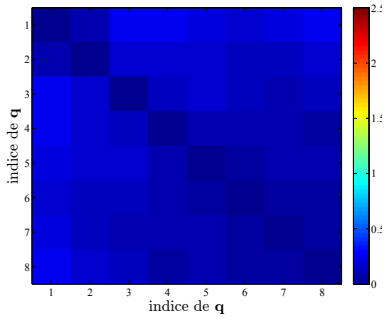
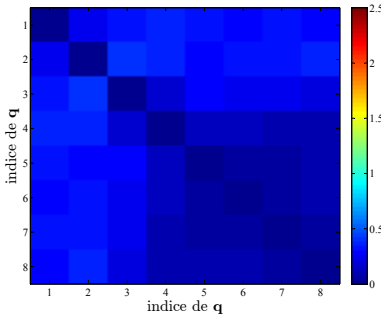
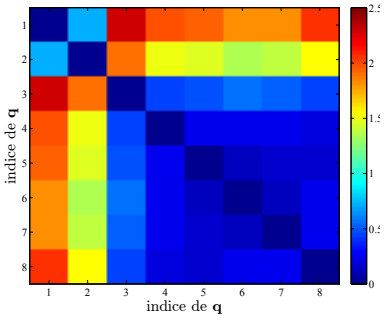
— Dans un premier temps, nous déterminons la courbe du nombre des vecteurs propres q nécessaires pour capter un taux d'inertie donné en fonction des valeurs de σ du noyau. Ces paramètres permettent de sélectionner un intervalle de valeurs raisonnables pour σ, comme nous avons vu dans la figure 4.8. La figure 4.13 montre cette courbe.

— Ensuite, nous exécutons notre approche proposée dans la section 4.4 (détection des données atypiques + clustering) avec chacune de ces valeurs de σ et avec plusieurs valeurs de q.

Afin d'ignorer les formes transitionnelles (données atypiques), nous faisons l'hypothèse que 20% des données sont des données atypiques. Comme notre premier objectif est de trouver les conformations principales plutôt que de déterminer de façon précise le nombre correct des données atypiques, nous pouvons autoriser une surestimation du nomre de données atypiques sans trop affecter les distributions de données des conformations principales. En parallèle, ce nombre de données atypiques peut être modifié en fonction des contraintes et des exigences de l'expert du domaine. Dans notre cas, ce pourcentage est justifié d'un point de vue biologique, en se basant sur la réalité qui prouve que les peptides traités dans notre cas sont des peptides très élastiques et ne sont pas stables. Un tel taux de détection est donc raisonnable. Dans le chapitre suivant, nous traitons cette question et nous analysons la stabilité de ces peptides.

Le tableau 4.1 représente la stabilité des positions des centres des trois clusters obtenus après la classification. Chaque ligne ou colonne de ces six illustrations du

TABLE 4.1 – Évolution des centres des trois clusters qui sont obtenus par la classification hiérarchique après application de la kernel PCA avec 3 valeurs de σ différentes et avec un ensemble des valeurs de $q = [3\ 5\ 12\ 20\ 39\ 64\ 114\ 473]$. Pour chaque figure, nous avons une valeur de σ fixe. Chaque colonne et chaque ligne ont un résultat de q différent.

	Centre C_1	Centre C_2	Centre C_3
$\sigma = 4$			
$\sigma = 5$			
$\sigma = 7$			

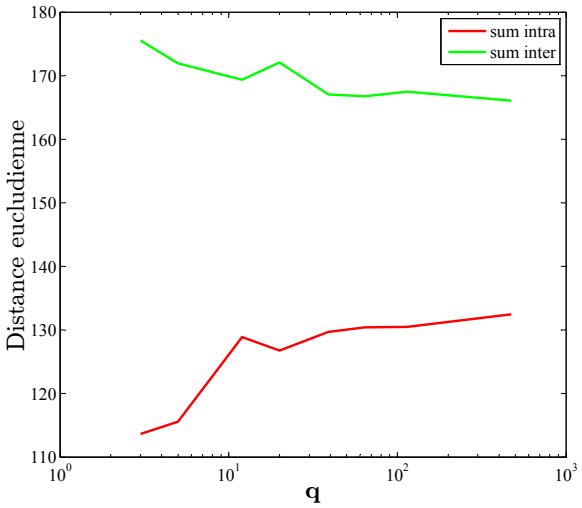


FIGURE 4.14 – Distances intraclasse et interclasse en fonction des q pour les clusters obtenus après application de l'ACP à noyau avec $\sigma = 5$.

tableau 4.1 correspondent à la position du centre obtenue avec une valeur différente de q. Ensuite, un pixel d'indice (i, j) correspond à la différence entre les positions obtenues avec le i-ième et le j-ième valeur de q.
Par conséquent, après avoir exécuté notre méthode avec ces différentes valeurs de q, nous observons qu'après la troisième valeur, les trois centres de ces clusters deviennent presque stables. Par exemple, pour les deux premières figures qui correspondent à $\sigma = 5$ et qui représentent respectivement l'évolution des positions des deux premiers centres C_1 et C_2, ils montrent que ces positions sont presque stables à partir de la première valeur de q. En revanche, pour le troisième centre C_3, il est clair que le centre n'est stable qu'après la troisième valeur de q qui est égale à 12 ; les deux positions de C_3 obtenues avec les deux premières valeurs de q ne sont pas les mêmes que celles qui suivent.
De plus, ces résultats sont confirmés par la mesure de la proportion d'éléments communs dans les trois clusters obtenus et par le calcul de la dispersion intraclasse et interclasse (Éq. (1.8) et (1.9)). Les figures 4.14, 4.15 illustrent ces mesures. Finalement, nous relevons une valeur σ égale à 5, et une valeur à q égale à 12 qui assure une bonne stabilité des résultats sur ces 5 000 observations. C'est sans doute une approche assez conservative, mais en l'absence de validation terroir ce choix nous a semblé plus prudent.

4.6 Analyse et description des résultats

Dans cette section, je ne présente que les résultats obtenus après l'application de notre méthode sur les 5 000 observations (d'un seul peptide) qui sont utilisées dans la

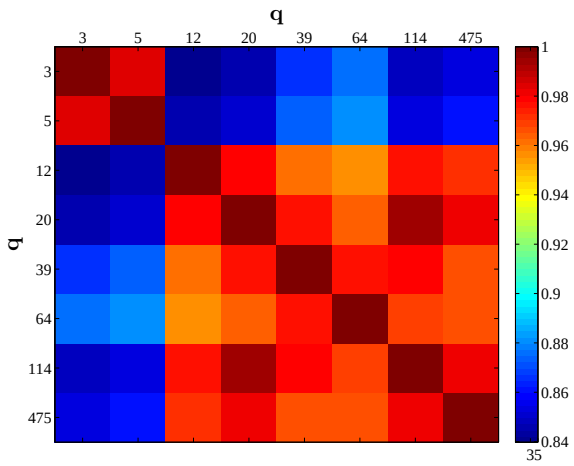


FIGURE 4.15 – Représentation de la matrice de pourcentage de similarité des éléments classés dans les trois clusters après l'application de l'ACP à noyau avec $\sigma = 5$.

partie d'évaluations de la méthode de l'ACP à noyau et qui sont présentées par leur matrice de similarité dans la figure 4.12. Nous appliquons la détection des données atypiques avec $\sigma = 5$ et $q = 12$. Les 1 000 observations ayant la plus grande distance à l'espace propre défini sont considérées comme des données atypiques et elles sont éliminées des données (20% des 5 000). Ensuite, la classification hiérarchique est appliquée sur les 4 000 observations restantes pour trouver les conformations principales. Le dendrogramme de la figure 4.16 montre clairement que ces dernières observations peuvent être divisées en trois clusters. De plus, nous avons utilisé plusieurs indices de validation pour confirmer ce nombre de clusters (indice Davies-Bouldin, indice de Dunn, indice de silhouette, indice de Calinski....) [143]. La figure 4.17 montre les trois conformations principales présentes dans ces observations. En outre, ces trois structures sont choisies de manière à être les plus proches structures 3D aux observations qui sont considérées comme des centres pour les trois clusters résultants.

Pour s'assurer que ce dendrogramme reflète exactement la différence entre les structures des peptides, chacun de ces trois groupes principaux a été séparé en deux sous-groupes. À l'issue, les centres de ces deux sous-groupes sont illustrés et comparés les uns aux autres. La figure 4.18 montre les deux structures les plus proches des centres des deux sous-groupes du groupe 1 qui ont les numéros 4 et 5 dans la figure 4.16. Il est clair dans la figure 4.18 que ces deux conformations ont presque la même forme de backbone (en vert). Seuls les 20 premiers atomes en haut qui sont des chaînes latérales ont changé de direction passant de verticale à horizontale. Cela signifie que les distances entre les 20 premiers atomes sont

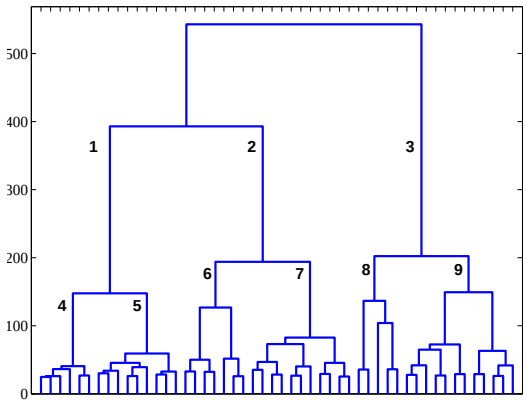
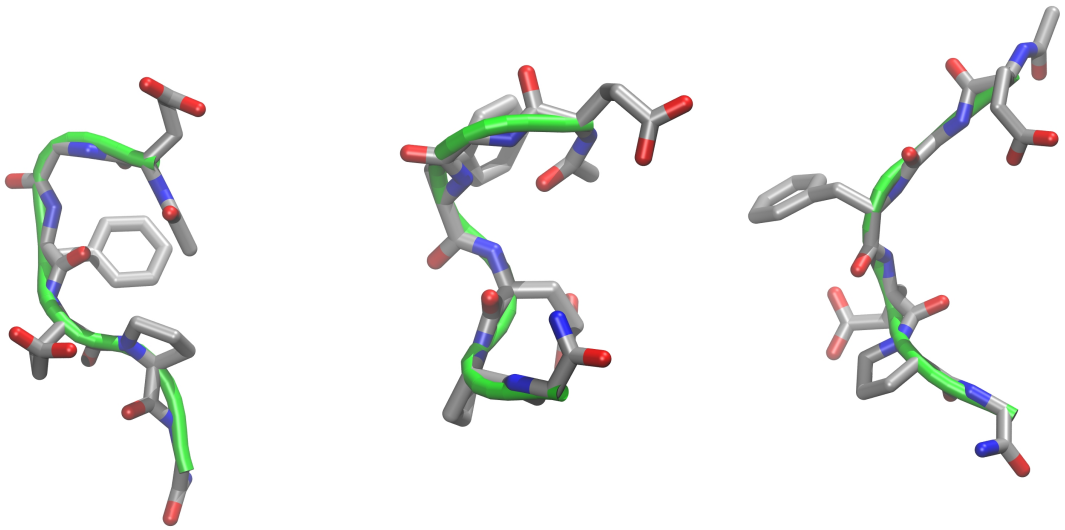


FIGURE 4.16 – Dendrogramme obtenu après l'application de la classification par hiérarchie sur les 4 000 structures sans les données atypiques.



(a) Classe numéro 1. (b) Classe numéro 2. (c) Classe numéro 3.

FIGURE 4.17 – Illustration des conformations représentatives de trois clusters obtenus par notre approche dans les 5 000 observations du peptide EGFEPG.

TABLE 4.2 – Représentation du nombre de clusters et des observations atypiques obtenues avec la méthode DBSCAN. Les résultats sont présentés sous cette forme : (nombre des données atypiques, nombre de clusters). Les colonnes correspondent au nombre minimal de points $MinPts$ et les lignes correspondent au rayon des cercles ϵ.

	.. ⩽2	5	10	15	20	30	50	70	100⩾..
.. ⩽1	(5 000 , 0)	(5 000 , 0)	(5 000 , 0)	(5 000 , 0)	(5 000 , 0)	(5 000 , 0)	(5 000 , 0)	(5 000 , 0)	(5 000 , 0)
2	(4968 , 16)	(5 000 , 0)	(5 000 , 0)	(5 000 , 0)	(5 000 , 0)	(5 000 , 0)	(5 000 , 0)	(5 000 , 0)	(5 000 , 0)
4	(221 , 82)	(469 , 33)	(934 , 18)	(1308 , 13)	(1620 , 11)	(2134 , 10)	(3008 , 6)	(3666 , 5)	(4307 , 1)
6	(0 , 1)	(0 , 1)	(0 , 2)	(20 , 4)	(59 , 3)	(117 , 5)	(247 , 4)	(289 , 4)	(347 , 5)
10⩾..	(0 , 1)	(0 , 1)	(0 , 1)	(0 , 1)	(0 , 1)	(0 , 1)	(0 , 1)	(0 , 1)	(0 , 1)

les mêmes et nous pouvons les considérer comme une seule boîte, bien que les distances entre cette dernière boîte et les atomes restants de la structure ne sont pas les mêmes. La figure 4.19 prouve notre interprétation et représente la différence entre la matrice de distance $\mathcal{M}$ qui correspond au centre du sous-groupe numéro 4 et celle qui correspond au centre du sous-groupe numéro 5. Avec ces résultats nous sommes capables d'assurer l'efficacité de notre méthode dans la description des relations entre les différentes conformations principales obtenues. Nous avons aussi la capacité de comparer le comportement des conformations des chaînes latérales ou de sous chaînes en fonction de la séquence du peptide étudié.

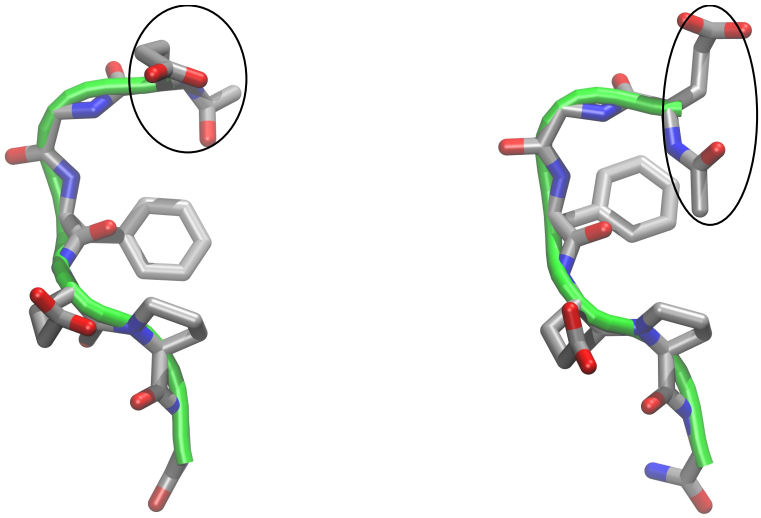
Enfin, pour mettre en évidence la qualité des résultats du clustering, nous pouvons analyser la distribution des distances entre les structures peptidiques et le centre d'une classe conditionnellement à leur affection au cluster. Soit $A^j_{k,i}$ la distance entre la structure i du cluster j et le centre du groupe k définie comme suit :

$$A^j_{k,i} = \sum_h \sum_l \left(\mathcal{M}^j_{\mathcal{S}_i}(h,l) - \mathcal{M}_{\mathcal{S}_k}(h,l)\right)^2 \tag{4.12}$$

où $\mathcal{M}_{\mathcal{S}_k}$ représente la matrice de distance du centre du k-ième cluster, $k = \{1,2,3\}$. La figure 4.20 illustre les distributions des $A^j_{k,i}$ pour les trois clusters obtenus. Ces distributions montrent bien la séparation entre les clusters trouvés et confirment la qualité de la partition ainsi définie par notre méthode de clustering.

Comparaison avec DBSCAN

Nous nous proposons de comparer notre méthode avec la méthode DBSCAN qui est l'une des méthodes les plus populaires dans le clustering que nous avons cité dans l'état de l'art du chapitre 3. Cette méthode est souvent utilisée pour la classification



(a) Conformation représentative du cluster numéro 4.

(b) Conformation représentative du cluster numéro 5.

FIGURE 4.18 – Représentation des conformations représentatives de deux sous-groupes qui sont dans le cluster numéro 1.

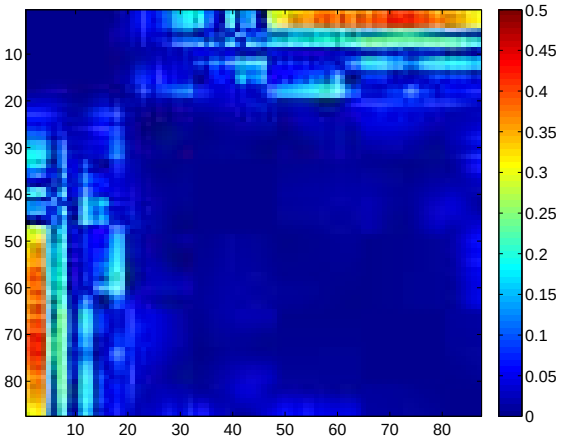
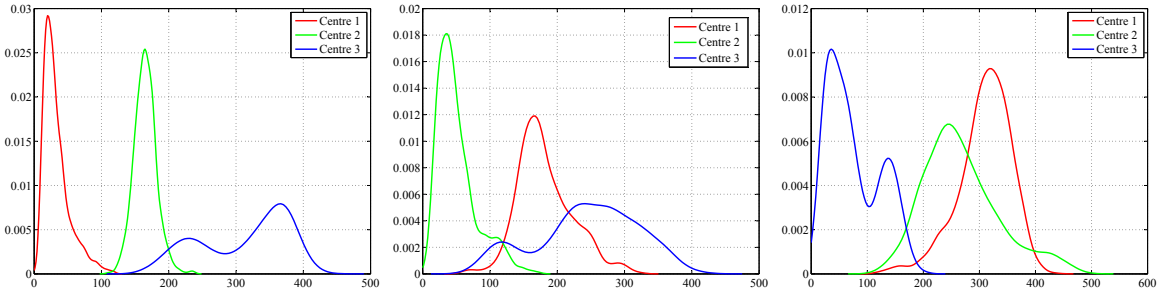


FIGURE 4.19 – Illustration de la différence entre les deux matrices des distances $\mathcal{M}$ des centres des sous-groupes 4 et 5.



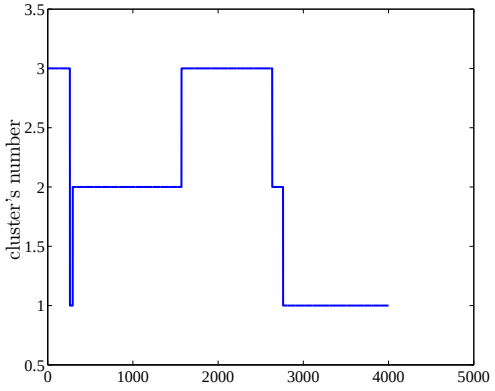
(a) La distribution de A_k^1. (b) La distribution de A_k^2. (c) La distribution de A_k^3.

FIGURE 4.20 – Représentation des distributions des distances entre les éléments de chaque cluster et les trois centres. (a) la distribution des distances entres les éléments du premier cluster et les 3 centres. (b) la distribution des distances entres les éléments du deuxième cluster et les 3 centres. (c) la distribution des distances entres les éléments du troisième cluster et les 3 centres.

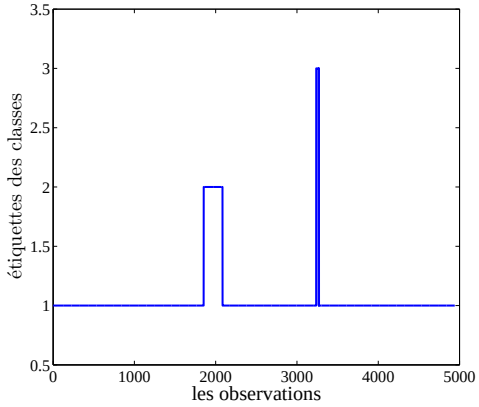
TABLE 4.3 – Représentation du nombre de clusters et des observations atypiques obtenues avec la méthode DBSCAN. Les résultats sont présentés sous cette forme : (nombre des données atypiques, nombre de clusters). Les colonnes correspondent au nombre minimal de points $MinPts$ et les lignes correspondent au rayon des cercles ϵ.

	18	22	25	28	32	36	38	43	49
6.2	(34 , 3)	(39 , 03)	(61 , 3)	(91 , 3)	(99 , 4)	(102 , 4)	(109 , 4)	(123 , 4)	(141 , 4)
6.4	(27 , 2)	(32 , 2)	(42 , 3)	(54 , 3)	(87 , 3)	(92 , 3)	(95 , 4)	(98 , 4)	(114 , 4)
6.6	(7 , 2)	(12 , 2)	(27 , 2)	(38 , 2)	(72 , 2)	(76 , 2)	(81 , 2)	(85 , 3)	(88 , 3)
6.8	(2 , 2)	(7 , 2)	(8 , 2)	(24 , 2)	(65 , 1)	(67 , 1)	(68 , 1)	(73 , 1)	(75 , 2)

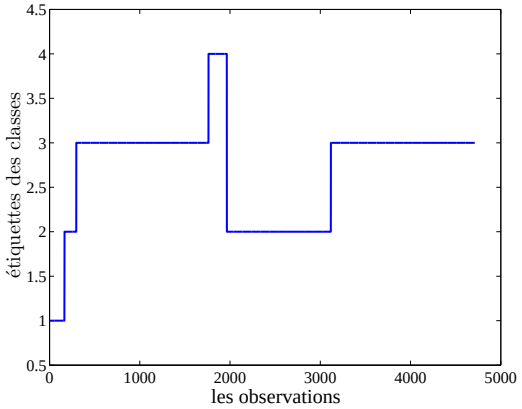
de grande base de données [144]. Elle peut identifier les clusters en examinant la densité locale des éléments de la base de données à l'aide de deux paramètres d'entrée, ϵ et $MinPts$. Elle peut ensuite déterminer les données atypiques à partir d'une règle simple présentée dans en section 3.4.3. Chaque échantillon qui satisfait les conditions liées à ϵ et $MinPts$ est considéré comme un membre d'un cluster, sinon il est considéré comme une donnée atypique. La mode de fonctionnement de cette méthode a été expliquée dans la partie de l'état de l'art des méthodes de classification qui se trouve dans la chapitre 3 section 3.4.3. Les deux tableaux 4.2 et 4.3 montrent les résultats obtenus après avoir exécuté DBSCAN sur les 5 000 observations qui sont considérées dans les simulations précédentes, pour différentes valeurs de ϵ et $MinPts$. Le tableau 4.3 montre les résultats obtenus après une échantillonnage plus fine autour de $\epsilon \in [6, 7]$ et $MinPts \in [18, 50]$. Dans les deux tableaux, les colonnes correspondent aux valeurs de $MinPts$ et les lignes correspondent aux valeurs de ϵ. Dans le tableau 4.2, il n'y a qu'un seul cas qui a l'index $(6, 20)$, où cet algorithme a détecté trois clusters. Pour cela, nous avons ajouté plus des échantillons autour l'index $(6, 20)$ pour voir s'il y en a d'autre cas où nous trouvons trois clusters. Comme nous voyons dans le tableau 4.3, il est clair qu'il existe plusieurs cas où nous avons ce nombre de clusters. Les classifications de la figure 4.21 représentent les distributions des données dans les clusters pour ces différents cas de détection. La figure 4.21b illustre les trois clusters résultants qui sont obtenus par DBSCAN dans tous les cas où il y a trois clusters dans les deux tableaux 4.2 et 4.3. En comparant ces trois clusters avec ce qui est obtenu avec notre approche (figure 4.21a), nous trouvons que la distribution des données dans les clusters est totalement différente. Seulement 59 observations sont considérées comme des données atypiques et quasiment toutes les observations restantes sont regroupées dans le cluster numéro 1 (50 observations classées dans le deuxième cluster et 5 classées dans le troisième cluster). Même dans le cas où nous avons obtenu 4 et 5 clusters, les classifications qui sont dans la figure 4.21c et la figure 4.21d, il est clair que ces distributions ne sont pas équilibrées sur tous les clusters. La totalité des données est quasiment classifiée dans deux groupes. Par conséquent, ils en résultent une classification déraisonnable par rapport à la figure 4.21a qui illustre les 4 000 observations qui sont classées avec notre méthode. Suite à ces différentes configurations de clustering, nous concluons que nos résultats sont plus précis et surtout plus cohérents et robustes que ceux obtenus par DBSCAN.



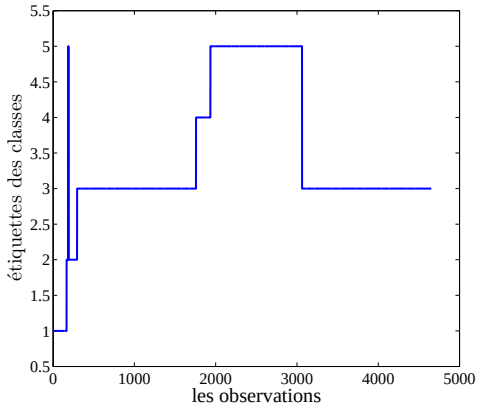
(a) Classification avec notre approche.



(b) DBSCAN avec $\epsilon = 6$ et $MinPts = 20$.



(c) DBSCAN avec $\epsilon = 6$ et $MinPts = 70$.



(d) DBSCAN avec $\epsilon = 6$ et $MinPts = 100$.

FIGURE 4.21 – Illustration des distributions des clusters obtenus avec notre méthode et les 3 configurations différentes de la méthode DBSCAN.

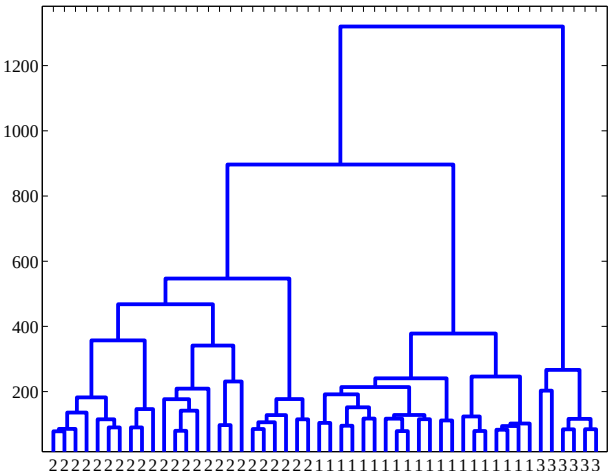
4.7 Clustering avec les atomes du backbone seul et/ou avec tous les atomes du peptide

4.7.1 Problématique

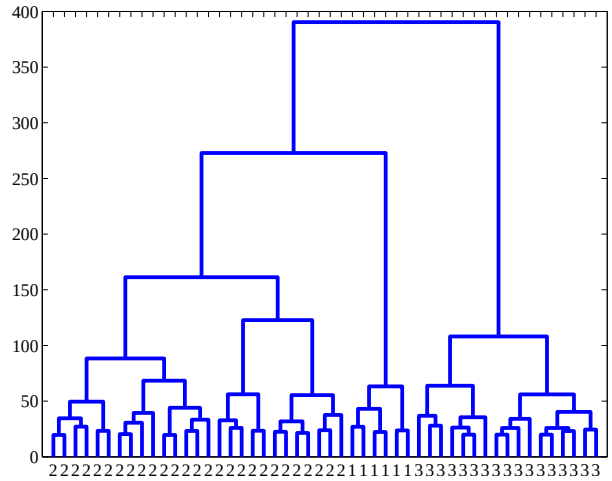
En regardant les recherches et les études statistiques qui ont déjà été réalisées sur les protéines dans le but de classifier et identifier leurs conformations principales, il apparaît que les chercheurs appliquent leurs méthodes à la chaîne principale uniquement (backbone) de la protéine. Ils ne prennent pas en compte la chaîne latérale de la protéine, impliquant indirectement qu'elle n'a pas d'effets sur la classification et le regroupement de peptides.
Le backbone est la suite d'atomes qui est présente dans tous les acides aminés et qui en forme la colonne vertébrale. La chaîne latérale est l'extension du backbone et varie d'un acide aminé à un autre : elle est notée par le symbole R dans la figure 2.9. Nous avons observé dans la section 4.6 qu'en divisant une des classes principales en deux sous-groupes, nous obtenons deux sous-groupes qui ont les mêmes backbone (en vert dans la figure 4.18), tandis qu'ils ont deux chaînes latérales avec des directions différentes. Ces distinctions questionnent la pertinence d'utiliser les chaînes latérales dans la classification. Pour cela, il nous paraît l'importance de comparer les classifications obtenues en considérant tous les atomes des peptides d'un côté et les atomes du backbone de l'autre.

4.7.2 Mise en œuvre

Pour réaliser cette comparaison entre les deux modes de traitements des peptides, nous les avons comparés en utilisant deux peptides de notre base de données : un peptide considéré actif, le VGVAPG et un non-actif, le GVGVAP. Pour chacun d'eux, nous avons appliqué la procédure suivante : nous avons extrait les positions 3D (x, y, z) des atomes des backbones des structures originales de la base de données. Notant que les atomes de base du backbone sont C, O, N et CA, où CA est le carbone alpha de l'acide aminé et ils sont ordonnés en suivant toujours cet ordre. Puis nous avons appliqué notre algorithme de classification (section 4.4) comme décrit précédemment pour extraire les conformations principales. Afin d'être cohérents dans notre traitement, nous avons proposé d'utiliser l'indice Calinski et l'indice de Davies [143] pour déterminer le nombre optimal de clusters pour chaque peptide. Cependant, selon les biologistes, ces indices d'évaluation ne peuvent pas nous donner le nombre logique des conformations existantes dans les structures de ces peptides. En effet, ces peptides sont très élastiques et peuvent avoir des dizaines de clusters, tandis qu'en utilisant les indices statistiques d'évaluation de clustering [143], nous tombons toujours sur cinq clusters au maximum. Enfin, nous affichons la



(a) Dendrogramme obtenue avec tous les atomes.



(b) Dendrogramme obtenue avec les atomes du backbone.

FIGURE 4.22 – Schéma représentant les dendrogrammes obtenues après l'application de la classification par les deux façons de traitement des structures du peptide VGVAPG.

Clusters	A1	A2	A3
B1	0	6.03%	93.9%
B2	15.09%	84.52%	0.39%
B3	77.46%	22.53%	0

TABLE 4.4 – Représentation de pourcentage de similarité entre les différents résultats de classification obtenus avec Backbone seul (B) et tous les atomes - pour le peptide VGVAPG.

similarité entre les structures obtenues par chaque classification et nous identifions les différences.

4.7.3 VGVAPG (actif)

Après l'application de l'algorithme illustré dans la figure 4.3 qui considère 20% des données comme des données atypiques, il reste 32 000 structures que nous classons avec la classification hiérarchique. Dans la figure 4.22 nous illustrons les deux dendrogrammes obtenus avec les deux méthodes de traitement des structures du peptide VGVAPG. Le premier dendrogramme 4.22a illustre la classification obtenue avec tous les atomes et le deuxième dendrogramme 4.22b avec seulement les atomes du backbone. Il est normal qu'ils n'aient pas la même échelle au niveau de l'axe des ordonnées, qui représente la différence entre les classes, puisque nous ne sommes pas dans des espaces de même dimension (nous avons mentionné précédemment que la dimension de l'espace avec notre méthode dépend du nombre des atomes de la structure peptidique). Pour avoir la même référence dans la coupure des dendrogrammes, l'indice Calinski et l'indice de Davies [143] sont utilisés pour obtenir le nombre optimal de clusters pour ces deux dendrogrammes : ils sont de 3. Pour cela, nous coupons chacun d'eux en trois clusters, puis nous identifions les éléments des clusters en commun entre les différentes distributions, tel que le pourcentage de la différence entre les centres obtenus. Après cette classification, 28 738 de 40 000 observations sont classées de façon analogues. En cherchant les éléments qui sont communs entre les 2 classifications, nous obtenons le tableau 4.4. A1, A2, A3 et B1, B2, B3 sont les 2 groupes des clusters obtenus après la classification de tous les atomes d'un côté et les atomes du backbone de l'autre côté.

D'après le tableau 4.4, il est clair que les deux regroupements sont significativement différents, ce qui implique que dans notre cas des peptides une classification basée seulement sur les atomes de backbone fait l'impasse sur les variations induites par les chaînes latérales et donne un résultat simplifié. Par conséquent, il nous

clusters	A1	A2	A3	A4	A5
B1	0	0	99.65%	0.069%	0.27%
B2	0	6.924%	0.27%	0	92.80%
B3	89%	3.23%	0	0.01%	7.75%
B4	4.22%	50.69%	0	2.59%	42.49%
B5	0.012%	13.64%	0	84.64	1.69%

TABLE 4.5 – Représentation de pourcentage de similarité entre les différents résultats de classification obtenus avec Backbone seul (B) et tous les atomes - pour le peptide GVGVAP.

semble préférable d'utiliser tous les atomes du peptide quand nous travaillons avec des structures qui appartiennent au même peptide et qui ont le même nombre d'atomes. Ainsi, travailler avec des protéines qui ont des milliers d'atomes est différent du traitement des peptides qui n'en ont que des dizaines. En effet, le nombre d'atomes pris en compte affecte directement les résultats de classification comme il change la dimension de l'espace (distance inter-atomes) dans lequel nous travaillons et donc la présence du problème de la distance entre observations dans les grandes dimensions [145]. Ceci peut expliquer que les méthodes qui sont citées dans la section 4.2 et qui sont appliquées sur des protéines avec des milliers d'acides aminés ne sont pas adaptées à notre problème et réciproquement.

4.7.4 GVGVAP (non actif)

Pour confirmer les résultats obtenus avec le peptide VGVAPG, nous avons refait la comparaison avec le peptide GVGVAP : celui-ci a 29 184 structures en commun sur 40 000 entre les deux méthodes de classification (avec tous les atomes et les atomes du backbone seul). De plus, nous n'avons pas la même répartition des données dans les clusters obtenus. Les résultats de classification sont affichés dans le tableau 4.5.

À ce stade, l'accumulation d'indices convergents nous permet de valider notre méthode et d'affirmer l'importance de prendre en compte les chaînes latérales pour des petites structures comme les peptides surtout si les structures appartiennent au même peptide. Dans la suite de notre travail, nous rencontrons de nouveau ce problème au niveau de la comparaison entre les conformations principales obtenues avec des peptides différents, ce qui repose la question du mode de comparaison des conformations.

4.8 Conclusion

Ce chapitre étudie la classification des structures 3D des peptides réalisée à partir des simulations numériques qui traduisent le comportement de leur dynamique moléculaire. Ce chapitre introduit un moyen entièrement automatisé pour détecter les principales conformations d'un peptide pendant sa trajectoire moléculaire. Il propose d'utiliser la matrice des distances entre les atomes propres à chaque structure peptidique afin d'éviter de prendre en compte les effets des translations et rotations de la structure au cours du temps et évitant ainsi les opérations de recalage couteuses et imprécises. De plus, la méthode présentée dans ce chapitre propose de détecter les données atypiques qui sont les structures ne ressemblant à aucune autre conformation, en utilisant la méthode de l' ACP à noyau qui n'avait jamais été appliquée à ce type de données. L'élimination de ces données atypiques est très importante, car elle assure une classification qui n'est pas affectée par les observations qui se trouvent essentiellement entre les conformations les plus fréquemment trouvées. Par ailleurs, la méthode de classification hiérarchique sur les données restantes est utilisée pour avoir une grande flexibilité dans la classification des différentes conformations. En effet, en fixant un seuil de coupure de dendrogramme, nous offrons à l'utilisateur la possibilité de choisir entre une classification microscopique ou une plus fine. Nous avons également la capacité de visualiser la différence entre les centres des clusters pour aider l'utilisateur à ajuster la méthode en fonction de ses données et de ses exigences. Les résultats expérimentaux qui sont obtenus sur la base de données jouet et sur les données de simulations de dynamique moléculaire de peptides ont montré que cette méthode (matrice de distance + ACP à noyau + classification hiérarchique) est performante et efficace. Finalement, nous avons confirmé l'importance d'exploiter, dans le cas des classifications des structures appartenant au même peptide, tous les atomes des peptides y compris les atomes des chaînes latérales, celles-ci ayant prouvé leur importance et leur impact direct sur les résultats des classifications.

Chapitre 5

Analyse de l'activité des peptides

Sommaire

5.1 Introduction

Dans ce chapitre, nous travaillons avec une base de données qui contient 12 peptides, dont 4 sont inactifs (APGVGV, GVGVAP, PGVGVA, VAPGVG), 5 actifs (GVAPGV, PGAIPG, VGVAPG, EGFEPG, LGTIPG) et 3 pour lesquels l'activité n'est pas encore connue (PGAYPG, VGLAPG, VVGPGA). Chaque peptide est constitué de 40 000 structures, comme celui traité dans la section 4.5.2.

Comme cité précédemment, l'activité des peptides est reliée aux conformations qu'ils peuvent explorer au cours du temps. Cette activité est identifiable biologiquement par les interactions qu'ils peuvent ou pas avoir avec un récepteur spécifique, nommé Complexe Récepteur de l'Élastine (CRE) [40]. De plus, ces peptides sont des produits issus de la dégradation de l'élastine contenue dans la paroi artérielle, et c'est cette origine "élastique" qui leur donne potentiellement une "hyper-élasticité"

par rapport à d'autres peptides qui seraient issus de la dégradation d'autres tissus qui potentiellement ne sont pas aussi élastiques. Ce potentiel élastique va donc leur conférer une aptitude particulière à pouvoir se déformer.

Dans ce présent chapitre, nous nous concentrons sur l'objectif principal de la thèse. C'est-à-dire que nous tentons d'identifier les particularités structurales et dynamiques qui distinguent les peptides actifs des peptides inactifs. En d'autres termes, identifier les "déterminants" de l'activité d'un peptide. En nous basant sur l'idée que la fonctionnalité des peptides est reliée à leurs structures 3D, nous supposons que les peptides sont actifs du fait de leur capacité à pouvoir adopter certaines conformations spécifiques. A nous maintenant d'être capables de repérer ces déterminants de l'activité par déductions sur les formes actives.

Dans ce travail, nous avons à traitons au préalable plusieurs problèmes :
1) L'instabilité des peptides, car elle implique l'existence de structures transitionnelles, mais aussi il est nécessaire de détecter ces observations "atypiques" avant d'identifier les conformations principales de chaque peptide.
2) La comparaison de peptides différents, présentant de plus des tailles différentes.
Pour ces raisons, dans ce chapitre, nous travaillons en premier lieu sur le reclassement des données atypiques afin d'analyser plus spécifiquement la stabilité des peptides. Nous vérifions ensuite si cette stabilité est pertinente ou pas pour l'identification du nombre de conformations principales au cours du temps. De plus, nous sélectionnerons la méthode la plus appropriée de comparaison des peptides, notamment lors de la comparaison de conformations de peptides entre peptides de différentes tailles. Enfin, nous proposons une méthode pour inférer quant à l'activité des peptides dont l'activité biologique est encore inconnue.

5.2 Reclassement des données atypiques

La stabilité d'un peptide se réfère à l'étude du comportement dynamique de celui-ci. Avant toute analyse de sa stabilité, il est nécessaire d'effectuer un étiquetage précis de toutes les structures transitionnelles. Dans le chapitre 4, 20% des observations ont été rejetées afin de définir de manière nette les différents clusters. Ce taux de rejet a été volontairement choisi élevé pour induire un fort contraste entre les clusters.

La figure 5.1 montre la distribution des structures sur 10 clusters (i.e. conformations) qui sont entre les instants 8700 et 9700 de la trajectoire moléculaire du peptide nommé VAPGVG (nous le représentons avec 10 clusters pour mieux voir les

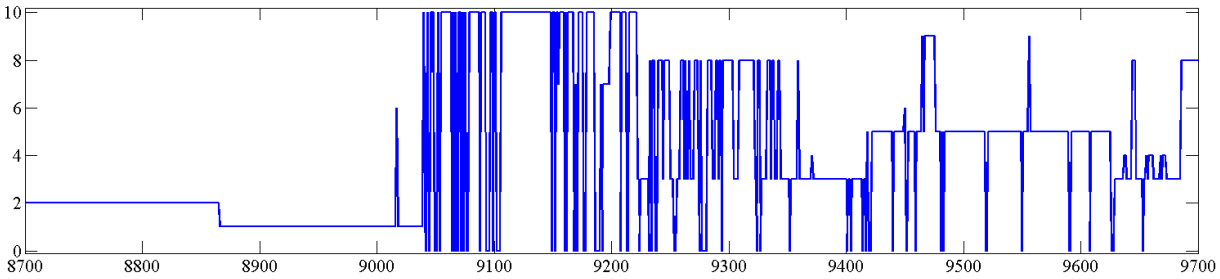


FIGURE 5.1 – Illustration des étiquettes des données distribuées sur 10 clusters, où 0 est l'étiquette des 20% de structures qui ont été rejetés dans un premier temps.

transitions). L'axe vertical représente les étiquettes des cluster, 0 étant l'étiquette des structures considérées comme atypiques après l'application de l'ACP à noyau. Dans cette illustration, il est clair que les données atypiques sont ubiquitaires. Le but est donc d'essayer de mieux classer les données pour en avoir une vue plus réaliste au niveau de leur stabilité. Dans cet objectif, nous avons utilisé les distributions des distances entre les plus proches voisins pour trouver un seuil qui définit le rejet en distance [146]. Le rejet en distance est basé sur la notion de distance aux clusters. L'idée consiste à rejeter une observation si la distance à un cluster plus grand qu'un seuil donné. Dans notre cas, on mesure la distance à un cluster par la mesure de la distance au plus proche voisin appartenant à ce cluster.

Les courbes de la figure 5.3 illustrent les distributions des distances proposées pour le peptide VAPGVG. Par exemple, la figure 5.3a illustre les distributions des distances entre les éléments du premier cluster avec leurs plus proches voisins du même cluster, ainsi que leur plus proche voisin dans des autres clusters. Dans chaque figure, il y a autant de courbes que de clusters (les indices d'évaluation de clustering ont conduit à retenir 3 clusters). Nous avons reclassé les données atypiques par rapport à ces trois clusters. Il est clair qu'à partir de chaque illustration de la figure 5.3, il est possible de définir un seuil de distance de rejet pour reclasser les données atypiques par rapport à chaque cluster séparément. L'illustration 5.3a nous permet, par exemple, de définir un seuil pour reclasser les données atypiques par rapport au premier cluster.

En bref, la procédure de reclassification peut être décrite par les étapes suivantes :

- Choisir un seuil s_i pour chaque cluster C_i par rapport à la distribution empirique des distances aux voisins.
  Dans notre cas, nous avons pris un seuil qui est égal à : (moyenne + 2.écart-type)
- Calculer les distances entre les données atypiques et les plus proches voisins

de chaque cluster.

- Assigner l'observation atypique au cluster C_i si la distance avec le plus proche voisin de ce cluster est inférieure à s_i.
- Si une observation atypique est associée à plusieurs clusters, nous la conservons comme une observation atypique - situation d'ambiguïté.

Ainsi, deux types de données atypiques sont enlevés : 1) celles qui sont loin de l'ensemble de la donnée (rejet de distance). 2) Celles qui sont entre les clusters (rejet d'ambiguïté). Pour le peptide VAPGVG, nous avons reclassé 2306 de 8000 structures initialement considérées comme atypiques. La figure 5.2 représente les distributions des clusters après la reclassification. Si nous superposons les figures 5.1 et 5.2, nous voyons que les données atypiques restantes sont présentes dans les parties où il y a des transitions entre les clusters (exemple : entre les instants 9000 et 9200). Par conséquent, cette opération permet de mieux préciser la notion de structures atypiques. Ceci permet d'analyser plus clairement le comportement des peptides au cours du temps et de manière plus rigoureuse qu'en définissant les données atypiques par une quantité arbitraire (20%).

5.3 Analyse de stabilité du peptide

5.3.1 Vérification de l'hypothèse de l'existence de transitions

Comme mentionné dans les parties précédentes, nous nous sommes basés dans notre travail sur l'hypothèse que les peptides alternent entres des états stables en passant par des états de transitions (instables) au cours du temps. Dans cette section, nous proposons une méthode qui nous permet d'illustrer ce changement dans le comportement du peptide et de confirmer cette hypothèse. La méthode se base

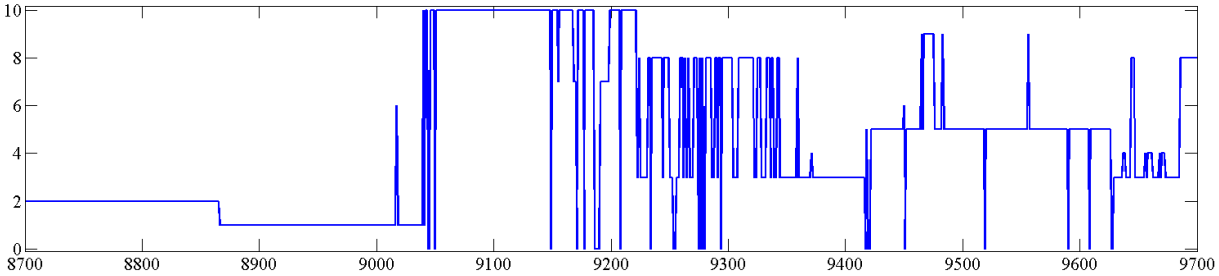
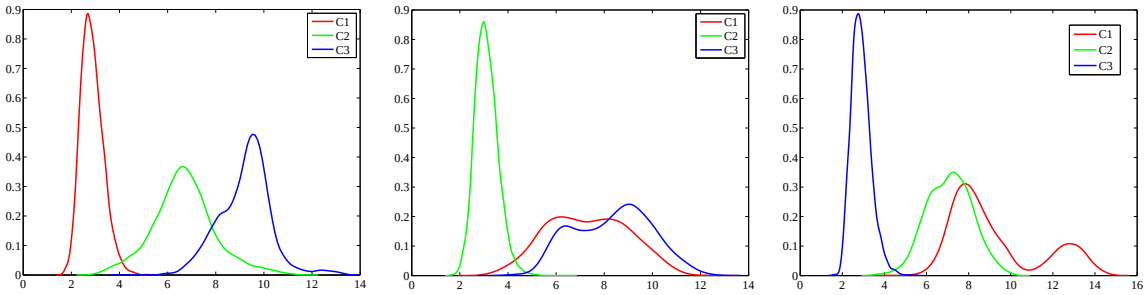


FIGURE 5.2 – Illustration des étiquettes des données reclassées sur 10 clusters, avec 0 pour étiquette des données rejetées.



(a) Par rapport aux éléments du premier cluster C1. (b) Par rapport aux éléments du deuxième cluster C2. (c) Par rapport aux éléments du troisième cluster C3.

FIGURE 5.3 – Schéma illustrant les distributions des distances entre les éléments de chaque cluster et la totalité des éléments du peptide considéré, C_i est i-ième cluster.

sur la mesure de distance moyenne entre chaque observation et les k observations qui les précèdent. Donc, k est la taille de la fenêtre que nous définissons derrière chaque observation pour illustrer le comportement du peptide à un instant donné t. Les observations dans la fenêtre ont les indices $[t-k-1, ...t-1]$. La figure 5.4 montre un exemple de cette mesure avec $k = 10$. En mettant cette dernière figure en correspondance avec la figure 5.5, nous remarquons que, quasiment à chaque saut dans la courbe bleue de la figure 5.4, il s'opère un changement entre les clusters dans la figure 5.5. En notant que la courbe de la figure 5.5 illustre la distribution des observations sur 10 clusters. Quelques exemples de ces changements sont illustrés par des flèches orange dans les deux figures ci-dessous. Ces illustrations montrent que notre hypothèse est confirmée et ceci nous a conduit à nous questionner sur la validation du nombre des conformations principales à l'aide de cette stabilité.

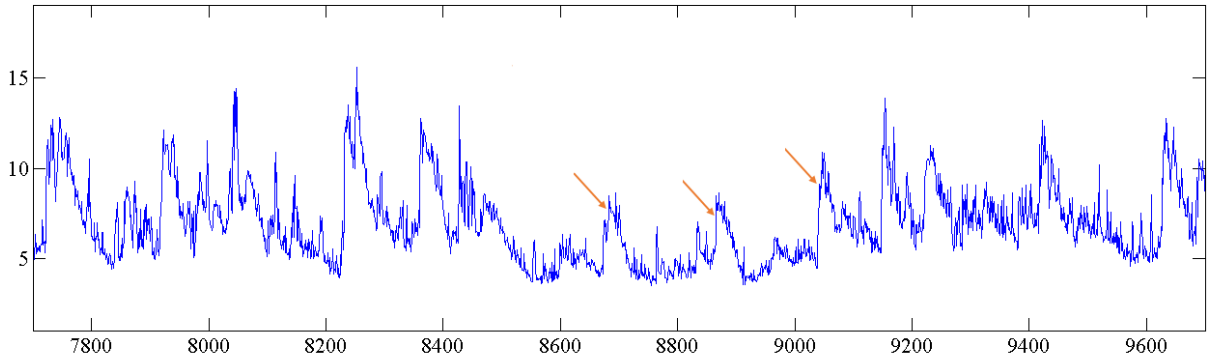


FIGURE 5.4 – Schéma représentant le comportement temporel d'un peptide entre les instants $[7700 - 9700]$. Cette courbe représente la distance moyenne entre chaque observation et les 10 observations qui les précèdent.

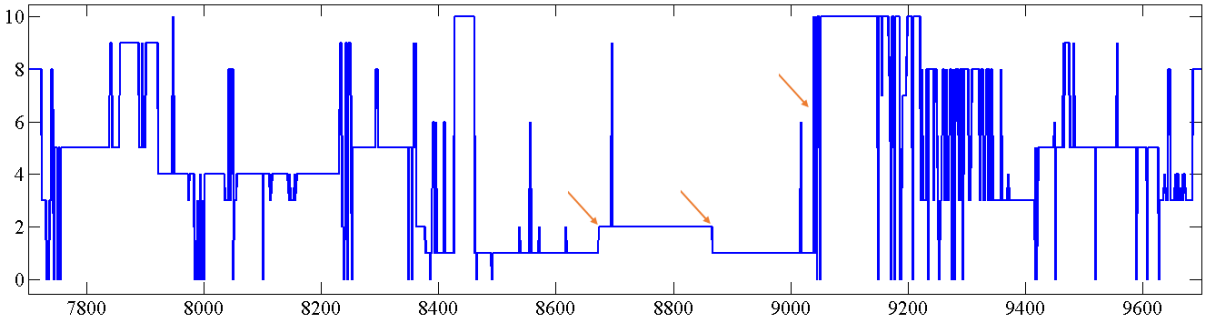


FIGURE 5.5 – Illustration de la distribution des structures du peptide sur 10 conformations principales entre les instants $[7700 - 9700]$. Les flèches orange indiquent les transitions entre les clusters.

5.3.2 Détermination du nombre des conformations principales

La complexité du clustering est due à l'instabilité du mouvement des peptides. Cette propriété rend l'estimation du nombre de conformations principales d'un peptide très difficile. Il est mentionné précédemment qu'avec les indices d'évaluation des méthodes de clustering, le nombre optimal de clusters obtenu pour n'importe quel peptide n'est pas cohérent. En effet, selon les biologistes, il est impossible d'avoir moins d'une dizaine de conformations pour chaque peptide. Pour cette raison, une autre approche a été imaginée pour estimer le nombre de conformations principales.

L'idée de cette approche provient de l'observation du comportement temporel du peptide dans les zones des transitions. Notre proposition consiste à tenter de corréler les ruptures temporelles (figure 5.4) de la conformation d'un peptide et les diagrammes de clusters pour voir si nous pouvons estimer le nombre des conformations principales dans les peptides.

Approche proposée

Cette approche se base sur l'idée que nous sommes sûrs qu'un peptide opère un changement dans son comportement temporel en passant d'une conformation à une autre. La problématique était donc de voir si ce changement pouvait nous aider à estimer le nombre de conformations principales. Nous supposons qu'en détectant les ruptures de la courbe qui représente les distances moyennes d'une conformation à la fenêtre glissante des conformations qui le précédent (illustrer dans la figure 5.4) et en les comparant aux instants des transitions obtenus après coupure du dendrogramme à différents niveaux, nous pouvons avoir une estimation du nombre de conformations

principales de chaque peptide. Par ailleurs, l'approche proposée est constituée des étapes suivantes :

- En premier lieu, pour un peptide donné (exemple VAPGVG), nous divisons le dendrogramme obtenu par notre méthode de clustering à plusieurs niveaux, où chaque niveau fournit une partition de données avec un nombre différent de clusters.
- En second lieu, pour chaque partition, nous déterminons les instants de transitions entre états (en rouge dans les illustrations de la figure 5.6).
  Ces transitions sont définies par le passage d'une zone stable à une zone instable (fig. 5.6a), ou d'une zone stable à une autre zone stable (fig. 5.6c), en se basant sur l'hypothèse qu'une zone stable se caractérise par la suite de 10 observations successives dans le même cluster. Ces transitions sont définies afin d'avoir une référence dans la comparaison qui sera appliquée dans la dernière étape.
- Ensuite, nous calculons la courbe qui illustre le comportement temporel du peptide en fixant une taille pour la fenêtre glissante et qui dans le cas présent est égale à 10 pour être en cohérence avec l'hypothèse du point précédent (Figure 5.4).
- Puis, nous appliquons la méthode de détection de ruptures décrite ci-dessous à la courbe de la figure 5.4 et on compare ces résultats avec les instants de transitions définies après découpage du dendrogramme à plusieurs niveaux (indiqué dans le deuxième point). Finalement, nous regardons à quels niveaux du dendrogramme correspondent ces détections.

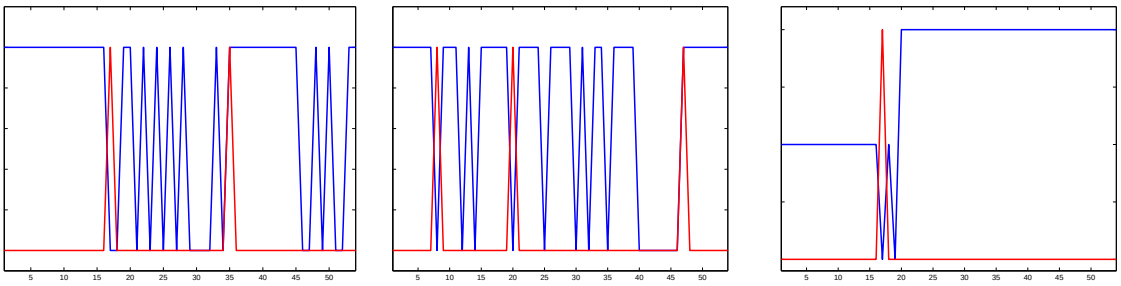
Méthode de détection de rupture

Il existe plusieurs méthodes pour la détection des ruptures [147]. Une de ces méthodes se base sur la séparation entre les distributions de données. Une façon de mesurer la séparation entre deux distributions empiriques est d'estimer la « divergence de Kullback-Leibler » [148]. Considérons :

— H_0 : $x_1, ..., x_i$ et $x_{i+1}, ..., x_j$ deux ensembles d'observations issues de la même distribution

— H_1 : $x_1, ..., x_i$ et $x_{i+1}, ..., x_j$ deux ensembles d'observations issues de 2 distributions différentes
  supposant que $x_1, ..., x_i$ suivent la loi de distribution $\widehat{f_0}$ et $x_{i+1}, ..., x_j$ suivent la loi $\widehat{f_1}$

— $D_{KL}(\widehat{f_0} \| \widehat{f_1}) \underset{H_1}{\overset{H_0}{\gtrless}} seuil$



(a) (b) (c)

FIGURE 5.6 – Exemples de transitions définies après coupure du dendrogramme. La courbe bleue représente les étiquettes de clusters. Les pics de la courbe rouge montrent les instants des transitions définies après classification. L'identification de ces instants est basée sur l'hypothèse qu'une zone stable se déduit par la présence de 10 structures successives dans le même cluster. Donc, une transition sera déterminée quand il y a un passage d'une zone stable à une zone instable, ou d'un passage d'une zone stable à une autre zone stable.

En fixant un seuil, une décision est prise pour indiquer quelle hypothèse est vraie H_0 ou H_1. Dans notre cas, nous l'appliquons comme ceci :

- A l'instant (**t**) de la courbe qui présente la distance moyenne entre chaque conformation et les 10 conformations qui la précédent (figure 5.4) :
  $\rightarrow \widehat{f_0}$ sont les données de **t**-10 à **t**-1
  $\rightarrow \widehat{f_1}$ sont les données de **t** à **t**+10
- Nous prenons $\widehat{f_0}$ comme référence, puis nous mesurons la distance de Kullback-Leibler entre les distributions empiriques de $\widehat{f_0}$ et $\widehat{f_1}$, en considérant que ces deux ensembles de données suivent une loi normale.

$$D_{KL}(\widehat{f_0}\|\widehat{f_1}) = (\frac{\nu_0}{\nu_1} + \frac{(m_1 - m_0)^2}{\nu_1} - 1 + ln\frac{\nu_1}{\nu_0}) \tag{5.1}$$

avec ν_0, ν_1 et m_0, m_1 les variances et les moyennes estimées de $\widehat{f_0}$ et $\widehat{f_1}$ respectivement.

- Ensuite, il faut appliquer un seuil pour détecter les pics qui sont obtenus avec la distance de Kullback-Leibler (courbe verte de la figure 5.7) afin de déterminer les ruptures de la courbe représentant le comportement temporel des conformations (courbe rouge de la figure 5.7).

Résultats de l'estimation du nombre de conformations

Pour vérifier l'utilité de cette approche, nous testons les résultats qui sont obtenus par cette proposition, grâce à la courbe COR qui évalue la correspondance

entre les ruptures détectées et les transitions déterminées à différents niveaux du dendrogramme.
L'algorithme utilisé pour accomplir la mesure peut être décrit par ces points :

- Pour chaque niveau de clustering dans le dendrogramme et pour n'importe quel peptide,
    - ∗ Nous identifions les instants de transitions obtenus par le clustering comme dans la figure 5.6.
        - ⋆ Puis nous prenons plusieurs valeurs de seuil pour détecter les ruptures de la courbe du comportement temporel (via la méthode de détection des ruptures).
        - ⋆ Ensuite pour chaque seuil :
            - ◇ Nous comptons les ruptures détectées,
            - ◇ Nous estimons la probabilité de détection (la correspondance entre les indices des ruptures détectées par la distance de Kullback-Leibler et les indices des transitions définis manuellement après clustering),
            - ◇ Nous calculons la probabilité de fausse alarme.

Finalement, pour chaque niveau du dendrogramme nous obtenons une courbe COR. La figure 5.8 représente l'AUC (Area Under Curve) des courbes COR en fonction du nombre de clusters obtenu par le clustering.

À travers la courbe d'AUC obtenue dans la figure 5.8, il apparaît que les résultats ne font pas apparaître de valeurs particulières par rapport au différent nombre de cluster, impliquant des résultats imprécis. La valeur maximum d'AUC est obtenue avec un indice égale à 3, impliquant 3 clusters. Cela ne répond pas à nos besoins. Il

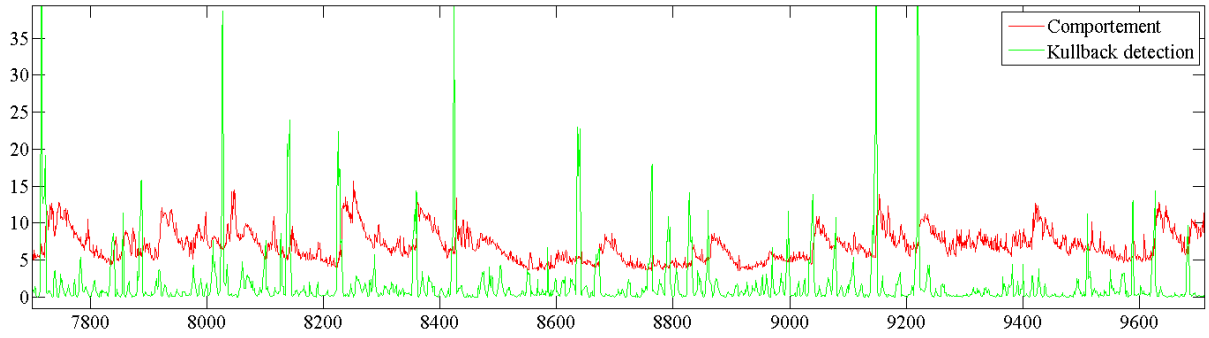


FIGURE 5.7 – illustration de la distance de Divergence de Kullback-Leibler (courbe verte) appliquée au comportement dynamique du peptide (courbe rouge) avec une fenêtre = 10.

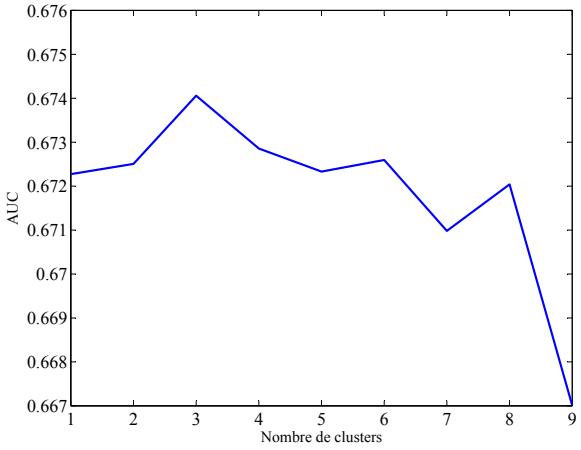


FIGURE 5.8 – Illustration de l'AUC en fonction du nombre de clusters obtenus à chaque niveau du dendrogramme.

était prévu d'avoir une valeur maximum d'AUC avec un nombre de clusters supérieur à trois. En revanche, cette approche a paru relativement sensible au choix des paramètres pris en compte, dont quatre sont à optimiser : taille de la fenêtre, taille de la zone de stabilité, le seuil de détection et le nombre d'observations utilisées dans le calcul de distance Kullback-Leibler. Cette optimisation requérait plus d'efforts que ne nécessitait la présente recherche. Nous avons donc interrompu ce travail et nous nous sommes concentrés sur l'objectif principal qui consistait à comparer les conformations répétitives de chaque peptide pour identifier la conformation associée à leur activité. Néanmoins, ces travaux d'optimisation et d'étude sur la stabilité des peptides peuvent être envisagés comme une perspective pour affiner ces travaux.

5.4 Comparaison des peptides

Dans le chapitre 4 nous avons étudié chaque peptide séparément. Notre objectif dans cette section est de comparer les conformations principales de tous peptides. La difficulté réside essentiellement dans la taille respective des peptides analysés. Pour comparer les conformations principales de différents peptides, il est crucial d'identifier quelle méthode de comparaison convient le mieux pour .
Dans le chapitre 4 et plus précisément dans la section 4.7.1, nous avons observé que les résultats du clustering ne sont pas les mêmes lorsque les atomes des chaînes latérales sont inclus en plus de ceux du backbone. Cependant, cette analyse a été menée avec des structures qui appartiennent au même peptide et donc de taille identique. La présente comparaison est réalisée sur des structures de tailles différentes et qui n'appartiennent pas au même peptide. De surcroît, les peptides ont des chaînes latérales différentes. En appliquant le même principe

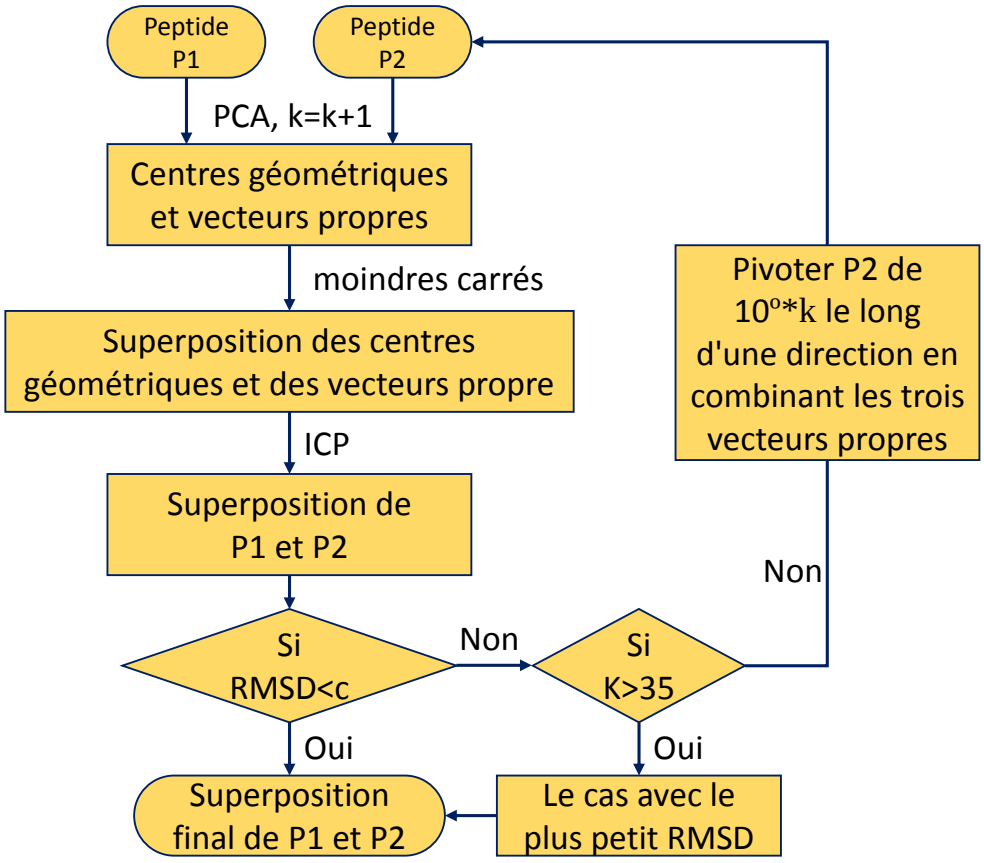


FIGURE 5.9 – Organigramme représentant l'algorithme de superposition de structures de protéines par paires pour les données manquantes.

qu'au chapitre 4 se pose la question de la mise en correspondance des atomes de ces chaînes latérales. Par conséquent, il est important d'utiliser des méthodes qui sont capables de comparer ces conformations en prenant en considération les différences dans la chaîne latérale et dans les backbones avec des atomes manquants.

Dans la littérature, une méthode récente traite la comparaison des structures protéiques en prenant en compte les problèmes mentionnés ci-dessus. Cette méthode se base sur la superposition des corps rigides pour aligner structuralement les conformations peptidiques. Elle a montré son efficacité par rapport aux méthodes antérieures [138]. Elle est présentée dans la section 4.2.4 et la figure 5.9 illustre son principe de traitement qui consiste à superposer les centres de gravité des 2 peptides comparés puis à choisir la superposition qui a la meilleure valeur de RMSD (Éq.(4.1)). Grâce à cette méthode, nous sommes capables de comparer tous les peptides et d'observer les similitudes entre leurs différentes conformations principales.

Peptide	APGVGV	GVGVAP	PGVGVA	VAPGVG	GVAPGV	PGAIPG	VGVAPG	EGFEPG	LGTIPG
Nombre des conformations	13	14	18	13	11	17	15	18	15

TABLE 5.1 – Nombre des clusters obtenus après découpage de dendrogramme au niveau 200 de tous les peptides.

5.4.1 Méthode de superposition rigide

Pour appliquer cette méthode de superposition, il faut avant tout disposer les conformations principales de chaque peptide. Cependant, nous ne connaissons pas pour l'instant le nombre de conformations principales qu'il faut extraire de chaque ensemble des structures. Nous avons vu que nous ne pouvions pas nous baser sur les indices d'évaluation de clustering pour obtenir cette information. De plus, avec nos études sur la stabilité des peptides (section 5.3.2), nous n'avons pas réussi à estimer cette valeur.

En parallèle, nous savons que, pour les biologistes, il n'est pas logique d'avoir moins d'une dizaine de conformations pour chaque peptide. Pour cette raison et dans un premier temps, nous avons fixé un niveau de dissimilarité pour tous les dendrogrammes de peptides et nous avons coupé chaque dendrogramme à ce niveau afin d'obtenir un ensemble de conformations plus grandes que 10 pour chaque peptide. Le tableau 5.1 montre le nombre des clusters obtenus pour un niveau de dissimilarité égal à 200 pour tous les peptides. Ensuite, nous avons appliqué la méthode de superposition illustrée par la figure 5.9 sur les conformations principales obtenues pour tous les peptides. Malgré la performance remarquable de cette méthode par rapport aux méthodes existantes de l'état de l'art, cette méthode n'a pas fourni de bons résultats dans ce cas précis. Elle a détecté des similarités entre des conformations qui en réalité, ne sont pas similaires.
La figure 5.10 confirme cette conclusion. Elle représente les valeurs de dissimilarité entre chaque conformation et la conformation la plus proche de chaque peptide. Chaque colonne appartient à un peptide et chaque ligne correspond à une conformation. Les conformations sont classées par ordre des peptides. Les quatre premiers peptides sont les peptides inactifs et les 5 suivants sont les peptides actifs. La case (**i**,**j**) contient la valeur de dissimilarité entre la conformation i et la conformation la plus proche du peptide j. Un seuil a été fixé à 0.1 par l'auteur de cette méthode pour

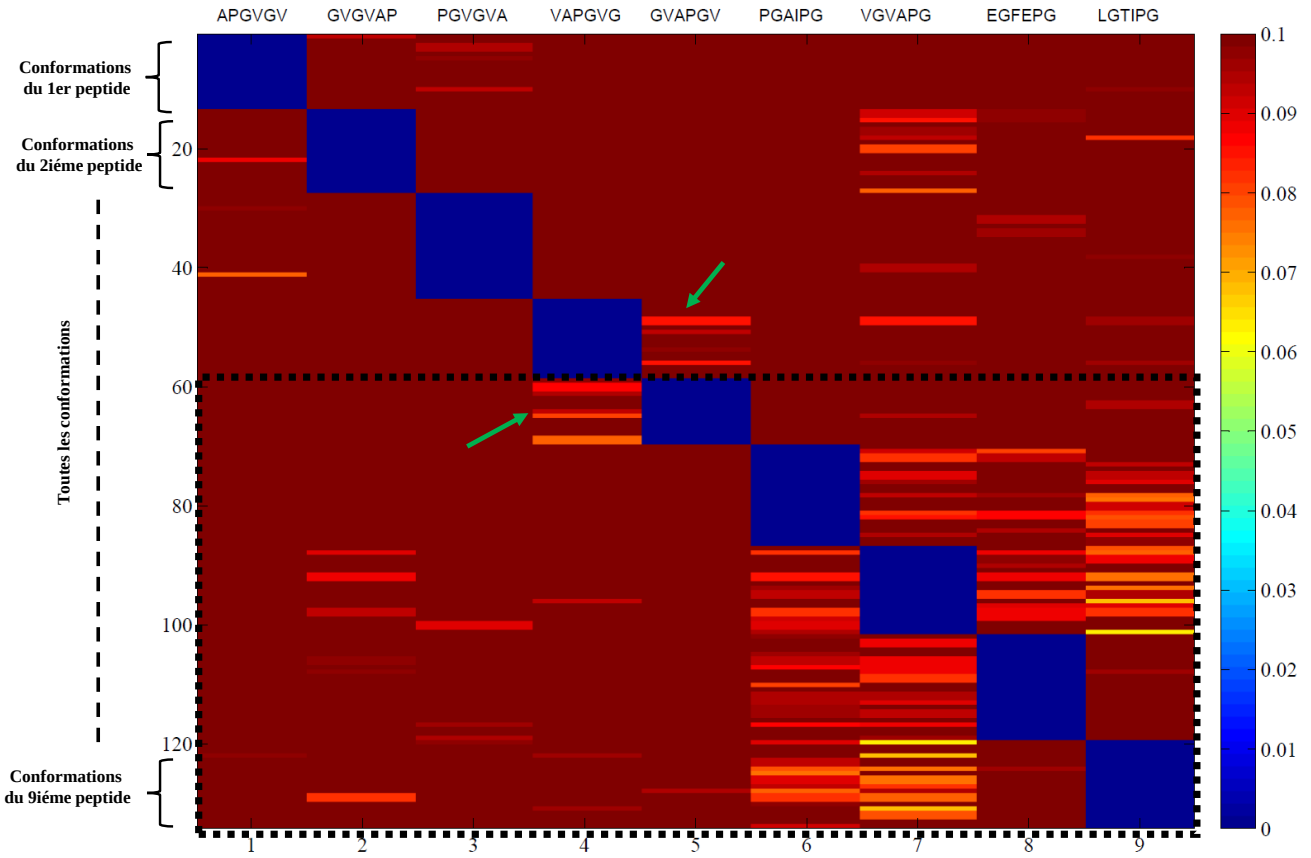


FIGURE 5.10 – Dissimilarité entre chaque conformation et son plus proche voisin dans les autres peptides. Chaque ligne représente une conformation. Chaque colonne représente un peptide. La valeur de dissimilarité qui a l'indice (**i**,**j**) représente la conformation **i** qui a un voisin dans le peptide **j**, avec une valeur de dissimilarité plus petite que 0.1.

identifier les conformations similaires. En fait, ce qui nous intéresse dans la figure 5.10 est la partie située dans le cadre en pointillé correspondant aux conformations où les peptides actifs se situent.

Cette méthode montre des similarités entre certaines conformations des peptides 4 et 5 (VAPGVG, GVAPGV) contrairement à nos hypothèses (les conformations marquées par les flèches vertes dans la figure 5.10). Donc, soit nos hypothèses sont fausses, soit la méthode établit des similitudes à tort. En examinant les peptides superposés, nous trouvons une similitude erronée. La figure 5.11 montre un exemple de cette mauvaise prédiction, mais il en existe beaucoup d'autres ; c'est donc la méthode qui n'est pas adaptée à nos données. Il y a plusieurs raisons possibles pour cette mal-prédiction de similarité : une de ces raisons revient à la sensibilité de cette méthode aux positions initiales prises pour commencer la superposition des conformations [138].

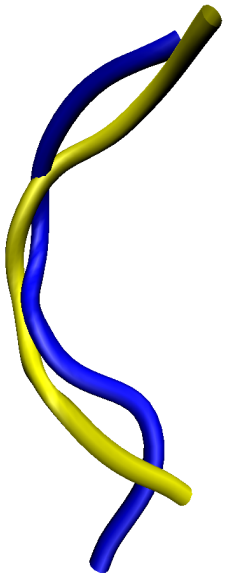


FIGURE 5.11 – Illustration des conformations qui sont prétendues être similaires par la méthode de superposition rigide, mais qui sont en réalité différentes. La couleur jaune revient au peptide VAPGVG et la couleur bleue foncée au peptide GVAPGV.

5.4.2 Détection des formes principales

En raison des mauvais résultats obtenus par la méthode utilisée ci-dessus, mais aussi du coût de calcul élevé de la procédure de superposition, nous avons proposé d'utiliser la méthode présentée au chapitre 4 pour faire la comparaison. Elle consiste à utiliser la matrice de distance pour représenter chaque conformation, puis calculer la distance euclidienne entre elles afin de mesurer la similarité, en gardant à l'esprit que seuls les atomes des backbones seront pris en compte dans le calcul étant donné que le nombre d'atomes dans tous les peptides n'est pas le même.

La figure 5.12 représente les valeurs de dissimilarité entre chaque conformation et la conformation la plus proche de chaque peptide. Les lignes et les colonnes sont les mêmes que dans la figure 5.10. Seule la mesure de dissimilarité change entre ces 2 figures. Elles n'ont pas également la même échelle, car elles ne représentent pas la même mesure. La figure 5.10 contient les valeurs RMSD entre les structures superposées, tandis que la figure 5.12 représente les distances euclidiennes entre les matrices de distances des conformations. Afin d'identifier les conformations similaires résultantes de notre méthode, il est nécessaire de définir un seuil qui identifie les conformations identiques au sein des cinq peptides actifs sans les confondre avec les peptides inactifs.

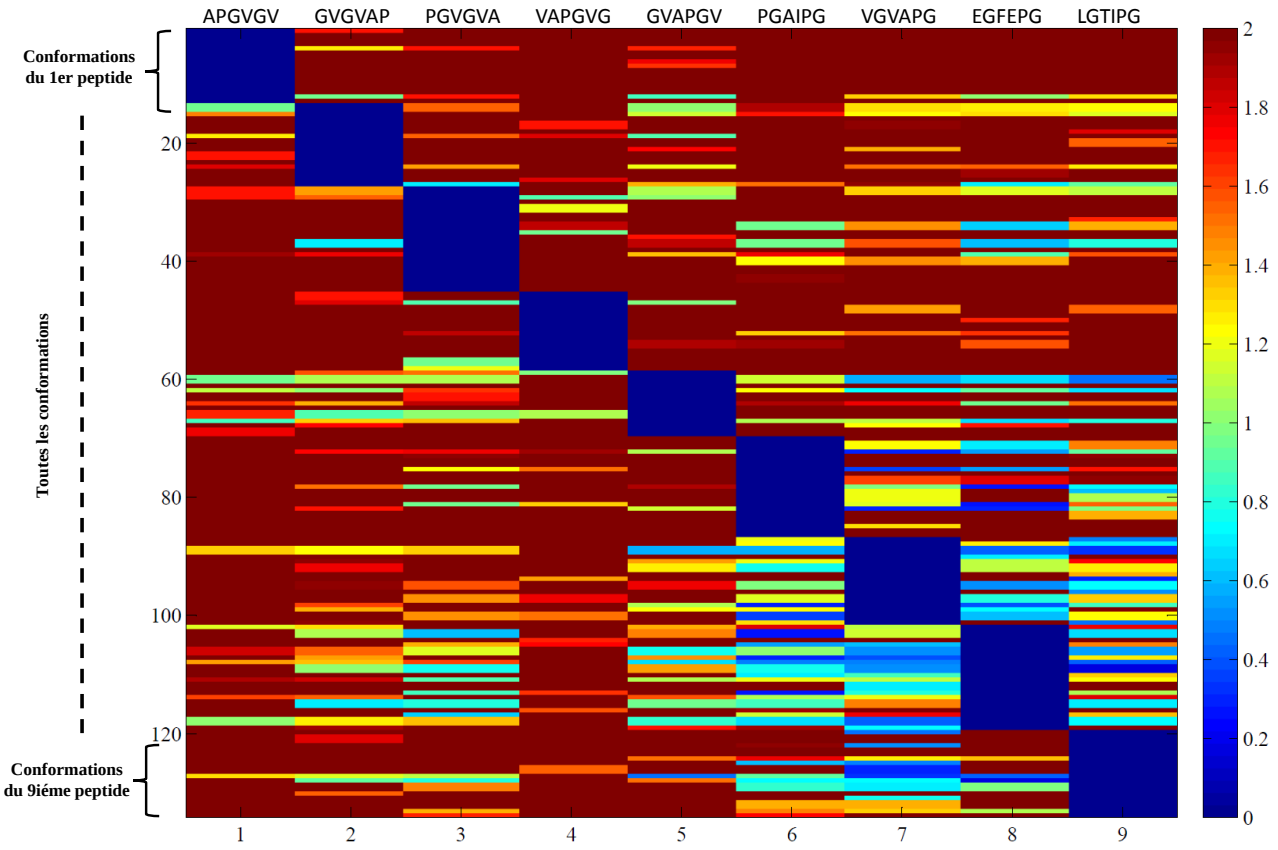


FIGURE 5.12 – Dissimilarité entre chaque observation et sa plus proche conformation dans chaque peptide. Chaque colonne correspond à un peptide. Chaque ligne correspond à une conformation. La couleur bleue indique une forte similarité et la couleur rouge montre une forte dissimilarité.

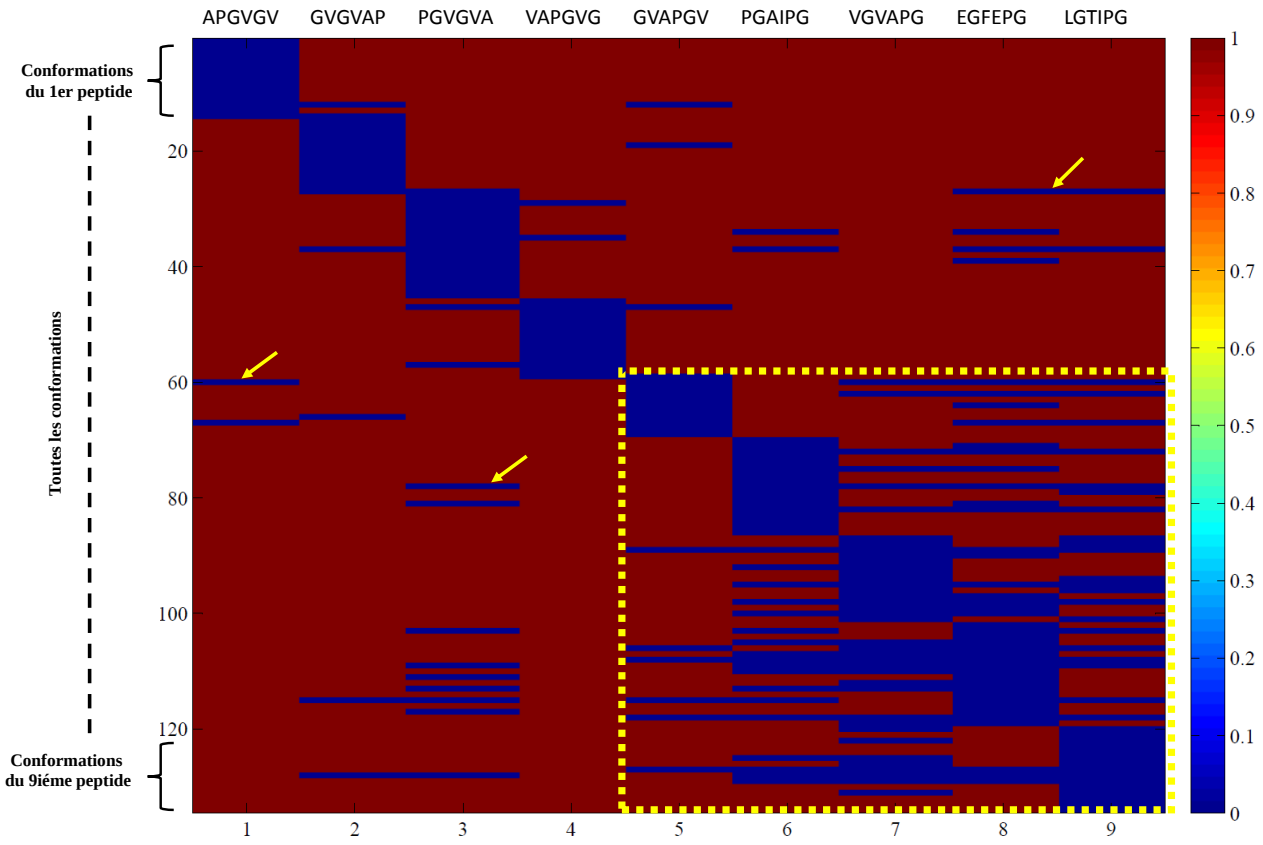


FIGURE 5.13 – Seuillage du tableau des dissimilarités de la figure 5.12. Cette figure représente les conformations identiques pour chaque peptide. La couleur bleue indique la similarité et la couleur rouge indique la dissimilarité.

Dans la figure 5.12, il est clair qu'il y a plus de similarité dans les cinq dernières colonnes (peptides actifs) que dans les quatre premières qui appartiennent aux peptides inactifs. Visuellement, cela permet de prédire la ressemblance entre les conformations des peptides actifs. Mais cette méthode nécessite également la définition d'un seuil. Nous avons besoin d'un seuil pour identifier les conformations similaires. Nous avons appliqué les étapes suivantes pour déterminer ce seuil : tout d'abord nous fixons une valeur importante pour le seuil, puis nous commençons à le réduire jusqu'à obtenir au moins une ligne bleue qui passe par les 5 dernières colonnes, si elle existe, sans avoir de mauvaises estimations de similarités avec les peptides inactifs. Cela conduit à trouver les conformations communes entre les peptides actifs (les 5 dernières colonnes). En suivant cette procédure, nous sommes parvenus à un seuil de 1.

La figure 5.13 représente les conformations identifiées identiques après seuillage des résultats de la figure 5.12. Il est apparent qu'il y a plusieurs conformations communes pour les cinq peptides actifs. Cela peut être déduit par les lignes bleues

qui passent conjointement par les cinq dernières colonnes. En effet, en regroupant les conformations qui correspondent à ces lignes bleues dans la partie des peptides actifs qui est entourée par le cadre en pointillés jaunes dans la figure 5.13, nous trouvons 6 conformations références qui sont présentées dans le tableau 5.2.

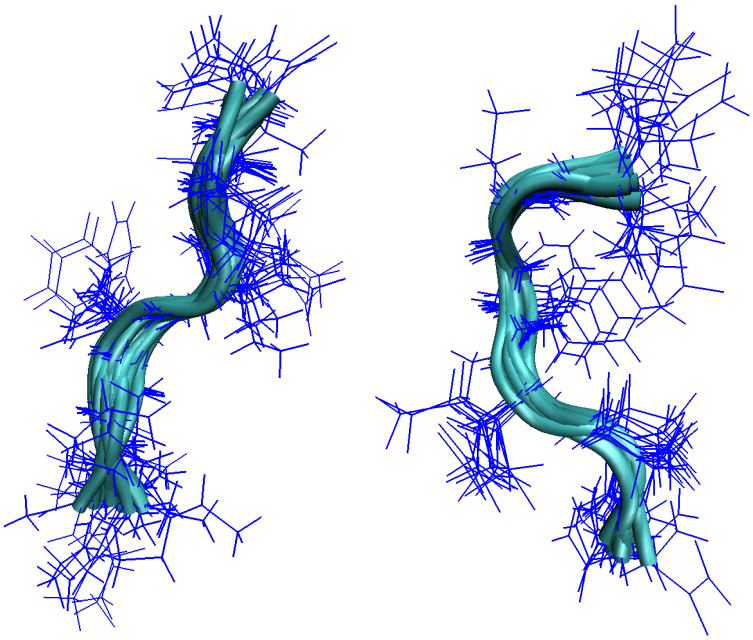
En effet, chaque tiret bleu dans cette matrice correspond à une conformation principale pour un peptide. Chaque ligne bleue dans le cadre pointillé correspond à une conformation référence (par exemple Conf 1 du tableau 5.2) regroupant les conformations principales similaires de peptides actifs. Il existe des lignes bleues dans ce cadre qui ne passent pas par tous les peptides actifs et représente aussi des conformations références. Ainsi, elles ont des conformations communes avec d'autres lignes. Par conséquent, en regroupant toutes ces lignes (impliquant le regroupement des conformations références correspondantes à toutes les lignes bleues du cadre jaune pointillé) nous obtenons les 6 conformations références présentées dans le tableau 5.2. Ce tableau représente les conformations principales de chaque peptide regroupées dans les 6 conformations références nommées de Conf 1 à Conf 6. Dans ce tableau, les peptides sont différenciées par des lettres (e, f, g, h, i), et les conformations principales de chaque peptide sont différenciées par des numéro. Prenant l'exemple du peptide EGFEPG, il a 5 conformations principales (5 clusters) labellisées par (h7, h17, h6, h5 et h4) groupés dans la conformation référence Conf 1, les 6 conformations principales labellisées par (h14, h2, h8, h9, h12 et h18) dans la conformation référence Conf 2 et enfin la conformation labellisée par (h11) dans la conformation référence Conf 4.

D'après ce tableau, les deux premières conformations références sont les plus récurrentes dans les peptides actifs. Pour chaque peptide, plusieurs conformations principales sont regroupées avec ces deux premières conformations références. Cela veut dire que ces conformations principales, qui appartiennent au même peptide (comme g3, g12, g6 et g14 du peptide VGVAPG), ont des similarités également entre elles-mêmes. La figure 5.14 illustre les deux premières conformations référence du tableau 5.2.

En se basant sur les résultats illustrés dans la figure 5.13 et le tableau 5.2, nous pouvons déduire que la conformation référence "conf 1", représentée dans la figure 5.14, est toujours présente dans les peptides actifs et n'est pas présente dans les peptides inactifs. Par contre, pour la deuxième conformation référence "conf 2", elle est similaire à plusieurs conformations des peptides inactifs. Un exemple de ces conformations est marqué par des flèches jaunes dans la figure 5.13. Donc, avec ces résultats nous pourrions affirmer la véracité de l'hypothèse

Peptide \ Conformation	Conf 1	Conf 2	Conf 3	Conf 4	Conf 5	Conf 6
GVGVAP	e2,e9,e4					
PGAIPG	f13,f3,f6	f9,f2,f12	f10			
VGVAPG	g3,g12,g6 g14	g9,g11,g2 g13	g1,g10	g4	g15	g8
EGFEPG	h7,h17,h6 h5,h4	h14,h2,h8 h9,h12,h18		h11		
LGTIPG	i8	i9,i10	i6,i3,i12		i1	i7

TABLE 5.2 – Représentation des conformations principales références (Conf 1,....,Conf 6) des peptides actifs obtenus après découpage de dendrogrammes à un niveau 200. Chaque colonne correspond à une conformation de référence représentant un ensemble des conformations similaires de différents peptides. fi est le label de la i-ième conformation principale (cluster) du peptide. Chaque ligne de ce tableau contient uniquement les clusters d'un seul peptide labellisés selon leurs ordres parmi les clusters obtenus.



(a) Première conformation (b) Deuxième conformation

FIGURE 5.14 – Les conformations qui sont obtenues au niveau 200 dans les dendrogrammes et qui sont prises des peptides actifs seulement.

Peptide	APGVGV	GVGVAP	PGVGVA	VAPGVG	GVAPGV	PGAIPG	VGVAPG	EGFEPG	LGTIPG
Nombre des conformations	9	9	10	10	7	9	8	9	10

TABLE 5.3 – Nombre des clusters obtenus après découpage du dendrogramme au niveau 300 de tous les peptides.

présentée au début de notre travail, qui liait l'activité du peptide aux structures qui existent dans les peptides actifs et qui n'existent pas dans les peptides inactifs, et la confirmer par la présence de la conformation référence Conf 1 uniquement dans les peptides actifs. Il faut donc maintenant proposer une solution pour inférer l'activité des peptides dont le niveau d'activité est inconnu en profitant de cette similarité entre les peptides actifs. Mais avant cela, ces résultats nous amènent aussi à la problématique de savoir s'il est réellement nécessaire d'identifier le vrai nombre de clusters dans chaque peptide afin de trouver la conformation clé (Conf 1).

En s'appuyant sur le mode de fonctionnement du dendrogramme, si nous découpons les dendrogrammes à un niveau plus haut que 200, il est sûr que le nombre de clusters qui appartiennent à la même conformation référence sera réduit par rapport à ceux qui sont obtenus avec ce même niveau. Cela nous conduit à nous demander si à un niveau de dissimilarité du dendrogramme plus haut (> 200), nous pourrions toujours retomber sur la même conformation principale référence Conf 1 ou pas. Notons que, la conformation référence Conf 1 est considérée maintenant comme la clé de l'activité des peptides actifs.

5.4.3 Influence du nombre des clusters retenus

Pour répondre à cette problématique qui propose d'identifier les conformations principales des peptides actifs malgré l'ignorance du nombre de conformations dans chaque peptide, nous avons coupé les dendrogrammes de tous les peptides à un niveau de dissimilarité supérieur à 200, 300 dans ce cas, et avec lequel nous pouvons obtenir un nombre de clusters raisonnable (autour de 10) pour chaque peptide. Le nombre de clusters obtenus avec ces clusterings est représenté dans le tableau 5.3.
Les mesures des dissimilarités entre chaque conformation et la conformation la plus proche de chaque peptide sont illustrées dans la figure 5.15 : il apparaît qu'il existe toujours des similarités entre les peptides actifs (cinq dernières colonnes), absentes dans les peptides inactifs (quatre premières colonnes).
En regroupant les conformations qui correspondent à toutes les lignes bleues de cette matrice de dissimilarité nous trouvons les mêmes conformations références obtenues avec le niveau 200. Le tableau 5.4 représente les conformations principales

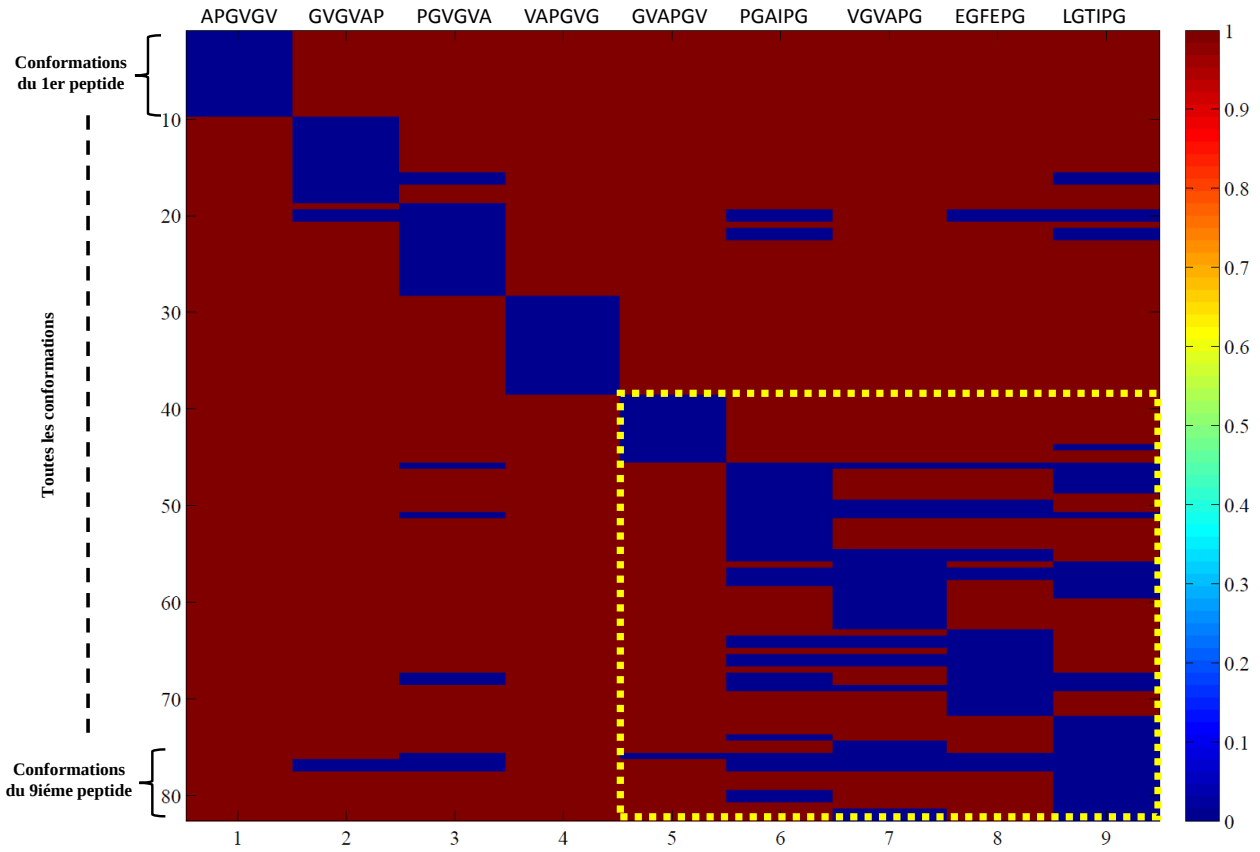
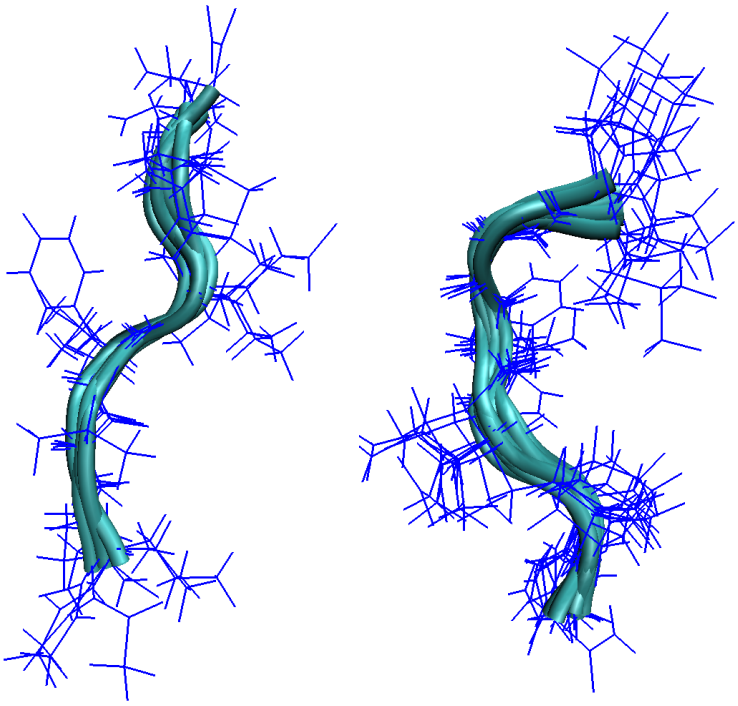


FIGURE 5.15 – Seuillage du tableau de similarité qui est obtenu au niveau 300 des dendrogrammes des peptides. Elle représente les conformations identiques par la couleur bleue.

Peptide \ Conformation	Conf 1	Conf 2	Conf 3	Conf 4	Conf 5	Conf 6
GVGVAP	e6					
PGAIPG	f6	f1, f5	f2	f3		
VGVAPG	g3	g4,g1			g2	g5
EGFEPG	h7	h6,h4				
LGTIPG	i5	i6,i8	i3	i9	i4	i11

TABLE 5.4 – Représentation des conformations principales des peptides actifs obtenues après découpage de dendrogrammes à un niveau 300.



(a) Première conformation (b) Deuxième conformation

FIGURE 5.16 – Représentation des 2 conformations principales qui sont obtenues au niveau 300 dans la dendrogramme.

de chaque peptide regroupées dans ces 6 conformations références. La conformation clé est aussi identifiée à ce niveau de dendrogramme (300) et elle est représentée par leur forme 3D dans la figure 5.16. Ceci valide notre proposition précédente. Suite à ces démonstrations, il apparaît que connaître le nombre exact de clusters n'est pas indispensable à l'identification des conformations principales communes chez les peptides actifs et qu'un nombre de clusters raisonnable et logique par rapport à la dynamique du peptide est suffisant.

Nous savons que notre peptide est très élastique et dynamique et que pour des raisons d'identification du nombre de conformations principales dans chaque peptide, il serait anormal d'avoir un nombre de clusters égaux à $3, 4$ ou 5 ; il est préférable que celui-ci soit plus grand pour identifier plus précisément les conformations existantes. De plus, cela aide à souligner les similarités entre peptides actifs si elles existent. Enfin, nous voulons maintenant essayer de profiter de cette similarité entre les conformations des peptides actifs afin de classifier les peptides inconnus comme étant actifs ou non. Dans la partie suivante, nous proposons une méthode qui peut être efficace pour détecter l'activité des peptides.

5.5 Classification des peptides inconnus

Pour finaliser notre travail, il fallait proposer une méthode permettant de classifier les peptides inconnus entre peptides actifs ou inactifs, ou tout du moins, de reconnaître leur comportement comme proche de celui des peptides actifs ou pas. Les trois peptides inconnus sont PGAYPG, VGLAPG, VVGPGA. Dans les sections précédentes, nous avons induit des exemples analysés l'hypothèse que l'activité des peptides est liée à leur capacité à adopter des conformations précises au cours de leur dynamique moléculaire. Pour cela, nous avons décidé d'ajouter les conformations de ces trois derniers peptides inconnus aux matrices qui illustrent les dissimilarités entre les 134 identifiées à l'aide des 9 peptides d'apprentissage. La figure 5.17 montre le résultat binarisé en appliquant le même seuillage que sur les peptides inconnus (seuil égal à 1).

À travers la figure 5.17, il est clair que les deux peptides PGAYPG et VGLAPG ont plus de similarités avec les peptides actifs qu'avec les peptides inactifs. De plus, chacun a une conformation similaire à la conformation référence 'Conf 1" ; impliquant le même comportement des peptides actifs. Par contre la dernière colonne est très différente des autres : elle n'a quasiment aucune conformation similaire avec les autres peptides. Il est donc nécessaire de rechercher une mesure statistique afin de vérifier ces résultats et mettre en évidence cette similarité.
Pour ce faire, nous proposons un détecteur d'activité des peptides. Ce détecteur se base sur la classification des peptides par rapport aux conformations de références résultantes de la figure 5.13. Les conformations similaires sont regroupées séparément pour les peptides actifs et les peptides inactifs. Une seule structure est utilisée pour représenter chaque conformation de référence. Pour les peptides actifs, nous obtenons six conformations de références qui sont indiquées dans le tableau 5.2. Pour les peptides inactifs, il n'y a pas de similarités entre eux. Pour cette raison, les 58 conformations des quatre peptides inactifs sont prises en compte. Finalement, le but est d'assigner la totalité des structures de chaque peptide à ses 64 conformations ($58 + 6$ conformations). Les figures 5.18a et 5.18b représentent le nombre d'individus des populations assignées à chaque conformation de référence pour les peptides actifs et les peptides inactifs respectivement. L'axe (x) représente les conformations de références. Les conformations de références des peptides non actifs sont numérotées de 1 à 58. Le reste couvre les conformations références des peptides actifs. À travers la figure 5.18b, il est très clair que les peptides inactifs n'ont pas de similarité avec les conformations des peptides actifs. Les quatre courbes qui sont dans la figure 5.18b ont un effectif négligeable de l'indice 59 à l'indice 64, indiquant que ces peptides n'adoptent quasiment jamais les conformations des peptides actifs. Par contre, pour les peptides actifs, dans la figure 5.18a, nous

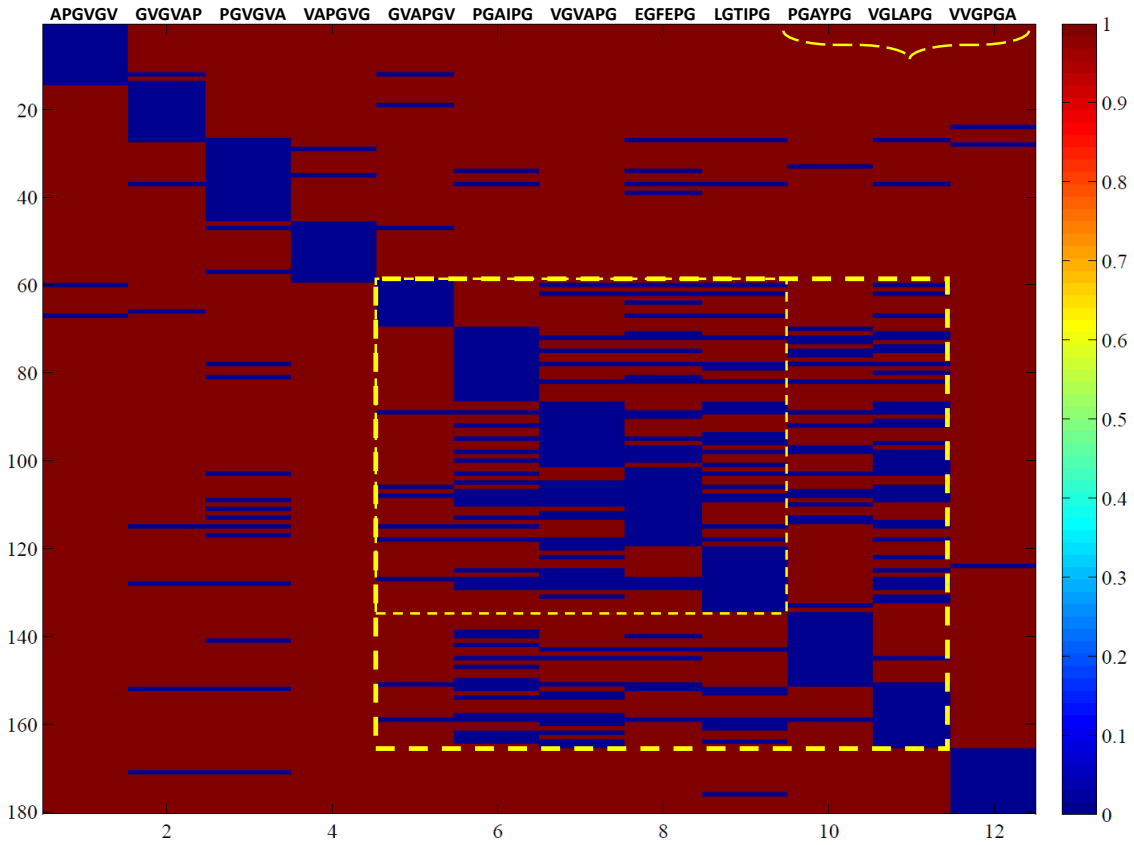
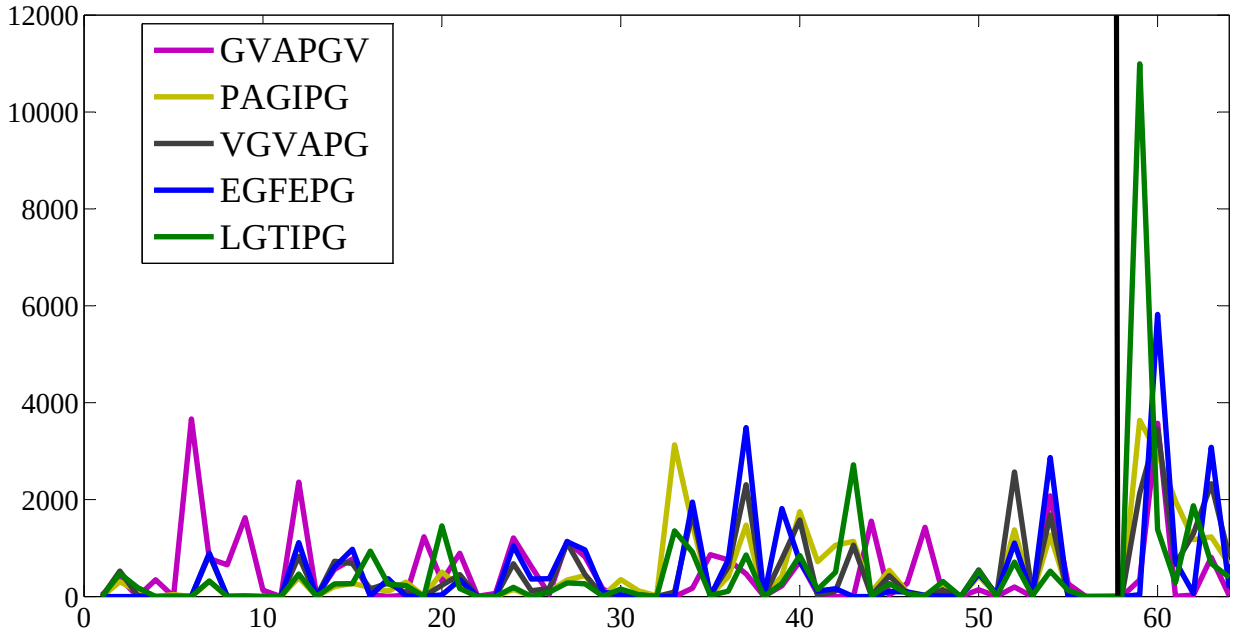
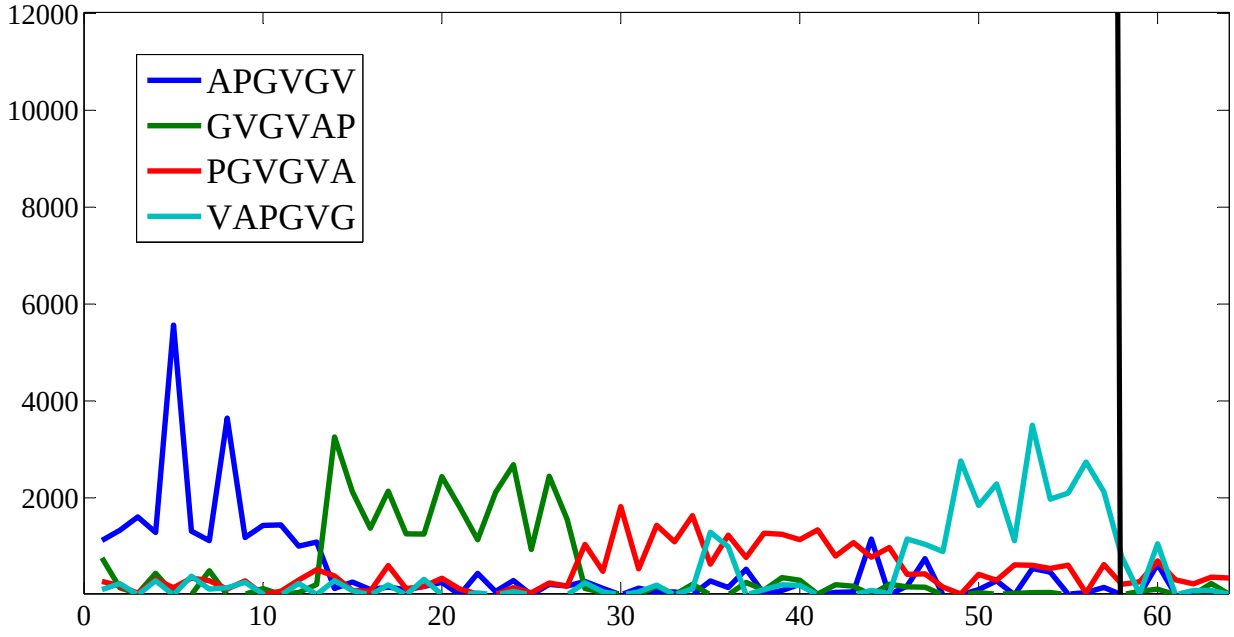


FIGURE 5.17 – La similarité entre chaque conformation est leur plus proche voisin dans chaque peptide. Chaque ligne correspond à une conformation. Chaque colonne correspond à un peptide. La couleur bleue indique qu'il existe deux conformations similaires. Les 3 dernières colonnes correspondent aux peptides que l'on cherche à classer comme actifs ou inactifs.



(a) Pour les peptides actifs.



(b) Pour les peptides inactifs.

FIGURE 5.18 – Représentation des effectifs des éléments proches à chaque conformation. L'axe x correspond aux conformations des peptides actifs et inactifs. Les premières 58 conformations appartiennent aux peptides inactifs et les 6 dernières conformations reviennent aux conformations des peptides actifs.

Peptide \ Indice des conformations	$1 \rightarrow 58$ (inactif)	$59 \rightarrow 64$ (actif)
APGVGV (inactif)	**539.70**	116.16
GVGVAP(inactif)	**544.12**	73.5
PGVGVA(inactif)	**513.60**	368.5
VAPGVG(inactif)	**529.89**	211
GVGVAP(actif)	469.27	**797**
PGAIPG(actif)	351.18	**1938.5**
VGVAPG(actif)	367.36	**1782.2**
EGFEPG(actif)	380.9	**1650.7**
LGTIPG(actif)	282.5	**2602.5**
PGAYPG(inconnu)	435.2	**1126.3**
VGLAPG(inconnu)	372.41	**1733.3**
VVGPGA(inconnu)	**516.5**	340

TABLE 5.5 – Représentation de l'effectif moyen des structures de chaque peptide dans les conformations qui appartiennent aux peptides inactifs (première colonne) et peptides actifs (deuxième colonne)

voyons qu'ils ont des conformations qui sont partagées avec les peptides inactifs. Cependant à partir de l'indice 59 de l'axe (x) ces peptides adoptent fréquemment ces conformations de références. La différence est tout à fait nette.

En définitive, nous remarquons que les peptides actifs ont une diversité dans leurs conformations, mais certaines sont plus récurrentes que d'autres. Pour les peptides inactifs, chacun possède ses propres conformations, différente des autres et celles qu'ils partagent avec les peptides actifs sont négligeables. Le tableau 5.5 représente l'effectif moyen des structures de chaque peptide dans les conformations qui appartiennent aux peptides inactifs d'une part et aux peptides actifs d'autre part. L'effectif moyen des peptides inactifs entre les indices 59 à 64 est minimal par rapport aux peptides actifs. D'ailleurs, pour les peptides inconnus, il semblerait que PGAYPG et VGLAPG ont le comportement de peptides actifs. Par contre, le peptide VVGPGA a le comportement d'un peptide inactif. Ces trois peptides nous avaient été donnés en aveugle pour le test, la validité biologique ayant été testée.

Pour conclure au niveau de cette classification, grâce aux résultats obtenus dans

cette section, nous pouvons voir les différences entre les peptides actifs et les peptides inactifs. Pour avoir un système de classification performant, il est nécessaire d'avoir un très grand nombre de peptides pour réaliser le modèle de classification et un nombre important des peptides pour le tester. Avec la base de données existante, nous ne pouvons pas encore généraliser notre système de classification, mais grâce à l'application sur les peptides inactifs, nous avons vu une différence remarquable entre PGAYPG et VGLAPG et les peptides inactifs, qui est très similaire à la différence entre les peptides actifs et les peptides inactifs. Avec les données actuelles, nous pouvons pronostiquer que PGAYPG et VGLAPG sont actifs et que VVGPGA est inactif. Finalement, ces résultats sont validés par l'équipe de biologie de notre projet.

5.6 Conclusion

Dans ce chapitre, le but était d'analyser les conformations des peptides actifs et des peptides inactifs afin d'identifier leurs différences et de déterminer la structuration clé de l'activité d'un peptide. Au début de notre travail, l'hypothèse émise était que l'activité des peptides est due à la présence de quelques conformations dans les peptides actifs au cours de leurs trajectoires moléculaires et que leur absence cause l'inactivité de ces peptides. Cette hypothèse a été construite sur l'idée avérée que l'activité des peptides est fortement liée à leurs structures 3D et à leur interaction avec les récepteurs de ces peptides.

Afin d'atteindre cet objectif, il était crucial de passer par des étapes intermédiaires pour préparer les données traitées. Nous avons été confrontés aux problèmes du nombre inconnu de conformations stables de chaque peptide et à celui de la comparaison des conformations de peptides différents. Nous nous sommes efforcés de résoudre ces questionnements.

En premier lieu, parce qu'ils ne sont pas cohérents avec l'expertise métier, nous avons pu observer que les critères d'évaluation de clustering ne sont pas pertinents pour nos données. Par conséquent, il était très important de trouver une autre approche. Tout d'abord, nous avons confirmé l'hypothèse qui indique que les peptides ont des états de stabilité et passent entre eux via des états de transitions. Cela a été montré par l'illustration du comportement temporel du peptide en utilisant une fenêtre glissante sur toutes les observations afin de mesurer le mouvement du peptide au cours de temps. Puis, nous nous sommes basés sur cette mesure pour essayer d'estimer le nombre de conformations principales. Malgré la relation observée entre cette mesure et les distributions des clusters au cours de temps, nous n'avons pas réussi à estimer ce nombre inconnu.

Pour la comparaison des conformations des peptides qui sont de différentes tailles, nous avons utilisé une méthode de comparaison basée sur la superposition de corps rigides qui a montré son efficacité sur des bases de données de protéines bien connues. Mais en l'appliquant sur nos peptides, cette méthode permet de conclure à la similarité de structures différentes au sens de l'activité, ce qui n'est pas satisfaisant. La taille et la flexibilité extrême de nos objets, par rapport à des protéines, sont certainement à l'origine de ce non-fonctionnement. Par conséquent, nous avons utilisé le même concept que dans le chapitre 4, qui consiste à comparer les peptides dans un espace des distances inter-atomes. Bien que dans le chapitre 4, nous ayons noté que la chaîne latérale a un rôle très important dans le clustering, pour la comparaison inter-peptides, nous sommes obligés de ne considérer que les atomes du backbone. Grâce à cette approche, nous avons vu que les peptides actifs se caractérisent par une similarité entre certaines de leurs conformations qui n'existe pas entre les peptides inactifs qui sont très différents. De plus, il existe une conformation principale (clé) dans les peptides actifs qui n'existe pas avec les peptides inactifs. Ainsi, le nombre exact des conformations n'est pas indispensable : tant que nous avons une idée du comportement des peptides, nous pouvons en proposer un nombre raisonnable en lien avec l'expertise pour enfin trouver les mêmes conformations communes entre les peptides. D'ailleurs, plus nous augmentons le nombre de clusters dans chaque peptide, plus nous obtenons de conformations similaires dans le même peptide. De ce fait, il n'est pas nécessaire de couper le dendrogramme ni à un niveau très bas, ni à un niveau très haut, comme nous l'avons expérimentalement montré.

Finalement, il suffit d'avoir une estimation assez vague du nombre de conformations principales pour identifier les conformations de références des peptides actifs. Ainsi, les peptides actifs ont une "identité" qui est différente des peptides inactifs. Cette identité se caractérise par la convergence des conformations des peptides vers quelques conformations précises, au contraire des peptides inactifs qui ont tous des conformations différentes les unes des autres. Ceci nous a permis de proposer une méthode pour créer un détecteur d'activité afin de classifier les trois peptides inconnus (PGAYPG, VGLAPG, VVGPGA). Ce détecteur est basé sur la classification des peptides par rapport aux conformations références qui sont obtenues avec les peptides actifs et les peptides inactifs. Un peptide sera classé actif, s'il a une présence moyenne dans les conformations des peptides actives plutôt que dans les conformations qui appartiennent aux peptides inactifs, à l'inverse des peptides inactifs.

Chapitre 6

Conclusion

Ce travail de thèse a pour finalité biologique de mieux comprendre les signaux envoyés par le corps lorsque son système vasculaire se dégrade, lors du vieillissement ou de pathologies chroniques. La partie élastique des parois vasculaire doit son élasticité à une protéine nommée élastine qui, lorsqu'elle est dégradée, relargue dans l'organisme de petits morceaux appelés peptides qui vont ensuite jouer le rôle de messagers, indiquant à l'organisme qu'une dégradation s'opère. Le système cardiovasculaire est l'un des systèmes vitaux le corps humain, puisqu'il est responsable de la distribution du sang dans toutes les cellules. Ainsi toute maladie qui affecte le système cardiovasculaire peut être une cause de la détérioration de l'ensemble du corps. Ces peptides envoient des signaux qui peuvent être bénéfiques, néfastes, ou anodins pour le reste de l'organisme. Certains d'entre eux peuvent avoir des effets sur l'évolution de pathologies cardiovasculaires. De nombreuses recherches sont réalisées pour identifier et comprendre le rapport entre cette dégradation et l'apparition et/ou l'évolution des maladies cardiovasculaires. La plupart d'entre-elles sont faites du côté biologie/physiologie. Récemment, les spécialistes du domaine ont commencé à utiliser les outils mathématiques existants pour prédire les fonctionnalités liées aux protéines et aux peptides. Ces outils ont surtout été utilisés pour identifier la forme spatiale des protéines et des peptides, afin d'essayer de déduire leurs interactions entre protéines, entre protéines et peptides, afin de mieux appréhender la cascade d'évènements moléculaires qui mènent à une réponse physiologique de l'organisme. Ce travail de thèse a aussi pour but de contribuer à "ce dialogue entre biologistes et statisticiens appliqués" en créant une méthodologie qui permet d'analyser les conformations des peptides très flexibles produits lors de la dégradation des élastines, et d'ainsi essayer d'isoler in silico les peptides qui auront un rôle biologique de ceux qui n'en auront pas. Ceci ferait gagner un temps immense aux biologistes en pré-triant les molécules d'intérêt thérapeutique.

Dans le deuxième chapitre de ce manuscrit, les motivations biologiques de notre travail ont été introduites. Nous avons expliqué l'importance d'étudier le

système cardiovasculaire, ainsi que ses composants, ses constituants cellulaires et structuraux. Enfin, nous avons décrit les motivations du notre sujet de thèse et les solutions proposées afin d'aborder ce problème.

Après avoir exposé les motivations de notre sujet et les solutions possibles pour le traiter, un état de l'art sur les méthodes possibles est abordé dans le troisième chapitre. La notion de noyaux a été définie, ainsi que l'ensemble des méthodes de détection des données atypiques. Finalement, des méthodes de classification multiclasse utiles à notre travail ont été présentées en détail.

Dans le chapitre 4, les problèmes de l'analyse de conformations au niveau de chaque peptide ont été posés. En se basant sur l'hypothèse qu'un peptide, au cours de temps, alterne entre un certain nombre de conformations principales, passant de l'une à l'autre via des états de transitions, l'objectif de notre travail a été de déterminer ces conformations dites "principales". Pour le faire, nous proposons une méthode de classification des structures peptidiques tridimensionnelles, dérivées de l'élastine, qui sont hautement flexibles. Nous avons introduit une stratégie originale qui combine différentes méthodologies statistiques afin de détecter les conformations principales d'un peptide au cours d'un laps temps dans une expérience de simulation de dynamique moléculaire. La matrice des distances entre les atomes de chaque structure est utilisée comme espace de représentation invariant aux translations et rotations pour éviter la mise en œuvre de méthode de recalage. De plus, l'utilisation de l'ACP à noyau pour détecter les formes transitoires entre les conformations principales est appliquée pour la première fois. Elle permet de limiter l'impact de ces structures transitoires sur la méthode de classification qui suit.La méthode de classification hiérarchique a été utilisée sur les données restantes pour assurer une grande flexibilité dans la classification des différentes conformations et afin de laisser la liberté à l'utilisateur d'ajuster la méthode de classification selon ses données et ses exigences. Des tests et des comparaisons sur la classification avec et sans les chaînes latérales ont été réalisés et montrent l'importance des chaînes latérales dans la procédure de détermination des conformations principales pour ce type de peptides.

Dans le dernier chapitre, le travail portait sur l'analyse des conformations principales de différents peptides actifs et inactifs. Le but était d'identifier la différence entre les peptides actifs et inactifs en se basant sur leurs conformations principales. Des méthodes de comparaison des conformations de différentes tailles ont été utilisées pour accomplir les travaux, mais celles-ci n'ont pas fourni les résultats attendus. Pour cette raison, nous avons ré-utilisé la méthode du chapitre

4 qui fait la comparaison dans l'espace des distances inter-atomes, en ne prenant en compte que les atomes des backbones. Les résultats obtenus après comparaisons montrent que les peptides actifs se caractérisent par des similarités entre ces conformations qui sont différentes de celles qui caractérisent les peptides inactifs. Ceci nous a poussés à profiter de cette signature pour créer un détecteur d'activité pour les peptides, où ce détecteur se base sur la classification des peptides par rapport aux conformations principales des peptides actifs. Finalement, cela nous a permis de classifier les peptides inconnus et de prévoir leur activité.

Suite aux travaux présentés dans le cadre de cette thèse, nous avons identifié certaines améliorations qui pourraient faire l'objet de perspectives à court terme :

— Classifier les peptides en utilisant uniquement les chaînes latérales.

— Créer une interface Web permettant à d'autres utilisateurs d'essayer cette méthode sur leurs propres banques de peptides.

Parmi les perspectives à plus long terme de cette thèse figurent :

— Trouver une méthode de comparaison inter-peptides plus adaptée aux conformations qui ont des petites tailles et qui est capable de prendre en compte la différence dans les chaînes latérales.

— Analyser les conformations transitionnelles enlevées par l'ACP à noyau et essayer de voir si les peptides ont un parcours bien défini ou bien changent arbitrairement.

— Améliorer le détecteur d'activité de peptides en essayant de trouver un critère d'évaluation d'activité en se basant seulement sur les conformations des peptides actifs.

— Appliquer le détecteur d'activité proposé dans notre travail sur d'autres peptides et mesurer sa puissance de détection.

Bibliographie

[1] circulatory system, 2012. URL https://www.slideshare.net/tristan87/unit-b-section Online ; accessed 2-juin-2018.

[2] Peeter Värnik. Suicide in the world. *International journal of environmental research and public health*, 9(3) :760–771, 2012.

[3] Heart Anatomy. Internal structure of human heart, August 28, 2017. URL https://anatomybodysystem.com/internal-structure-of-human-heart/. Online ; accessed 20-April-2018.

[4] Anatomy and Physiology. The cardiovascular system : Blood vessels, December 21, 2013. URL http://anatomyandphysiologyi.com/cardiovascular-system-blood-vessels/. Online ; accessed 2-juin-2018.

[5] Sarah Lowe Dan Rogers. The arterial system, December 21, 2013. URL http://teachmeanatomy.info/the-basics/ultrastructure/blood-vessels/. Online ; accessed 2-juin-2018.

[6] Collagen Health Benefits. The arterial system, Janvier 27, 2018. URL https://www.uninvesthub.com/article/archives/01-2018. Online ; accessed 3-Mars-2018.

[7] Suzanne M Mithieux and Anthony S Weiss. Elastin. In *Advances in protein chemistry*, volume 70, pages 437–461. Elsevier, 2005.

[8] Clemens Žváček, Gerald Friedrichs, Leonhard Heizinger, and Rainer Merkl. An assessment of catalytic residue 3d ensembles for the prediction of enzyme function. *BMC bioinformatics*, 16(1) :359, 2015.

[9] www.ck12.org. The arterial system, Feb 23, 2012. URL https://www.ck12.org/book/CK-12-Chemistry-Second-Edition. Online ; accessed 3-Mars-2018.

[10] Mounzer Boubou. *Contribution aux méthodes de classification non supervisée via des approches prétopologiques et d'agrégation d'opinions*. PhD thesis, 2007.

URL http://www.theses.fr/2007LYO10287. Thèse de doctorat dirigée par Lamure, Michel Informatique Lyon 1 2007.

[11] Liisa Holm and Chris Sander. Protein structure comparison by alignment of distance matrices. *Journal of molecular biology*, 233(1) :123–138, 1993.

[12] William HE Day and Herbert Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1) :7–24, 1984.

[13] HN Sallam. L'ancienne école de médecine d'alexandrie. *Gynécologie Obstétrique & Fertilité*, 30(1) :3–10, 2002.

[14] J R Levick. Chapter 1 - overview of the cardiovascular system. In *An Introduction to Cardiovascular Physiology*, pages 1 – 12. 1991. URL http://www.sciencedirect.com/science/article/pii/B9780750610285500044.

[15] Neha Jadeja Pagidipati and Thomas A Gaziano. Estimating deaths from cardiovascular disease : a review of global methodologies of mortality measurement. *Circulation*, 127(6) :749–756, 2013.

[16] World Health Organization et al. Global atlas on cardiovascular disease prevention and control. 2011.

[17] Melanie Nichols, Nick Townsend, Peter Scarborough, and Mike Rayner. *European cardiovascular disease statistics*. European Heart Network, 2012.

[18] Edward G Lakatta. Arterial and cardiac aging : major shareholders in cardiovascular disease enterprises : Part iii : cellular and molecular clues to heart and arterial aging. *Circulation*, 107(3) :490–497, 2003.

[19] S Martin and R Andriantsitohaina. Mécanismes de la protection cardiaque et vasculaire des polyphénols au niveau de l'endothélium. In *Annales de Cardiologie et d'Angéiologie*, volume 51, pages 304–315. Elsevier, 2002.

[20] Julian F Thayer and Richard D Lane. The role of vagal function in the risk for cardiovascular disease and mortality. *Biological psychology*, 74(2) :224–242, 2007.

[21] Organisation mondiale de la Santé. Régime alimentaire, nutrition et prévention des maladies chroniques. *Genève : Rapport d'une Consultation OMS/FAO d'experts. OMS, Série de Rapports techniques*, 916, 2003.

[22] Marina Cecelja and Phil Chowienczyk. Role of arterial stiffness in cardiovascular disease. *JRSM cardiovascular disease*, 1(4) :1–10, 2012.

[23] Henry E Kim, Seema S Dalal, Erik Young, Marianne J Legato, Myron L Weisfeldt, and Jeanine D'Armiento. Disruption of the myocardial extracellular matrix leads to cardiac dysfunction. *The Journal of clinical investigation*, 106 (7) :857–866, 2000.

[24] William D Tucker and Steve S Bhimji. Anatomy, blood vessels. 2017.

[25] Kenneth S Saladin and Leslie Miller. *Anatomy & physiology*. WCB/McGraw-Hill New York (NY), 1998.

[26] Fabrice Schneider. *Remodelage de la paroi artérielle : étude des aspects de destruction et de reconstruction*. PhD thesis, Université Paris-Est, 2011.

[27] Marie-Paule Jacob. Matrice extracellulaire et vieillissement vasculaire. *médecine/sciences*, 22(3) :273–278, 2006.

[28] Wassim Fhayli. *Evaluation de l'action de traitements chroniques par l'extrait d'aneth ou le minoxidil en tant que nouvelles pharmacothérapies antivieillissement du système cardiovasculaire chez la souris*. PhD thesis, 2013. URL `http://www.theses.fr/2013GRENV039`. Thèse de doctorat dirigée par Faury, Gilles Physiologie physiopathologies pharmacologie Grenoble 2013.

[29] Peter Fratzl. Collagen : structure and mechanics, an introduction. In *Collagen*, pages 1–13. Springer, 2008.

[30] E Petersen, F Wågberg, and K-A Ängquist. Serum concentrations of elastin-derived peptides in patients with specific manifestations of atherosclerotic disease. *European journal of vascular and endovascular surgery*, 24(5) :440–444, 2002.

[31] Cay M Kielty, Michael J Sherratt, and C Adrian Shuttleworth. Elastic fibres. *Journal of cell science*, 115(14) :2817–2828, 2002.

[32] Cay M Kielty, Clair Baldock, David Lee, Matthew J Rock, Jane L Ashworth, and C Adrian Shuttleworth. Fibrillin : from microfibril assembly to biomechanical function. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 357(1418) :207–217, 2002.

[33] JETAL Rosenbloom, WR Abrams, and R Mecham. Extracellular matrix 4 : the elastic fiber. *The FASEB Journal*, 7(13) :1208–1218, 1993.

[34] FX Maquart, G Bellon, S Pasco, and JC Monboisse. Matrikines in the regulation of extracellular matrix degradation. *Biochimie*, 87(3-4) :353–360, 2005.

[35] Jean-Michel Claverie and Cedric Notredame. *Bioinformatics for dummies*. John Wiley & Sons, 2011.

[36] Srinivasan Damodaran. Amino acids, peptides and proteins. *Fennema's food chemistry*, 4 :217–329, 2008.

[37] Andrii Stanovych. *Synthèse et études structurales de γ-peptides synthétisés à partir d'acides β, γ-diaminés*. PhD thesis, Paris 11, 2014.

[38] Viachaslau M Barodka, Brijen L Joshi, Dan E Berkowitz, Charles W Hogue Jr, and Daniel Nyhan. Implications of vascular aging. *Anesthesia and analgesia*, 112(5) :1048, 2011.

[39] Cay M Kielty. Elastic fibres in health and disease. *Expert reviews in molecular medicine*, 8(19) :1–23, 2006.

[40] Charlotte Kawecki, Laurent Duca, Sébastien Blaise, Béatrice Romier, Hassan el Btaouri, Philippe Gillery, Laurent Debelle, and Pascal Maurice. Vieillissement matriciel et impacts vasculaires. *Hématologie*, 21(4) :221–229, 2015.

[41] S Baydanoff, G Nicoloff, and Ch Alexiev. Age-related changes in the level of circulating elastin-derived peptides in serum from normal and atherosclerotic subjects. *Atherosclerosis*, 66(1) :163–168, 1987.

[42] Jes Sanddal Lindholt, Lene Heickendorff, EW Henneberg, and H Fasting. Serum-elastin-peptides as a predictor of expansion of small abdominal aortic aneurysms. *European journal of vascular and endovascular surgery*, 14(1) : 12–16, 1997.

[43] Edward R Smith, Laurie A Tomlinson, Martin L Ford, Lawrence P McMahon, Chakravarthi Rajkumar, and Stephen G Holt. Elastin degradation is associated with progressive aortic stiffening and all-cause mortality in predialysis chronic kidney disease. *Hypertension*, pages HYPERTENSIONAHA–111, 2012.

[44] Robert M Senior, Gail L Griffin, Robert P Mecham, David S Wrenn, Kari U Prasad, and Dan W Urry. Val-gly-val-ala-pro-gly, a repeating peptide in elastin, is chemotactic for fibroblasts and monocytes. *The Journal of cell biology*, 99(3) :870–874, 1984.

[45] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2014.

[46] Maria Laura Maag, Ludovic Denoyer, and Patrick Gallinari. Graph anonymization using machine learning. In *Advanced Information Networking and Applications (AINA), 2014 IEEE 28th International Conference on*, pages 1111–1118. IEEE, 2014.

[47] Christine L Lisetti and Fatma Nasoz. Affective intelligent car interfaces with emotion recognition. In *Proceedings of 11th International Conference on Human Computer Interaction, Las Vegas, NV, USA*. Citeseer, 2005.

[48] Tsung Fu Lin and Yan Ping Chi. Application of webpage optimization for clustering system on search engine v google study. In *Computer, Consumer and Control (IS3C), 2014 International Symposium on*, pages 698–701. IEEE, 2014.

[49] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[50] Alex J Smola and Bernhard Schölkopf. *Learning with kernels*, volume 4. Citeseer, 1998.

[51] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *Journal of machine learning research*, 12(Jul) :2211–2268, 2011.

[52] Antoine Cornuéjols and Laurent Miclet. *Apprentissage artificiel : concepts et algorithmes*. Editions Eyrolles, 2011.

[53] Charles A Micchelli. Interpolation of scattered data : distance matrices and conditionally positive definite functions. In *Approximation theory and spline functions*, pages 143–145. Springer, 1984.

[54] Ralf Herbrich. *Learning Kernel classifiers : theory and algorithms (adaptive computation and machine learning)*. MIT press, 2002.

[55] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3) :293–300, 1999.

[56] Vladimir Cherkassky and Yunqian Ma. Practical selection of svm parameters and noise estimation for svm regression. *Neural networks*, 17(1) :113–126, 2004.

[57] Albert D Shieh and David F Kamm. Ensembles of one class support vector machines. In *International Workshop on Multiple Classifier Systems*, pages 181–190. Springer, 2009.

[58] David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54(1) :45–66, 2004.

[59] Heiko Hoffmann. Kernel pca for novelty detection. *Pattern Recognition*, 40 (3) :863–874, 2007.

[60] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7) :1443–1471, 2001.

[61] David MJ Tax and Piotr Juszczak. Kernel whitening for one-class classification. In *Pattern recognition with support vector machines*, pages 40–52. Springer, 2002.

[62] David MJ Tax and Robert PW Duin. Support vector domain description. *Pattern recognition letters*, 20(11-13) :1191–1199, 1999.

[63] Douglas M Hawkins. *Identification of outliers*, volume 11. Springer, 1980.

[64] John A Quinn and Christopher KI Williams. Known unknowns : Novelty detection in condition monitoring. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 1–6. Springer, 2007.

[65] Lei Clifton, David A Clifton, Peter J Watkinson, and Lionel Tarassenko. Identification of patient deterioration in vital-sign data using one-class support vector machines. In *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*, pages 125–131. IEEE, 2011.

[66] Animesh Patcha and Jung-Min Park. An overview of anomaly detection techniques : Existing solutions and latest technological trends. *Computer networks*, 51(12) :3448–3470, 2007.

[67] V Jyothsna, VV Rama Prasad, and K Munivara Prasad. A review of anomaly based intrusion detection systems. *International Journal of Computer Applications*, 28(7) :26–35, 2011.

[68] Christopher P Diehl and John B Hampshire. Real-time object classification and novelty detection for collaborative video surveillance. In *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on*, volume 3, pages 2620–2625. IEEE, 2002.

[69] Markos Markou and Sameer Singh. A neural network-based novelty detector for image sequence analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10) :1664–1677, 2006.

[70] Hugo Vieira Neto and Ulrich Nehmzow. Real-time automated visual inspection using mobile robots. *Journal of Intelligent and Robotic Systems*, 49(3) :293–307, 2007.

[71] Boris Sofman, Bradford Neuman, Anthony Stentz, and J Andrew Bagnell. Anytime online novelty and change detection for mobile robots. *Journal of Field Robotics*, 28(4) :589–618, 2011.

[72] Yang Zhang, Nirvana Meratnia, and Paul Havinga. Outlier detection techniques for wireless sensor networks : A survey. *IEEE Communications Surveys & Tutorials*, 12(2) :159–170, 2010.

[73] Sugato Basu, Mikhail Bilenko, and Raymond J Mooney. A probabilistic framework for semi-supervised clustering. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 59–68. ACM, 2004.

[74] Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99 :215–249, 2014.

[75] BS Everitt and A Skrondal. The cambridge dictionary of statistics. 2002. *Cambridge, Cambridge*.

[76] Baback Moghaddam. Principal manifolds and probabilistic subspaces for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6) :780–788, 2002.

[77] Jérémie Kellner. *Gaussian models and kernel methods*. PhD thesis, Lille 1, 2016.

[78] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3) :337–404, 1950.

[79] Guy Cohen. Optimisation des grands systemes. 2004.

[80] Yunqiang Chen, Xiang Sean Zhou, and Thomas S Huang. One-class svm for learning in image retrieval. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 1, pages 34–37. IEEE, 2001.

[81] David MJ Tax and Robert PW Duin. Data domain description using support vectors. In *ESANN*, volume 99, pages 251–256, 1999.

[82] Ian T Jolliffe. Principal component analysis and factor analysis. In *Principal component analysis*, pages 115–128. Springer, 1986.

[83] Bernhard Schölkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John C Platt. Support vector method for novelty detection. In *Advances in neural information processing systems*, pages 582–588, 2000.

[84] Harry H Harman. *Modern factor analysis*. University of Chicago Press, 1976.

[85] Peter HA Sneath, Robert R Sokal, et al. *Numerical taxonomy. The principles and practice of numerical classification.* 1973.

[86] Daniel Barbará and Sushil Jajodia. *Applications of data mining in computer security*, volume 6. Springer Science & Business Media, 2002.

[87] Depa Pratima and Nivedita Nimmakanti. Pattern recognition algorithms for cluster identification problem. *International Journal of Computer Science & Informatics*, 1(1) :2231–5292, 2012.

[88] Anil K Jain and Patrick J Flynn. *Image segmentation using clustering*. IEEE Press, Piscataway, NJ, 1996.

[89] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[90] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering : a review. *ACM computing surveys (CSUR)*, 31(3) :264–323, 1999.

[91] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3) :645–678, 2005.

[92] Glenn W Milligan and Martha C Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2) :159–179, 1985.

[93] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Cluster validity methods : part i. *ACM Sigmod Record*, 31(2) :40–45, 2002.

[94] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Clustering validity checking methods : part ii. *ACM Sigmod Record*, 31(3) :19–27, 2002.

[95] Anil K Jain and Richard C Dubes. Algorithms for clustering data. 1988.

[96] Richard Dubes and Anil K Jain. Clustering techniques : the user's dilemma. *Pattern Recognition*, 8(4) :247–260, 1976.

[97] Anima Majumder, Laxmidhar Behera, and Venkatesh K Subramanian. Emotion recognition from geometric facial features using self-organizing map. *Pattern Recognition*, 47(3) :1282–1293, 2014.

[98] Godfrey N Lance and William Thomas Williams. A general theory of classificatory sorting strategies : Ii. clustering systems. *The computer journal*, 10 (3) :271–277, 1967.

[99] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining : concepts and techniques*. Elsevier, 2011.

[100] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301) :236–244, 1963.

[101] Noam Slonim and Naftali Tishby. Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 208–215. ACM, 2000.

[102] Athman Bouguettaya, Qi Yu, Xumin Liu, Xiangmin Zhou, and Andy Song. Efficient agglomerative hierarchical clustering. *Expert Systems with Applications*, 42(5) :2785–2797, 2015.

[103] Hugo Steinhaus. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, 1(804) :801, 1956.

[104] G Ball and Isodata Hall Dj. A novel method of data analysis and pattern classification. isodata, a novel method of data analysis and pattern classification. tch. report 5ri, project 5533, 1965.

[105] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[106] Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(Dec) :2651–2667, 2006.

[107] Grigorios F Tzortzis and Aristidis C Likas. The global kernel k-means algorithm for clustering in feature space. *IEEE Transactions on Neural Networks*, 20(7) :1181–1194, 2009.

[108] Greg Hamerly and Charles Elkan. Learning the k in k-means. In *Advances in neural information processing systems*, pages 281–288, 2004.

[109] M Emre Celebi, Hassan A Kingravi, and Patricio A Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert systems with applications*, 40(1) :200–210, 2013.

[110] Grigorios Tzortzis and Aristidis Likas. The global kernel k-means clustering algorithm. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 1977–1984. IEEE, 2008.

[111] Teuvo Kohonen. Learning vector quantization. In *Self-Organizing Maps*, pages 175–189. Springer, 1995.

[112] Teuvo Kohonen. Neurocomputing : Foundations of research. *chapter Self-organized formation of topologically correct feature maps*, pages 509–521, 1988.

[113] T Kohonen. Self-organizing maps.-springer series in information sciences, v. 30, springer. 2001.

[114] Teuvo Kohonen, Erkki Oja, Olli Simula, Ari Visa, and Jari Kangas. Engineering applications of the self-organizing map. *Proceedings of the IEEE*, 84(10) : 1358–1384, 1996.

[115] Chantal Hajjar and Hani Hamdan. Cartes auto-organisatrices pour la classification des données de type intervalle en se basant sur la distance city-block. *Revue des Nouvelles Technologies de l'Information*, pages 119–131, 2015.

[116] Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases : The algorithm gdbscan and its applications. *Data mining and knowledge discovery*, 2(2) :169–194, 1998.

[117] Daoying Ma and Aidong Zhang. An adaptive density-based clustering algorithm for spatial database with noise. In *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*, pages 467–470. IEEE, 2004.

[118] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 1(3) :231–240, 2011.

[119] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[120] Sanghamitra Bandyopadhyay. An efficient technique for superfamily classification of amino acid sequences : feature extraction, fuzzy clustering and prototype selection. *Fuzzy Sets and Systems*, 152(1) :5–16, 2005.

[121] Jong cheol Jeong, Xiaotong Lin, and Xue-Wen Chen. On position-specific scoring matrix for protein function prediction. *IEEE/ACM transactions on computational biology and bioinformatics*, 8(2) :308–315, 2011.

[122] Swati Vipsita, Bithin Kanti Shee, and Santanu Kumar Rath. An efficient technique for protein classification using feature extraction by artificial neural networks. In *2010 Annual IEEE India Conference (INDICON)*, pages 1–5. IEEE, 2010.

[123] Stephanie Leavitt and Ernesto Freire. Direct measurement of protein binding energetics by isothermal titration calorimetry. *Current opinion in structural biology*, 11(5) :560–566, 2001.

[124] Wei Wu, Anuj Srivastava, Jose Laborde, and Jinfeng Zhang. An efficient multiple protein structure comparison method and its application to structure clustering and outlier detection. In *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*, pages 69–73. IEEE, 2013.

[125] Anuj Srivastava, Eric Klassen, Shantanu H Joshi, and Ian H Jermyn. Shape analysis of elastic curves in euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7) :1415–1428, 2011.

[126] Christine A Orengo and William R Taylor. Ssap : sequential structure alignment program for protein structure comparison. *Methods in enzymology*, 266 : 617–635, 1996.

[127] Domenico Fraccalvieri, Alessandro Pandini, Fabio Stella, and Laura Bonati. Conformational and functional analysis of molecular dynamics trajectories by self-organising maps. *BMC bioinformatics*, 12(1) :158, 2011.

[128] Giovanni Bottegoni, Walter Rocchia, Maurizio Recanatini, and Andrea Cavalli. Aclap, autonomous hierarchical agglomerative cluster analysis based protocol to partition conformational datasets. *Bioinformatics*, 22(14) :e58–e65, 2006.

[129] Lilia V Nedialkova, Miguel A Amat, Ioannis G Kevrekidis, and Gerhard Hummer. Diffusion maps, clustering and fuzzy markov modeling in peptide folding transitions. *The Journal of chemical physics*, 141(11) :09B611_1, 2014.

[130] Frédéric Chazal, Leonidas J Guibas, Steve Y Oudot, and Primoz Skraba. Persistence-based clustering in riemannian manifolds. *Journal of the ACM (JACM)*, 60(6) :41, 2013.

[131] Deborah Goldman, Sorin Istrail, and Christos H Papadimitriou. Algorithmic aspects of protein structure similarity. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 512–521. IEEE, 1999.

[132] Adam Godzik. The structural alignment between two proteins : is there a unique answer ? *Protein science*, 5(7) :1325–1338, 1996.

[133] Hitomi Hasegawa and Liisa Holm. Advances and pitfalls of protein structural alignment. *Current opinion in structural biology*, 19(3) :341–348, 2009.

[134] Wolfgang Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A : Crystal Physics, Diffraction, Theoretical and General Crystallography*, 34(5) :827–828, 1978.

[135] Mauro Rustici and Arthur M Lesk. Three-dimensional searching for recurrent structural motifs in data bases of protein structures. *Journal of Computational Biology*, 1(2) :121–132, 1994.

[136] Simon K Kearsley. On the orthogonal transformation used for structural comparisons. *Acta Crystallographica Section A : Foundations of Crystallography*, 45(2) :208–210, 1989.

[137] Andrew D McLachlan. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallographica Section A : Crystal Physics, Diffraction, Theoretical and General Crystallography*, 28(6) :656–657, 1972.

[138] Jianbo Lu, Guoliang Xu, Shihua Zhang, and Benzhuo Lu. An effective sequence-alignment-free superpositioning of pairwise or multiple structures with missing data. *Algorithms for Molecular Biology*, 11(1) :18, 2016.

[139] R Daras, Dimitrios Zarpalas, Dimitrios Tzovaras, and Michael G Strintzis. 3d shape-based techniques for protein classification. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 2, pages II–1130. IEEE, 2005.

[140] Charles E Metz. Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier, 1978.

[141] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7) :1145–1159, 1997.

[142] Steven M Holland. Principal components analysis (pca). *Department of Geology, University of Georgia, Athens, GA*, pages 30602–2501, 2008.

[143] Malika Charrad, Nadia Ghazzali, Veronique Boiteau, Azam Niknafs, and Maintainer Malika Charrad. Package 'nbclust'. *Journal of Statistical Software*, 61 :1–36, 2014.

[144] Derya Birant and Alp Kut. St-dbscan : An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1) :208–221, 2007.

[145] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer, 2001.

[146] B Dubuisson. Decision with reject options. In *5. European Signal Processing Conference.*, volume 3, pages 1715–1718, 1990.

[147] Flore Harlee. *Détection de ruptures multiples dans des séries temporelles multivariées : application à l'inférence de réseaux de dépendance.* PhD thesis, Grenoble Alpes, 2016.

[148] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1) :79–86, 1951.

Azzam ALWAN

Doctorat : Optimisation et Sûreté des Systèmes

Année 2018

Vieillissement du système cardio-vasculaire – Étude de l'activité des peptides d'élastine.

Ce travail vise à proposer une méthodologie statistique pour étudier les peptides produits par la dégradation des protéines d'élastine artérielle. Les approches biologiques indiquent que certains de ces peptides peuvent être considérés comme des signaux moléculaires et peuvent influencer l'évolution des pathologies vasculaires. De plus, des expériences montrent que les propriétés biologiques des peptides sont liées à leurs structures 3D. Dans ce contexte, l'objectif de nos travaux consiste à analyser les structures 3D de ces peptides pour identifier les conformations liées à leurs activités biologiques et pronostiquer l'activité de nouveaux peptides. Il s'agit donc d'identifier les conformations "clés" pour l'activité de ces peptides à l'aide d'une base de données de simulations dynamiques du comportement moléculaire des peptides de l'élastine. Parmi les peptides simulés, certains sont connus pour avoir un effet biologique alors que d'autres ne le sont pas. Dans un premier temps, il est extrêmement important d'identifier les principales conformations de chaque peptide à partir des simulations moléculaires. Un processus combinant plusieurs méthodes statistiques a été proposé dans ce but et a démontré son efficacité sur la base de données existante. Dans un second temps, un détecteur d'activité peptidique a été proposé. Il est capable de prévoir l'activité de nouveaux peptides non étiquetés. Le détecteur proposé est simple et peut être appliqué à de grandes bases de données.

Mots clés : apprentissage automatique – détection des anomalies (informatique) – reconnaissance des formes (informatique) – peptides – dynamique moléculaire

Aging of the Cardiovascular System – Study of the Elastin Peptides Activity.

This work aims to propose a statistical methodology to study the degradation products of arterial elastin. The proposed approach consists in analyzing simulation data of molecular dynamics of peptides resulting from the degradation of elastin proteins. Biological approaches indicate that some of these peptides can be considered as molecular signals and can influence the evolution of vascular pathologies. Moreover, experiments show that the biological properties of peptides are linked to their 3D structures. In this context, the objective of our work consists in analyzing the 3D structures of these peptides to identify the structures (conformations) related to their biological activities and next predict the activity of new peptides. It is therefore necessary to identify the "key" conformations for the activity of these peptides using a database of dynamic simulations of their molecular behaviour. Among the simulated peptides, some are known to have a biological effect while others are not. First, it is extremely important to identify the main conformations of each peptide from molecular simulations. A process combining several statistical methods is proposed for this purpose and demonstrates its effectiveness on the basis of existing data. Second, a peptide activity detector is proposed. It is able to predict the activity of new unlabelled peptides. The proposed detector is simple and can be applied to large databases.

Keywords: machine learning – anomaly detection (computer security) – pattern recognition systems – peptides – molecular dynamics