L'ANALYSE DES POPULATIONS DE CONFORMATIONS PEPTIDIQUES HYPERSURFACES CONFORMATIONNELLES DE MOLÉCULES COMPLEXES

par

Catherine Jeandenans

Thèse présentée au département de chimie en vue
de l'obtention du grade de docteur ès sciences (Ph. D)

FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, Janvier 1996.

SOMMAIRE

Dans le cadre de la mécanique moléculaire, nous somme capable de générer des échantillons représentatifs de toutes les conformations accessibles aux molécules complexes et flexibles que sont les peptides. La diversité de conformations possibles conduit à des échantillons qui, pour être représentatifs, doivent comporter plusieurs centaines voire des milliers de conformations.

Nous proposons d'utiliser les méthodes d'analyses de données pour l'étude de ces échantillons. Ces méthodes sont très largement appliquées aujourd'hui dans de très nombreux domaines et en particulier en chimie ou la sophistication croissante des appareils de mesure et de calcul conduit à traiter des quantités massives de données. L'abondance des données devient un obstacle au traitement direct et masque les relations existant entre ces données. L'utilisation des méthodes d'analyse des données permet d'ordonnancer l'ensemble des données en dégageant les structures qui les caractérisent.

Ces méthodes ont été appliquées sur des échantillons de conformations de trois peptides de taille croissante, le dernier présentant une diversité conformationnelle maximum. Les échantillons ont été générés par le programme PEPSEA précédemment développé au laboratoire. Nous présentons les résultats obtenus par les méthodes d'analyse de données adaptées au traitement de nos échantillons. Les performances de ces méthodes sont attestées par comparaison des résultats obtenus avec les études expérimentales disponibles.

Les résultats obtenus montrent que les méthodes d'analyse de données sont particulièrement utiles dans le traitement de gros échantillons de conformation nécessaires pour traduire la flexibilité conformationnelle extrême des peptides. Ces résultats sont particulièrement

intéressants quand il s'agit d'expliquer les relations structure-activité pour les peptides biologiquement actifs ainsi que nous le montrons avec le dernier peptide étudié.

REMERCIEMENTS

Je tiens à exprimer mes remerciements aux personnes et organismes qui ont contribué à la réalisation de cette thèse en particulier:

Mon professeur, Monsieur André G. Michel qui m'a initié à la modélisation moléculaire. Le département de chimie pour le soutien professionnel et matériel. Tous les membres du laboratoire qui m'ont apporté amitié, aide et encouragements tout au long de ces années en particulier Marc Drouin, Gérald Villeneuve et Gaston Boulay.

Je remercie tout particulièrement mes parents, Françoise et Pierre Jeandenans, mes frères, Frédéric et Luc et tous les amis qui m'ont apporté un soutient tant matériel que moral depuis le début de cette aventure. Je citerais particulièrement Nathalie Bredin, Nathalie Jourdan et Ludovic Andrivon, Franco Cau, Patricia Dolez, Lucie Plamondon et Martin Piotte, ainsi que Pascale Laville et Florence Roussey.

Que tout ceux que je n'ai pas nommé par manque de place ici se voient chaleureusement remercié ici, quelque soit la manière dont ils m'ont encouragée.

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ANNEXES

INTRODUCTION

Les peptides et les protéines sont impliqués dans de nombreux processus biologiques ce qui suppose une grande variété d'environnement auquel la molécule doit s'adapter. Il a été montré expérimentalement que ces molécules possèdent une structure qui peut subir des fluctuations conformationnelles parfois importantes (1 , 2). Les méthodes d'études expérimentales de ces molécules comme la diffraction des rayons-X (3), la RMN (Résonnance Magnétique Nucléaire) (4) et le dichroïsme circulaire permettent de déterminer la structure de ces molécules dans les cas où on peut les synthétiser et obtenir éventuellement des cristaux. Néanmoins, ces techniques sont difficiles à utiliser en raison de la complexité des molécules à étudier: les nombreux degrés de liberté d'un peptide rendent sa cristallisation très difficile, certaines molécules biologiques sont difficiles à isoler, à synthétiser et à obtenir en quantité suffisantes. De plus, dans certains cas, le phénomène ne peut simplement pas être observé par ces méthodes: traversée d'une membrane par une protéine ou un ion par exemple (5). Toutes ces difficultés ont conduit au développement de méthodes d'optimisation et de simulation aptes à reproduire la structure et le comportement de ces molécules. Ces méthodes vont de celles basées sur la résolution de l'équation de Schrödinger accessibles pour une petite molécule jusqu'aux méthodes basées sur les lois de la mécanique classique et qui permettent de reproduire par un ajustement empirique des paramètres du champ de force, les structures observées expérimentalement. Ces dernières, réunies sous le nom de "mécanique moléculaire", beaucoup moins exigeantes en calculs complexes permettent la simulation de molécules de grande taille.

Les premières utilisation de la mécanique moléculaire ont privilégié la découverte du minimum global d'une molécule, censé reproduire toutes les caractéristiques observées expérimentalement. Il est maintenant admis et vérifié que pour traduire le comportement réel d'une molécule, il est nécessaire de considérer un ensemble de conformations de celle-ci (6 ,

7). Les propriétés observées expérimentalement sont une moyenne sur un certain nombre de conformations métastables (8) car le passage entre les différents états conformationnels est trop rapide pour que l'on puisse isoler et étudier séparément ces états.

D'autre part, la ou les conformations actives de la molécule ne sont pas forcément celles possédant la plus basse énergie potentielle telle que fournie par la simulation. Nous touchons ici un problème commun à toutes les méthodes de simulation: l'énergie conformationnelle calculée à partir d'un champ de force en mécanique moléculaire n'est pas un critère quantitatif de stabilité. En thermodynamique statistique, le concept de stabilité équivaut à une probabilité élevée d'existence pour un état conformationnel, en accord avec la loi de distribution de Boltzman (9). Il faut donc évaluer la probabilité d'un état conformationnel ce qui n'est possible que si l'on connaît tous les états conformationnels d'une molécule. Or les méthodes de simulation qui prennent une conformation moléculaire et la minimisent éventuellement, conduisent à déterminer l'énergie potentielle d'une molécule isolée ce qui ne permet évidement pas d'évaluer la probabilité d'existence de cette conformation et par suite, empêche l'accès à l'entropie et finalement à l'énergie libre qui représente le vrai critère de stabilité pour une molécule. Ce problème de déterminer l'énergie libre est essentiel si on veut élucider de nombreux processus chimiques et biologiques comme, par exemple, la liaison d'une molécule active avec son récepteur: si nous préparons plusieurs agonistes de cette molécule, lequel produira le complexe ligand-récepteur le plus stable thermodynamiquement (10 , 11)? Ce problème s'inscrit plus fondamentalement dans l'étude des processus qui gouvernent le repliement des peptides et protéines puisque qu'il a été proposé que ce repliement soit contrôlé soit thermodynamiquement (12) soit cinétiquement (13).

On se dirige ainsi de plus en plus vers une approche globale dans la simulation des conformations moléculaires et de nombreux auteurs soulignent l'importance d'un échantillonnage complet des états conformationnels d'une molécule ceci à la fois dans le but de rencontrer les observations expérimentales (14) et de prédire la stabilité des molécules par

le calcul de l'énergie libre (8, 9, 10, 11, 15 , 16 , 17). Pour ce dernier objectif, les différentes méthodes de simulations sont plus ou moins aptes à fournir l'échantillon nécessaire à l'estimation de l'entropie. En effet, ainsi que le souligne Berendsen (15) les méthodes Monte-Carlo et de dynamique moléculaire produisent un ensemble de conformations représentatif mais fini dans l'espace des phases ce qui ne permet pas de calculer la fonction de partition totale donc l'entropie. Le même auteur propose une méthode d'échantillonnage à partir de simulations par dynamique moléculaire à température élevée suivie d'un refroidissement ce qui permet une exploration plus large de la surface d'énergie de la molécule (8). Les méthodes de mécanique moléculaire où une conformation est générée puis minimisée ne sont pourvues d'aucun système permettant de franchir les barrières énergétiques et se heurtent au problème dit des minima multiples: la molécule peut rester bloquée dans un minimum local (11). Il faut donc trouver un moyen d'échantillonner tout l'espace conformationnel. Différentes techniques ont été proposées dont le "grid scan" qui consiste à faire varier systématiquement selon un incrément déterminé les degrés de liberté de la molécule et calculer l'énergie de chaque molécule correspondante. Les méthodes d'exploration de la surface d'énergie ont été comparées par Saunders (6) qui a conclut à la supériorité des méthodes stochastiques et par Kollman (18) qui note lui aussi l'importance de l'échantillonnage dans la reproduction des propriétés physico-chimiques des molécules.. Les recherches menées au laboratoire ont conduit à la même conclusion (19). Scheraga souligne également l'importance d'une composante aléatoire dans l'échantillonnage qui traduit le coté entropique dans la recherche de minima (2).

Les différentes méthodes développées à partir des simulations pour calculer l'énergie libre souffrent toutes d'un manque de généralité soit théorique, soit par les calculs très lourds auxquels elles conduisent. Les méthodes développées par Go et Scheraga (20) d'une part et Hagler (21) d'autre part peuvent calculer l'entropie conformationnelle de molécules subissant de petites fluctuations harmoniques et sont inaptes à comparer des états conformationnels très différents. C'est le cas également de la méthode proposée par Karplus qui traite les cas

extrêmes de petites ou de très grandes fluctuations conformationnelles (16). Des approches rigoureuses comme la technique des perturbations (22) ou le calcul de l'énergie libre absolue (23) sont en pratique limitées à l'étude de petits changements chimiques ou conformationnels à cause de l'ampleur des calculs nécessaires. Récemment, une méthode appelée méthode LS ("local state") proposée par Meirovitch (17) a montré son efficacité dans ce type de calcul car elle n'est pas limitée à un certain type de fluctuations conformationnelles, néanmoins, elle requiert des approximations qui conduisent à négliger certaines corrélations entre les variables.

La méthode développée au laboratoire, et qui a prouvé son efficacité, destinée à explorer l'hypersurface conformationnelle des molécules peptidiques très flexibles, privilégie la génération aléatoire d'un grand nombre de conformations dont l'énergie est ensuite calculée dans un champ de force à géométrie rigide (les variables sont les angles dièdres) qui sont ensuite minimisées par un algorithme du gradient conjugué. Il faut noter que dans notre méthode, nous ne tenons pas compte de l'environnement dans le calcul de l'énergie. Il est d'autant plus important de considérer l'ensemble des minima possibles même si ceux-ci sont relativement élevés énergétiquement. Il est en effet possible que ces minima soient stabilisés dans un environnement particulier. En travaillant dans le vide, nous obtenons évidement que les molécules les plus stables (i.e. les plus basses en énergie) sont repliées soit celles où les interactions électrostatiques intramoléculaires sont minimisées et les liens hydrogène intramoléculaires sont favorisés. Or dans un solvant polaire par exemple, ces interactions intramoléculaires seront remplacées par des interactions molécules-solvant plus favorables ce qui conduira à favoriser les interactions étendues. Cette remarque justifie le fait que nous analysions la population totale qui représente l'ensemble des conformation possibles sans rejeter les plus énergétiques. Nous rappelons néanmoins que chaque conformations obtenue est le résultat d'une minimisation, la population est donc un ensemble de minima métastables ce qui nous assure d'éliminer tout "monstre chimique" de notre population.

Nous avons également souligné l'importance d'obtenir un échantillon représentatif de la

surface d'énergie conformationnelle dans l'optique d'un calcul d'énergie libre puisque nous devons calculer le poids statistique de chaque minimum grâce à la formule de Boltzman (24). Néanmoins, cette approche entraîne l'apparition d'un nouveau problème: lorsque la surface d'énergie potentielle est complexe (molécule flexible et/ou de grande taille) ce qui est le cas des peptides nous générons un grand échantillon de conformations et le problème de l'analyse de cet échantillon devient crucial. Nous désirons identifier des familles de conformations possédant les même caractéristiques structurales et éventuellement élucider les relations et interconversions éventuellement possibles entre les familles de conformères.

Les techniques statistiques d'analyses des données nous sont apparues parfaitement adaptées pour résoudre ce problème (25). En effet, ces techniques permettent de classer un échantillon de grande taille c-à-d de retrouver une organisation à l'intérieur de l'échantillon si cette organisation y existe intrinsèquement. De plus, il est possible de construire des variables non corrélées à partir des variables initiales. L'intérêt de ces techniques est d'analyser l'échantillon de conformations selon des critères exclusivement mathématiques à partir de variables de départ conformationnelles à l'exclusion de l'énergie. En effet, pour une hypersurface d'énergie complexe, nous observons que des molécules d'énergie identique peuvent être conformationnellement très différentes et vice versa. Aucune restriction ou a priori d'ordre chimique n'intervient dans la classification. Le résultat ne peut donc être "orienté" par l'utilisateur de la méthode qui rechercherait par exemple la présence de structures "connues" comme l'hélice α ou le tournant β. Grâce à ces techniques, nous pensons pouvoir décrire de manière qualitative et objective la structure tridimensionnelle des peptides et déterminer quantitativement la stabilité des différentes structures par le calcul de l'énergie libre.

Après un exposé général des techniques d'analyses de données, nous étudierons comment ces techniques peuvent s'appliquer à l'étude des conformations peptidiques sur une molécule test dont les caractéristiques conformationnelles sont bien connues. Nous appliquerons ensuite le processus choisi à l'étude d'un peptide de taille moyenne (quatre résidus d'acides aminées) et

finalement à l'exploration des caractéristiques conformationnelles d'un peptide de grande taille et flexibilité de manière à évaluer la généralité de la méthode d'analyse proposée et les difficultés éventuelles.

CHAPITRE 1

MÉTHODE

1.1 Introduction: l'analyse des données

L'analyse des données est un ensemble de méthodes visant à synthétiser l'information contenue dans des tableaux de données en mettant en relief les relations existant entre les individus d'un même échantillon, entre les paramètres qui caractérisent ces individus, entre les individus et les paramètres qui les caractérisent (26). Ces méthodes, nécessaires à l'exploitation des tableaux de données de grandes tailles, ont pris un essor important avec les progrès du calcul électronique qui permet le recueil et la mémorisation de vastes tableaux de données.

Les méthodes d'analyse des données ne font pas partie des méthodes statistiques classiques. En effet, ces dernières sont basées sur un modèle probabiliste et donc un ensemble d'hypothèses et de conditions doivent être posées et respectées lors d'une analyse statistique. Au contraire, l'analyse des données se propose d'extraire directement le modèle des données étudiées. En effet, dans cette analyse, l'échantillon n'a aucune condition particulière à remplir. Cette approche permet de réduire au maximum la subjectivité reliée à l'utilisateur "le modèle doit suivre les données et non l'inverse" dit J.P. Benzécri (27) un des instigateur de ces méthodes.

Nous utiliserons deux méthodes classiques appartenant aux deux grandes familles en analyse de données: les méthodes factorielles (28) et les méthodes de regroupement ou classification (29 , 30). La distinction se fait sur le plan mathématique et objectif: les méthodes factorielles

procèdent à l'aide de calculs d'ajustement utilisant l'algèbre linéaire pour produire une analyse essentiellement globale et descriptive de l'échantillon; les méthodes de regroupement mettent en jeu une formulation et des méthodes algorithmiques dont le but est de constituer des classes d'individus semblables. Dans la pratique, les deux types de méthodes sont utilisées conjointement. En effet, les méthodes factorielles analysent les corrélations entre les variables alors que par définition, dans les méthodes de regroupement, les paramètres caractérisant les individus se doivent d'être non corrélés entre eux. Ainsi que le fait Lebart (31), on peut comparer ces méthodes d'analyse des données à des instruments d'observation du multidimensionnel tel un appareil de radiographie qui fournit des images à partir d'une réalité inobservable.

De nombreuses méthodes sont disponibles aussi bien dans les méthodes factorielles que dans les méthodes de classification. Nous avons choisi dans les méthodes factorielle l'analyse en composantes principales. En effet, elle est adaptée au traitement de tableau rectangulaire c'est à dire où il y a beaucoup de variables caractérisant un petit nombre d'individus ou l'inverse (ce qui est notre cas). De plus, les variables peuvent être hétérogènes en moyenne et en dispersion mais de même nature ce qui est également notre cas puisque toutes les variables sont des distances interatomiques mais diversement distribuées en fonction de la proximité des atomes concernés dans la mesure. La nature de nos données nous a donc conduit à choisir cette méthode factorielle. Parmi les autres méthodes disponibles, on peut citer l'analyse des rangs ou les valeurs sont des échelles de classement par exemple ou l'analyse des correspondances où les valeurs sont des fréquences de réalisation d'une certaine propriété. L'analyse factorielle en facteurs commun et l'analyse discriminante ne sont pas utilisables non plus dans le cas qui nous occupe puisqu'elles supposent l'existence d'un modèle a priori pour la première ou affecte les individus à des classes connues a priori dans le cas de la seconde.

Les méthodes de classification sont également diverses non sur le principe mais la manière dont on calcule les proximité entre les individus (appelé la métrique) et dont on les agrège

ensuite. Les méthodes sont diversement biaisées selon la métrique utilisée comme nous le verrons en détail par la suite. Par conséquent, il faut utiliser conjointement plusieurs méthodes de classification. Nous avons donc choisit les méthodes de classification en fonction de leur caractéristiques: sont-elles biaisées et dans quel sens, sont-elles sensibles aux "outliers"... Le choix s'est porté sur un ensemble de 4 méthodes: l'une non hiérarchique correspondant à la procédure FASTCLUS dans le logiciel SAS, et trois méthodes de classification hiérarchiques: WARD, AVERAGE et CENTROID. Les méthodes hiérarchiques WARD et AVERAGE possède chacune un biais différents mais sont performantes dans la classification de grand échantillon, et la méthode CENTROID bien que moins performante possède l'avantage d'être peu sensible aux "outliers".

1.2 Une méthode factorielle: l'analyse en composantes principales

L'objectif de l'analyse en composantes principales (désignée par la suite "ACP") est de présenter sous forme graphique le maximum de l'information contenue dans un tableau de données. On peut, à l'aide de l'ACP, répondre à deux types de questions: comment se structurent les variables (quelles sont celles qui sont associées, celles qui ne le sont pas, celles qui vont dans le même sens, celles qui s'opposent...) et comment se répartissent les individus (quels sont ceux qui se ressemblent, qui sont dissemblables). Le tableau de données est constitué en lignes par des individus caractérisés par des variables quantitatives (ou pouvant être considérées comme telles). Ici les individus sont les conformations et les variables les angles dièdres caractérisant chaque conformation. Le problème d'approximation numérique à résoudre est le suivant: soit un ensemble de n individus caractérisés chacun par p paramètres, est-il possible de reconstituer les np valeurs du tableau de données par un nombre de valeurs inférieur? Ceci revient à chercher comment caractériser les n individus de l'échantillon par un nombre de variables inférieurs à p tout en perdant un minimum d'informations sur ces individus.

Si on reformule ceci de manière mathématique:
l'ensemble des individus est représenté dans un espace vectoriel de dimension p puisque chaque individu est caractérisé par p variables. On cherche d'abord une base de cet espace vectoriel. Par définition, la base de cet espace vectoriel sera constituée de p vecteurs, (u_1, u_2, ... , u_p), linéairement indépendants et l'ensemble des individus sera reconstitué en faisant des combinaisons linéaires de ces p vecteurs. Soit un individu $X=(x_1, x_2, ... , x_p)$, il s'écrira :

$$X = \sum_{i=1}^{p} x_i u_i \qquad [1]$$

Le but de l'ACP est de trouver une base de dimension inférieure (p-k) de manière à reconstituer les np valeurs de l'échantillon :

$$X = x_1 u_1 + x_2 u_2 + ... + x_{p-k} u_{p-k} + E \qquad [2]$$

E étant un vecteur n,p résiduel dont les termes sont petits. Ainsi, si nous voulons reconstituer l'individu X de manière approchée et satisfaisante, nous conservons les p-k premiers termes en négligeant le terme E.

Quel est l'intérêt de générer ces nouvelles variables?
Les n individus de notre échantillon peuvent théoriquement être représentés dans l'espace des variables initiales, soit dans p dimensions, ce qui rend impossible la visualisation d'une telle représentation. Le but de l'ACP sera de représenter au mieux les n individus dans un espace de dimension 1, 2 ou 3. On peut faire l'analogie avec la photographie: on y passe d'un espace à trois dimensions à un espace à deux dimensions (la photo). Pour cela, la base de représentation choisie sera construite de manière à ce que la variance des individus projetés dans cette base soit maximale. On expliquera cette caractéristique par un exemple simple:
On veut étudier un ensemble de n d'individus caractérisés par deux paramètres, p_1 et p_2. La base de représentation initiale est donc de dimension 2. On cherche la base de dimension

inférieure (soit de dimension 1) dans laquelle on observera au mieux les individus.
En faisant les combinaisons linéaires de p_1 et p_2, on construira 2 vecteurs indépendants, cp_1 ($cp_1 = a_1p_1 + b_1p_2$) et cp_2 ($cp_2 = a_2p_1 + b_2p_2$) (formellement appelés composantes principales) qui constituent une base de l'espace des n individus. Pour choisir une base de dimension inférieure, on déterminera quel est le vecteur (cp_1 ou cp_2) sur lequel les individus ont une variance maximale:

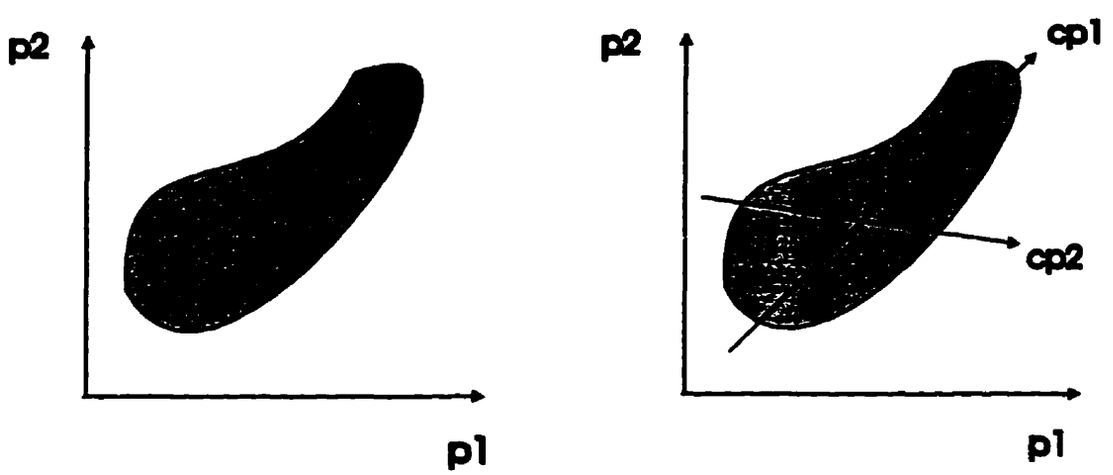


Figure 1. Projection d'un ensemble de n individus dans l'espace de leurs variables puis sur les nouvelles variables composantes principales.

On constate sur la figure 1. que la composante principales cp_1 est celle sur laquelle les individus projetés ont la plus grande variance. Si on doit décrire l'ensemble des n individus à partir d'une unique variable, on choisira donc cp_1.

L'intérêt de cette réduction des variables significatives est bien entendu faible lorsqu'on ne doit examiner qu'un petit nombre de variables. En revanche lorsque les dimensions du tableau de données sont un obstacle à sa lecture, l'ACP devient un outil essentiel qui permet de distinguer les variables essentielles à la description de l'échantillon, de celles qui sont accessoires. Le processus mathématique de l'ACP est décrit dans l'annexe A.

1.3 Les méthodes de regroupement

Le but de l'analyse de regroupement ou de classification (appelée aussi "clustering") est de classer les variables ou les individus en ensembles distincts. Il est possible de placer ainsi les individus dans des groupes suggérés par les données et non définis à priori. Le principe de la classification conduit à mettre dans une même classe les individus qui présentent une similitude. Ce degré de similitude entre individus est calculé sur les variables qui définissent chaque individu. Plusieurs types de regroupements sont possibles: en partition, hiérarchique ascendant ou hiérarchique descendant.

1.3.1 Quelques définitions

Il y a plusieurs manières de représenter les objets à classer:

. On peut utiliser une matrice de "similarité" où à la fois les lignes et les colonnes correspondent aux objets à classer (une matrice des corrélation est un exemple de matrice de similarité).

. On peut utiliser une matrice des coordonnées où les lignes sont les observations et les colonnes sont les variables. Dans ce cas, les observations ou les variables ou encore les deux peuvent être classées.

Comme souligné plus haut, il y a également plusieurs types de regroupement ou classes:

. disjointes: chaque objet est placé dans une et une seule classe.

. hiérarchiques: une classe peut être entièrement contenue dans une autre mais aucune superposition de classe n'est possible.

. superposées: le regroupement permet qu'un nombre d'objets (pré-défini ou non) appartiennent simultanément à deux classes.

. floues: classes définies par la probabilité qu'a chaque objet d'appartenir à chaque classe (peuvent être disjointes, hiérarchiques ou superposées).

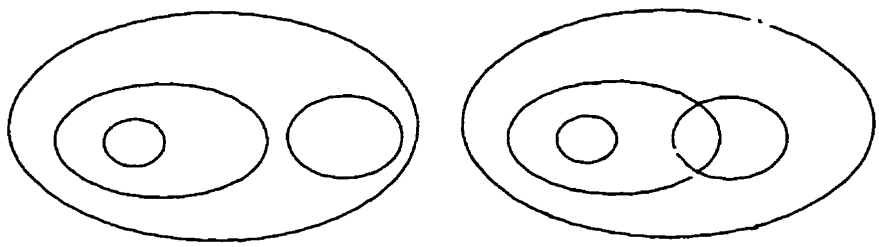


Figure 2. Exemple de classes hiérarchiquement emboitées (à gauche) et superposées (à droite).

1.3.2 Classement hiérarchique

La manière de procéder au classement hiérarchique est la suivante:

. On choisit au départ une "distance" dont on va munir l'ensemble des individus à classer et qui va mesurer adéquatement la similarité entre les objets.

. On suppose qu'il existe des règles de calcul entre les groupements disjoints d'objets.

Voici en guise d'illustration de la méthode, un exemple tiré de Lebart (31) :

Si x, y et z sont trois objets à classer et si les objets x et y sont regroupés en un seul élément h, on définira la distance de h à z par la plus petite distance des divers éléments de h, à z.

$$d(h,z) = Min\ [\ d(x,z),\ d(y,z)\ ] \qquad [3]$$

On pourrait également utiliser le critère de la distance moyenne:

$$d(h,z) = [\ d(x,z) + d(y,z)\ ] / 2 \qquad [4]$$

A l'étape suivante, si x et y désignent des sous-ensembles disjoints de l'ensemble des objets, possédant respectivement n_x et n_y objets, h sera un sous-ensemble formé de $(n_x + n_y)$ éléments et on définira:

$$d(h,z) = [n_x\ d(x,z) + n_y\ d(y,z)\ ] / (n_x + n_y) \qquad [5]$$

Ainsi, à chaque étape du regroupement, de nouveaux objets et ensembles sont formés. Un lien existe entre l'étape précédente et la suivante c'est pourquoi ce type de classification est appelée hiérarchique. On peut illustrer ce type de classification par un dendrogramme:

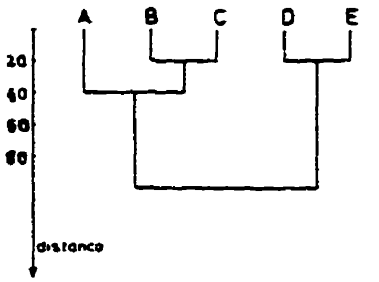


Figure 3. Dendrogramme illustrant le
processus de classification hiérarchique

Si on décide qu'il y a 2 classes dans notre échantillon initial de 5 objets (A, B, C, D, E) nous obtiendrons la classe "ABC" et la classe "DE". Si on opte pour trois classes, elle seront constituées par "A", "BC" et "DE". Ainsi l'ensemble initial des objets est muni d'une règle de calcul qui permet de mesurer la distance entre les objets eux mêmes et également entre les ensembles d'objets.

Deux variantes existent dans ce type de classification: la classification hiérarchique ascendante et la classification hiérarchique descendante. Dans le premier cas, on suppose que l'échantillon possède autant de classes n que d'objets et on agrège les objets pas à pas. Après n-1 itérations de ce processus, tous les objets appartiennent à la même classe. Dans le deuxième cas, on effectue le processus inverse: on suppose que tous les objets sont dans la même classe et on sépare cette classe en sous-ensembles jusqu'à obtenir n classes possédant chacune un seul objet.

1.3.3 Classification non-hiérarchique

Le but est ici de rechercher la partition optimale en q classes de l'ensemble des individus. Partition optimale signifie que l'on doit déterminer le nombre de classes qui existent naturellement à l'intérieur de notre échantillon lorsque ce dernier est muni d'une règle de calcul de distance entre individus. Ainsi, tout l'échantillon pourra être décrit à partir des caractéristiques des quelques ensembles trouvés. Le but est encore ici "d' y voir clair" à l'intérieur d'un échantillon ou il est impossible de décrire chaque objet. Pour cela, le nombre optimal d'ensembles doit être le plus réduit possible.

Le critère d'agrégation est de rendre minimale la variance à l'intérieur d'une classe en maximisant les distances entre les classes. La première étape est de générer q "germes" indépendant selon diverses méthodes, germes constitués par des individus pris à l'intérieur de la partition comme "centre de masse" provisoires des q classes. La distance euclidienne des autres individus à ces germes est ensuite calculée et les individus sont ainsi affectés aux classes. On détermine alors q nouveaux centres de classes qui seront les centres de gravité des q premières classes et on réitère le processus, ce qui va conduire à une nouvelle, et meilleure, partition de q classes. La décision d'arrêter la classification se fait de diverses façons. On peut arrêter le classement lorsque deux itérations successives conduisent à la même partition, lorsque un nombre d'itérations pré-déterminé a été atteint ou lorsque la variance intra-classe cesse de décroître de manière significative. Les caractéristiques mathématiques des différentes méthodes de regroupement sont décrites dans l'annexe B.

1.3.4 Les indices statistiques

Nous avons vu que les méthodes de classification procèdent d'une méthode très simple: on choisit une distance qui mesure l'écart entre les individus de l'échantillon et on agrège les plus proches. On itère le processus et on obtient ainsi une partition des objets en classes. Le problème essentiel de la classification est de décider le niveau de partition de l'échantillon.

Pour cela, il faut connaître le nombre de classes existant à priori dans l'échantillon. Or un des principe de l'analyse de données est d'extraire l'information des données, sans poser aucune hypothèse de départ. Pour résoudre ce problème, de nombreux tests (plus de trente) qui sont appelés indices statistiques ont été proposés. Les tests d'hypothèse habituels en statistique sont inutilisables (tels que le test de Fisher ou de Student) pour tester les différences entre les classes. En effet, les méthodes de classification tendent à maximiser la séparation entre les classes et les hypothèses sur la distribution des données qui assurent la validité des tests sont violées (32). Une hypothèse sur la distribution de l'échantillon (appelée hypothèse nulle) est posée (les données sont échantillonnées aléatoirement à partir d'une distribution normale multivariée ou à partir d'une distribution uniforme par exemple). Un critère est ensuite calculé dans chaque classe (par exemple le déterminant ou la trace de la matrice contenant la somme des carrés des variables caractérisant chaque individu) pour chaque niveau de classification. Ce critère est ensuite comparé avec celui calculé pour chaque classe d'une distribution théorique correspondant à l'hypothèse nulle. Les niveaux de classification pour lesquels les deux critères calculés sont en accord traduisent une partition correcte de l'échantillon. Néanmoins, tous les tests proposés, qui diffèrent par le choix de l'hypothèse nulle et du critère calculé, émettent des hypothèses sur la distribution ou la forme des classes à l'intérieur de l'échantillon ce qui porte à traiter les résultats avec prudence. C'est pourquoi, plutôt que de se fier à un test en particulier, on utilise et observe le résultat conjoint de plusieurs tests de manière à obtenir autant que possible un consensus entre les résultats. Néanmoins, il est possible qu'à cause des hypothèses choisies dans le calcul d'un test particulier celui-ci soit impropre pour traiter un échantillon particulier. Il est par suite nécessaire d'utiliser conjointement plusieurs tests sur un même échantillon de manière a obtenir un consensus sans que l'on puisse attendre un consensus parfait entre les tests quelque soit le problème de classification étudié. Dans les cas où aucune information sur le nombre de classe ne peut être retirée de l'étude des indices statistique il faudra examiner directement les résultats pour plusieurs niveaux de classification. Parmi tous les tests statistiques proposés pour prédire le nombre optimal de classes d'un échantillon, trois sont couramment utilisés car bien

documentés et validés: le CCC (Cubic Clustering Criterion) (33), le pseudo F (34) et le pseudo t^2 (35). La valeur de CCC, Pseudo F et Pseudo t^2 est tracée en fonction du nombre de classes possibles. On doit observer un pic pour CCC et Pseudo F combiné avec une faible valeur pour Pseudo t^2, cette dernière étant immédiatement suivie d'une valeur élevée. Un exemple de tracé des indices statistiques est donné à la figure suivante.

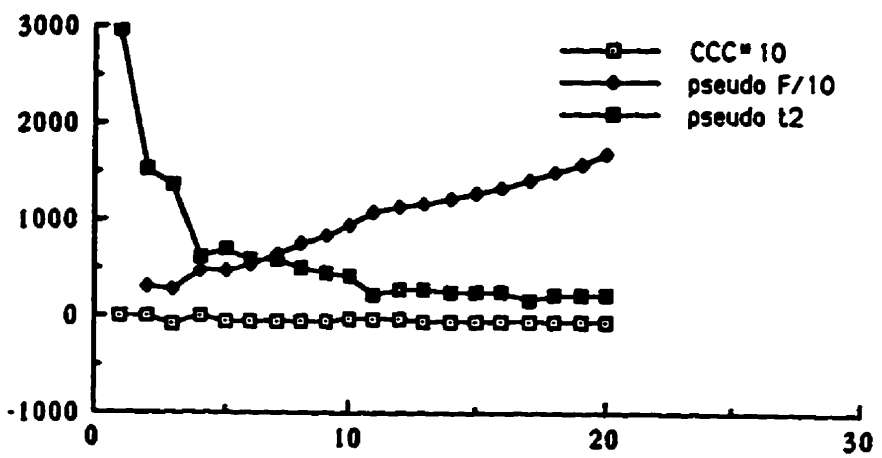


Figure 4. Les trois indices statistiques tracés par rapport au nombre de classes éventuelles pour une classification hiérarchique par la méthode AVERAGE .

Nous voyons ici que CCC et Pseudo F nous indiquent un nombre de classes optimal de 4 et Pseudo t^2 par un minimum immédiatement suivi d'un maximum confirme cette valeur. Le résultat ici est clair car les trois indices statistiques sont en accord pour une même valeur. En revanche, si nous observons la figure suivante, l'interprétation des indices statistiques n'est pas aussi évidente.

Dans ce cas, CCC indiquerait un nombre de classes optimal de 4, alors que Pseudo F et Pseudo t^2 suggèrent plutôt 5 classes. Pourtant, ces deux graphiques ont été tracés à partir du même échantillon pour un classement hiérarchique. La différence est la méthode choisie à

l'intérieur du classement hiérarchique c'est à dire la manière dont est calculée la distance entre les individus. En plus de comparer les résultats des trois indices statistiques, nous constatons qu'il est important de classer notre échantillon plusieurs fois avec différentes méthodes de manière à obtenir un consensus.

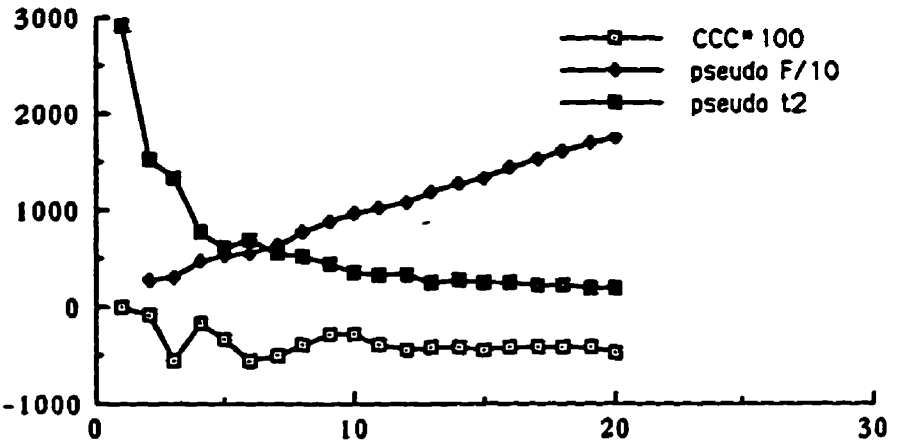


Figure 5. Les trois indices statistiques tracés par rapport au nombre de classes éventuelles pour une classification hiérarchique par la méthode WARD.

1.3.5 Influence des critères choisis sur la qualité de la classification

A partir des indices statistiques, nous avons constaté que les différentes méthodes de classification n'avaient pas les même performances. En effet, toutes les méthodes ne proposent pas le même nombre optimal de classes. Nous abordons le deuxième problème des méthodes de classification: l'existence de biais. Selon la manière dont est calculée la distance dont est muni l'échantillon, la classification produira des classes possédant diverses caractéristiques. On peut avoir production de classes ayant toutes la même variance, de classes ayant toutes le même diamètre, de classes de taille égales, ou inégales (36). Or on ne connaît pas, a priori, les caractéristiques de dimension, d'élongation, de variance, de dispersion, de

symétrie...etc des véritables classes existantes à l'intérieur de l'échantillon. Les différentes méthodes auront donc des performances inégales et qui dépendront étroitement du type d'échantillon à classer. Si les classes existantes dans l'échantillon sont approximativement de même variance, une méthode biaisée dans le sens de trouver des classes ayant cette caractéristique particulière sera très performante. Nous devons donc utiliser conjointement plusieurs méthodes de classification et comparer les résultats obtenus.

Il faut souligner que les méthodes de classification seront d'autant plus difficiles à mettre en oeuvre que les classes seront allongées et superposées. Si les classes sont suffisamment bien séparées, toutes les méthodes seront à peu près aussi performantes (37). De plus, les indices statistiques, qui utilisent nous le rappelons certaines hypothèses sur la distribution de l'échantillon et la distribution interne des classes, sont valables dans le cas de classes compactes ou légèrement allongées, de préférence avec une distribution interne de type normal multivarié dans les classes ce qui complique encore leur utilisation lors de l'étude d'un échantillon complexe.

Des comparaisons entre les méthodes ont été faites de manière à déterminer lesquelles sont les meilleures. Ces études ont été menées en générant des échantillons artificiels qui contiennent des classes connues au moyen de générateurs de nombres pseudo aléatoires. L'échantillon est analysé par les différentes méthodes de classification et le degré de performance de chaque méthode est mesuré (38) en comparant l'habileté des diverses méthodes à retrouver ces classes. Néanmoins, la plupart de ces études génèrent des classes compactes (souvent normales multivariées) approximativement de taille et dispersion égale. Il n'est pas surprenant que ces études concluent favorablement à la supériorité des méthodes que l'on sait performantes lorsque les classes possèdent ces caractéristiques de distribution (39). Certaines méthodes sont relativement moins biaisées que les autres et même si les études de simulation ne les notent pas comme les plus performantes, il est judicieux de les utiliser surtout lorsque l'échantillon est complexe (beaucoup d'individus à classer et/ou de

nombreuses variables caractéristiques).

1.4 Logiciels et temps de calcul

1.4.1 Echantillon de conformations peptidiques

La génération de l'échantillon de conformation des peptides est faite avec le logiciel PEPSEA développé et validé au laboratoire. Pour une description du logiciel et des procédures de calcul ainsi que la validation du résultats de calcul, se référer à (40). Brièvement, la procédure consiste en une génération aléatoire de conformations dont l'énergie est calculée dans un champ de force à géométrie rigide ECEPP/2 (Empirical Conformation Energy Program for Peptides). La minimisation de l'énergie est ensuite effectuée par un algorithme du gradient conjugué. Ceci conduit à un échantillonnage des minima locaux (et du minimum global éventuellement) de la surface conformationnelle de la molécule.

1.4.2 Analyse des données

Le logiciel utilisé pour le traitement statistique est le logiciel SAS (Statistical Analysis System) (41) version 6.08 implanté sur l'ordinateur ES9000 d'IBM au centre de calcul de l'Université de Sherbrooke. Le module SAS/STAT contient les programmes d'analyses de données. L'ACP se fait par la procédure PRINCOMP, la classification automatique par les différentes méthodes regroupées dans la procédure CLUSTER et la classification non-hiérarchique par la procédure FASTCLUS.

1.4.3 Temps de calcul pour l'ACP

Si n est le nombre d'individus, v le nombre de variables caractérisant chaque individu et c le nombre de composantes principales, le temps de calcul se répartit comme suit:

- calcul de la matrice des corrélations: temps proportionnel à nv^2
- calcul des valeurs propres: temps de calcul proportionnel à v^3
- calcul des vecteurs propres: temps de calcul proportionnel à cv^2

1.4.4 Temps de calcul pour la classification

1.4.4.1 Classification non-hiérarchique

Si n est le nombre d'observations, v le nombre de variables et c le nombre de classes, le temps de calcul est proportionnel à $nvc + vc^2$

1.4.4.2 Classification hiérarchique

Le temps de calcul dépend de la méthode employée c'est à dire du type de distance à calculer entre les individus. De manière générale, le temps pour la classification hiérarchique dépend du nombre d'individus à classer. Si celui-ci est égal à n, le temps de calcul sera de n^2 à n^3 selon la méthode. Nous constatons que la partie la plus exigeante et donc limitative dans l'analyse de données est la classification hiérarchique qui augmente très rapidement avec le nombre d'individus.

1.4.5 Traitement des valeurs numérique

L'utilisation des ordinateurs conduit à commettre deux types d'erreurs dans le traitement des valeurs numériques: les erreurs d'arrondi, indépendantes de l'utilisateur et reliées au type d'ordinateurs utilisé et les erreurs de troncature dépendant de la construction du logiciel traitant les données (42). Les erreurs d'arrondi concernent le stockage des valeurs réelles. En effet, ces dernières devraient être stockées avec une précision infinie ce qui n'est pas le cas. Selon le type d'ordinateurs, la précision du stockage sera plus ou moins grande avec plus ou

moins de décimales stockées pour une valeur réelle. Seuls les entiers seront stockés avec une précision parfaite. Le problème arrive lorsqu'une suite d'opérations arithmétique est effectuée sur ces valeurs. Dans un processus itératif, le nombre de décimales nécessaires pour obtenir un résultat exact double à chaque opération et dépasse rapidement la capacité de stockage de l'ordinateur. Plus il y a d'itérations, plus le résultat final sera éloigné de la valeur réelle. De plus, un nombre stocké avec une précision importante au départ verra une dérive plus rapide des résultats des opérations arithmétiques subséquente. Si l'erreur commise par l'ordinateur est alternée c'est à dire que ce dernier alternativement surestime et sous-estime le nombre réel, l'erreur finale sur le nombre réel sera de l'ordre de la racine carré du nombre d'opérations arithmétiques N effectuées, multiplié par l'erreur de stockage commise par l'ordinateur sur une valeur. Néanmoins, si l'erreur d'arrondi commise par l'ordinateur est biaisée au départ c'est à dire que ce dernier à tendance à systématiquement sous-estimer ou surestimer la valeur l'erreur finale sera N fois l'erreur de stockage. Pour contourner ce problème au maximum, il est possible à l'utilisateur d'arrondir les valeurs numériques. De cette manière, l'erreur initiale est connue et les valeurs dérivent moins vite des valeurs réelles puisqu'elles peuvent être stockée exactement plus longtemps dans la suite des opérations arithmétiques. De même, on peut transformer les valeurs initiales de manière à les stocker sous forme d'entiers.

Les erreurs de troncature sont reliées à la formulation algorithmique du programme de calcul numérique. L'évaluation numérique d'une intégrale par exemple est faite à partir d'un nombre fini de valeurs. Le nombre de valeurs devra fournir une solution suffisamment proche de ce qui serait obtenu si le nombre de valeurs était effectivement infini. Dans la formulation des opérations arithmétiques, les propriétés de commutativité, distributivité et associativité ne sont pas forcément respectés et le placement des parenthèse est à contrôler car diversement interprété par les ordinateurs et les compilateurs. Cette dernière observation peut d'ailleurs être utilisée pour détecter les erreurs dues à l'algorithmique d'un programme (43). Enfin, il est possible que les deux types d'erreurs (qui sont indépendantes et doivent s'additionner) se multiplient en fait si l'algorithme est instable. Cela signifie qu'une erreur d'arrondi introduite

au départ va être démesurément amplifiée dans la suite des calculs (comparaison ou multiplication entre très petits nombres par exemple) jusqu'à masquer la valeur finale. Il faut noter ici que les méthodes les plus précises mathématiquement (minimisation faisant intervenir la dérivée seconde d'une fonction par exemple) sont aussi les plus instables et sensibles aux erreurs d'arrondi. De plus, il faut remarquer que même si une méthode donne le même résultat sur plusieurs ordinateurs, cela ne signifie pas que le résultat obtenu est correct. En effet, on ne peut pas calculer l'évolution dans le temps d'un système dynamique sensible aux conditions initiales. Le calcul numérique présente donc une sensibilité aux conditions initiales et une sensibilité à la précision des calculs.

En ce qui concerne la génération et la minimisation des structures par PEPSEA les erreurs commises dues à l'algorithmique ont été minimisées par l'utilisation de banques de fonctions préprogrammées. Ceci prévient l'utilisation d'un algorithme instable. L'algorithme de minimisation ne fait pas intervenir la dérivée seconde des fonctions calculant l'énergie des molécules ce qui limite la complexité des calculs. Le critère de convergence est ajusté pour ne pas augmenter démesurément le nombre d'itérations nécessaire à la minimisation. Le programme a été compilé sur différents ordinateurs avec différentes précision de stockage des réels et différents paramètres et les résultats obtenus montrent que le programme est stable. La précision de l'énergie est consistante à deux chiffres après la virgule et donc utilisée ainsi pour la suite des calculs. Enfin, les résultats obtenus ont été validés par comparaison des structures obtenues et de leur énergie avec les résultats obtenus par divers programmes de modélisation ainsi que les résultats expérimentaux. Ceci a été vérifié sur des composés de tailles diverses conduisant à des minimisation (donc des itérations) plus ou moins longues.

Le logiciel SAS utilisé pour l'analyse de données est un logiciel commercial et optimisé dans sa conception et sa stabilité pour divers systèmes d'exploitation. Plusieurs études ont été effectuées pour tester la stabilité et la sensibilité du logiciel aux erreurs d'arrondi (44 , 45). Selon les fonctions utilisées dans SAS, les valeurs numériques sont diversement traitées selon

l'importance que les erreurs d'arrondi ou de troncature peuvent avoir sur les résultats (transformation en valeurs entières, arrondi des valeurs au départ...). Lors de nos calculs, nous nous sommes conformé aux recommandation quant au traitement des valeurs numériques que nous trouvons dans le chapitre 5 de (46). Les valeurs sont donc données avec une précision de plus ou moins un sur le dernier chiffre décimal mentionné.

1.5 Application à l'étude des populations peptidiques

1.5.1 Introduction

Les méthodes d'analyses de données ont été appliquées avec succès dans de nombreux domaines en chimie autant pour analyser des résultats expérimentaux que pour rationaliser et modéliser les phénomènes (47). Quelques utilisations en ont été faites dans l'analyse de population de conformations de molécules à commencer par le groupe de Benzécri (48) qui a appliqué l'analyse de données à l'étude de l'angiotensine et du polypeptide $(Ala)_6$-Pro-Ala, deux structures peptidiques très voisines. Cette étude constitue un travail exploratoire des possibilités de l'analyse des données dans le domaine de l'analyse de conformations chimiques. Les résultats obtenus, permettent de rationaliser l'étude d'un échantillon de conformations. Néanmoins, aucune mention n'est faite des conformations expérimentales pour ces molécules et, par conséquent, aucune validation des résultats obtenus n'est présentés. Le groupe de Maigret a ensuite étudié des échantillons de conformations en solution pour l'enképhaline (49) et l'angiotensine (50) à partir d'un échantillon généré par la méthode de Monte-Carlo. Ces deux études conduisent à determiner un ensemble réduit de conformations métastables en accord avec les observations expérimentales effectuées par RMN. Néanmoins, ces études sont limitées à un cas particulier de conformation puisqu'elles tentent de reproduire les observations en solution aqueuse acide ce qui ne donne pas un aperçu de l'ensemble des conformations accessibles pour ces molécules indépendamment du milieu. Quelques années plus tard, le substrat angiotensinogène (51) et un fragment CCK8 de l'hormone

cholécystokinine (52) a été étudié par le même groupe de chercheurs, toujours après génération d'un échantillon par méthode Monte-Carlo. Ces études sont caractérisées par le fait que l'échantillon généré est de petite taille (~ 3000 conformations), ou/et que la simulation se fait en solution, on accède donc à une partie seulement de l'espace conformationnel de la molécule. Ces caractéristiques facilitent le traitement par l'analyse de données et les résultats sur la description des structures des molécules étudiées obtenus par ces chercheurs recoupent correctement les travaux expérimentaux. Ces études attestent de l'applicabilité des méthodes d'analyse des données à l'étude d'échantillon de conformations. En outre, la performance de telles méthodes à retrouver les observations expérimentales est démontrée par l'accord observé dans ces études avec les observations expérimentales. Néanmoins, il est dommage qu'aucune de ces études ne s'attache à décrire l'ensemble des conformations accessibles à un composé. En effet, nous savons que lorsqu'il s'agit d'étudier des molécules potentiellement actives biologiquement, le problème de déterminer la (ou les) conformation active in vivo est majeur, et à la mesure de l'intérêt que cette information représente pour l'explication ou la prédiction de l'activité de ce composé et surtout le design de nouveaux composés. En effet, le milieu biologique est d'une complexité telle qu'aucune méthode expérimentale n'est capable de déterminer la conformation des molécules en condition réelle. Il s'ensuit que les conformations déterminées expérimentalement peuvent malheureusement être aussi proche de la réalité qu'elles peuvent en être éloignées, le degré de similitude étant imprévisible car propre à chaque composé étudié.

Notre propos diffère donc de celui des études préalablement mentionnées: nous désirons décrire la totalité de l'espace conformationnel de la molécule et découvrir les principes qui gouvernent le repliement des peptides. Nous devons par conséquent générer un échantillon important en taille. Le programme que nous utilisons génère une population qui représente tout l'espace conformationnel de la molécule (si la taille de l'échantillon est suffisante). Les méthodes d'analyse de données permettent-elles de trier correctement notre échantillon c'est à dire de manière à ce que les résultats obtenus aient un sens chimiquement? Nous allons

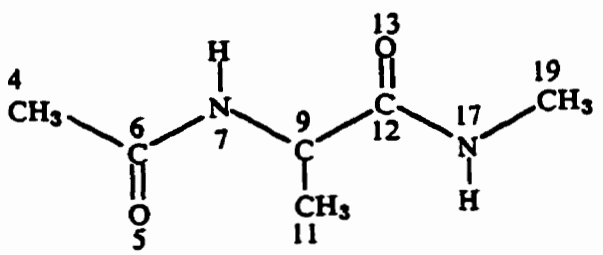
tester ces méthodes sur un peptide dont l'espace conformationnel est bien connu : l'alanine (53). Ce peptide a été étudié de manière exhaustive par diverses méthodes de simulation y compris avec le champ de force ECEPP/2 (voir paragraphe 1.4, 1.4.1) qui sert à calculer l'énergie conformationnelle dans le programme PEPSEA (54 , 55 , 56).

1.5.2 Etude du peptide Nacétyle-N' méthylamide-Alanine

A l'aide du programme PEPSEA et suivant la procédure décrite en 1.4, 1.4.1, une population de 1000 conformères a été générée en prenant tous les angles dièdres variables excepté les angles ω des liaisons peptidiques, fixés à 180°. Notre échantillon est donc constitué de 1000 individus caractérisés par 5 angles dièdres soit, pour suivre la nomenclature IUPAC, θ_1, ϕ_1, ψ_1, κ_1, θ_2. Néanmoins, il a été souligné que l'utilisation des angles dièdres comme variables introduisait des difficultés liées au caractère périodique de celles-ci (48). Nous utiliserons plutôt comme variables un ensemble de distances interatomiques. En choisissant correctement ces distances, on a l'avantage d'accéder directement aux caractéristiques conformationnelles qui décrivent traditionnellement les peptides et les protéines comme l'hélice α, le tournant γ ou le tournant β. En effet, toutes ces structures secondaires sont caractérisées par la formation de liens hydrogène entre les C=O et N-H des différents aminoacide constituant ces molécules. Nous choisirons donc comme distances caractéristiques les distances entre ces groupements. De plus, certaines mesures expérimentales apportent une information en terme de distances comme la fluorescence X qui permet de mesurer la distance entre les groupements aromatiques, ou la RMN dont la procédure noesy (57) permet d'obtenir des distances entre les différents groupements de la molécule. De plus, les études QSAR (Quantitative Structure Activity Relationships) permettent de déterminer les pharmacophores des molécules actives biologiquement soit les endroits où se lie la molécule au récepteur. On a donc accès à certaines distances caractéristiques entre les différents groupements d'une molécule. Si nous avons accès à ces informations pour des récepteur particuliers, nous pourrons directement déterminer si notre peptide a une activité biologique potentielle et ce quantitativement puisque

nous pouvons calculer le pourcentage de la population totale qui remplit les conditions de distances. Les distances utilisées comme variables initiales sont présentées dans la figure suivante.

Les statistiques élémentaires telles que la moyenne et l'écart-type ont été calculées pour ces distances. Nous devons vérifier si nos distances se trouvent dans l'intervalle μ-3σ et μ+3σ qui est le critère de Gauss traduisant l'homogénéité de l'échantillon.



D1	4-19	C_{Ac} $C_{mét}$	Distance bout à bout
D2	5-17	O_{Ac} $N_{mét}$	C_7 A
D3	7-13	N_1 O_1	C_5 A
D4	5-11	O_{Ac} $C_{\beta A}$	chaîne latérale-acétyle
D5	11-17	$C_{\beta A}$ $N_{mét}$	chaîne latérale-méthylamide

Figure 6. Distances interatomiques décrivant la molécule pour l'analyse de données.

Nous supposons ici que la distribution des distances converge vers une distribution normale pour un grand nombre de conformations dans l'échantillon. En utilisant l'intervalle de μ-3σ à μ+3σ, nous supposons que 99.7% des distances sont dans cet intervalle. L'existence d'une distance extérieure à cet intervalle signifie soit que l'hypothèse de la normalité de la distribution est incorrecte, soit que la distribution est effectivement normale mais non classique (possédant un troisième et quatrième moment important traduisant une asymétrie ou/et un aplatissement de la distribution de Gauss) soit que nous observons un individu "outlier".

Un tel outlier peut être diversement interprété. Il peut être relié à la procédure de génération des données: erreur de mesure ou dans notre cas, de calcul. Il peut également traduire une vraie caractéristique présente dans les données et insuffisamment échantillonnée par la procédure de mesure ou de calcul, ou la restriction sur le choix des individus mesurés. Un exemple illustre cette possibilité: il existe une relation parfaitement linéaire entre le nombre de carbone d'une chaîne aliphatique et la lipophilie de ces composés. Admettons que nous ayons mesuré, parmi un ensemble d'alcanes de différentes tailles, la lipophilie d'un alcool. Ce dernier va évidement apparaître comme un point déviant par rapport aux autres sans que cela traduise une mesure de lipophilie erronée. Cela signifie seulement que la mesure de lipophilie faite sur le composé outlier n'est pas interprétable par le modèle linéaire choisi. Dans notre cas, un tel outlier peut avoir également diverses significations. Une distance déviante peut être liée à une conformation chimiquement impossible. Ce type de conformation correspond à un échec de la procédure de minimisation et conduit à des conformations d'énergie très élevée, facilement détectable et éliminées par un examen de la distribution énegétique des conformations. Une telle distance peut également représenter une conformation tout à fait possible chimiquement. Dans ce cas, il se peut que l'échantillon soit trop restreint et que toute une partie des conformations soit manquante. On s'assure de cela en étudiant la convergence de la moyenne et de l'écart-type en fonction de la taille de l'échantillon. En plus des indicateurs statistiques, il faut donc regarder un graphique de la distribution des distances pour

détecter des caractéristiques particulières et ainsi aider à la détermination de la nature des outliers le cas échéant.

L'utilisation du critère de Gauss peut ainsi conduire à éliminer certains individus intéressants. De plus, on fait l'hypothèse a priori que la distribution est de type normale.

Tableau 1. Statistiques élémentaires sur les 5 distances.

Variable	N	Moyenne	Ecart-type	Minimum	Maximum
D1	999	5.92	0.62	5.27	7.07
D2	999	3.80	0.81	2.50	4.98
D3	999	3.21	0.33	2.66	3.51
D4	999	3.51	0.54	2.71	4.25
D5	999	3.35	0.32	2.80	3.74

1.5.2.1 Analyse en composantes principales

L'annexe B décrit en détail la procédure d'analyse en composantes principale ainsi que les équations nécessaires au traitement des données et aux différents calcul subséquents. En résumé, après avoir construit une matrice des corrélations à partir des variables initiales centrées et réduites (donne une matrice symétrique (5, 5)), on diagonalise cette matrice pour obtenir un ensemble de valeurs propres avec les vecteurs propres correspondant.

La valeur propre associée à chaque composante principale fournit le pourcentage "d'information" apporté par cette composante (le terme information n'est pas utilisé ici dans le sens mathématique qu'il possèderait en théorie de l'information par exemple).

Tableau 2. Résultats de l'analyse en composantes principales.

Matrice des corrélations

	D1	D2	D3	D4	D5
D1	1.0000	0.7720	-.8350	-.2216	-.4437
D2	0.7720	1.0000	-.3212	-.2254	-.2716
D3	-.8350	-.3212	1.0000	0.1418	0.4435
D4	-.2216	-.2254	0.1418	1.0000	0.2455
D5	-.4437	-.2716	0.4435	0.2455	1.0000

Valeurs propres

	Valeurs propres	Pourcentage de variance	Pourcentages cumulés
cp1	2.69374	0.538747	0.53875
cp2	0.94248	0.188496	0.72724
cp3	0.79599	0.159199	0.88644
cp4	0.55652	0.111303	0.99775
cp5	0.01127	0.002255	1.00000

Vecteurs propres

	cp1	cp2	cp3	cp4	cp5
D1	-.584106	0.217681	0.156738	-.136246	0.753860
D2	-.453963	0.063389	0.649286	0.431162	-.427114
D3	0.495745	-.274660	0.323247	0.570134	0.499256
D4	0.234822	0.912894	-.128933	0.307991	0.000812
D5	0.389658	0.199455	0.657836	-.612883	-.003219

La valeur propre associée à chaque composante principale fournit le pourcentage "d'information" apporté par cette composante (le terme information n'est pas utilisé ici dans le sens mathématique qu'il possèderait en théorie de l'information par exemple). Nous voyons ici que cp1 porte 53.87% de l'information totale portée par les variables initiales sur l'échantillon. Nous devons décider combien nous devons garder de composantes principales pour l'analyse de regroupement subséquente. Il est de règle pour ce type d'analyse de conserver un nombre de composantes principales correspondant à peu près à 90% de l'information totale ce qui conduit ici à garder les trois premières composantes. Le pourcentage mentionné ici est purement statistique. Le choix du nombre de facteurs dépend

étroitement du but de l'étude. Si le but est une représentation graphique des données de manière à fournir à l'utilisateur une vision de la structure de ses données, le choix est évident puisqu'on ne gardera que les 3 premières composantes principales au maximum. Selon la complexité du problème (soit approximativement le nombre de variables indépendantes qui le caractérisent complètement), la représentation obtenue sera plus ou moins fidèle à la réalité. Si le but est d'utiliser le résultat de l'ACP comme modèle prédictif, la démarche est tout autre. En effet, il est possible de prédire les valeurs des variables pour un individu nouveau que l'on introduit une fois l'ACP terminée. La qualité de la prédiction dépendra du nombre de facteurs conservés. Trop peu et trop de facteurs conduisent au même résultat: une mauvaise prédiction du nouvel individu. Si il y a trop peu de facteurs, l'information contenue dans les facteurs conservés est insuffisante pour constituer un modèle du phénomène mesuré. Si le nombre de facteur est trop élevé, il s'ensuit un "surffitage" des données. Le modèle établit reproduit parfaitement la structure des données mais perd toute généralité et qualité d'extrapolation. Ces problèmes sont bien connus et il existe des techniques comme la validation croisée qui permettent de s'en affranchir (consiste à extraire un par un chaque individu de l'ensemble des données et à faire une analyse à partir des individus restant qui servira a prédire l'individu éliminé. Une mesure de la qualité prédictive est obtenue en faisant la somme des erreurs obtenues pour tous les individus éliminés au moins une fois). Dans le cas qui nous occupe l'ACP est utilisée pour décorréler les variables initiales de manière à pouvoir utiliser une méthode de regroupement par la suite. Il s'agit donc de conserver suffisamment d'information sans garder trop de nouvelles variables composantes principales. En effet, ceci nuira à l'efficacité du regroupement subséquent. Le nombre de facteurs à observer ou à conserver dépend donc de l'objectif de l'étude. La règle des "90%" est basée sur des observations empiriques et le "bon sens" de l'utilisateur doit toujours décider si elle est adaptée à son cas particulier.

Traditionnellement, et bien qu'un nouveau modèle de représentation "topologique" de la surface d'énergie ait été récemment proposé (53, 57), l'espace conformationnel des peptides

et protéines est graphiquement décrit à l'aide d'une carte de Ramachandran (40, 54). Cette carte dont les axes sont les angles ϕ et ψ permet de représenter la répartition d'un résidu d'acide aminé particulier pour les différentes conformations obtenues lors d'une recherche conformationnelle. Des zones conformationnelles repérées par un code alphabétique de A à H traduisent un type particulier de conformation pour la chaîne principale du résidu. Un enchaînement de plusieurs résidus situés dans la zone "A" donnera par exemple une structure caractéristique en hélice α. La carte de Ramachandran présentée sur la figure suivante et tirée des publications (40, 54) montre les zones conformationnelles les plus peuplées pour l'alanine soit les zones où la structure du peptide conduit à une énergie conformationnelle minimale. Dans le cas de l'alanine, ce type de carte permet d'avoir une vue d'ensemble de toutes les conformations possibles de la chaîne principale du peptide puisque ce dernier est composé d'un unique résidu d'acide aminé.

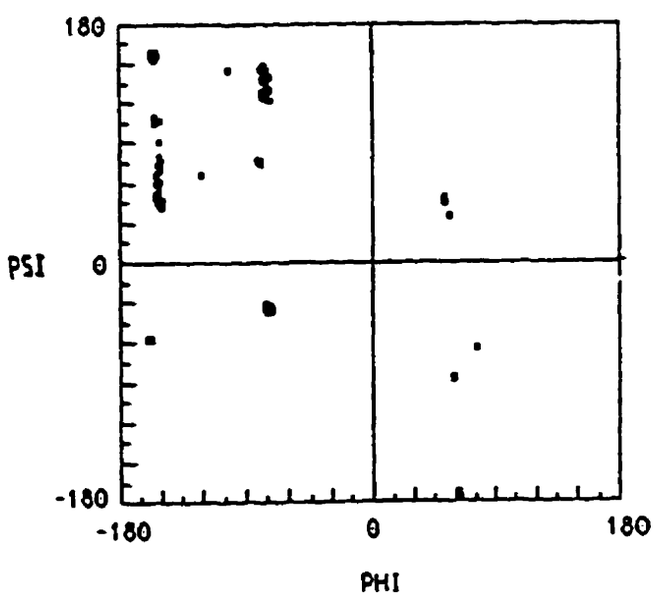


Figure 7. Carte de Ramachandran localisant les individus par rapport à leurs angles dièdres ϕ et ψ.

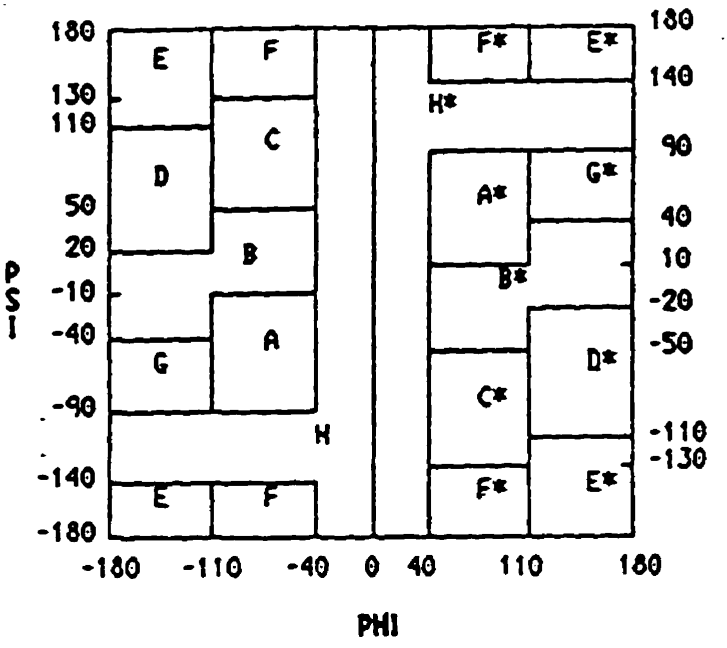


Figure 8. Carte de Ramachandran localisant les codes correspondant aux zones conformationnelles.

Chaque conformation pour l'alanine est décrite à l'aide des deux angles dièdres ϕ et Ψ de la chaîne principale du peptide. L'ensemble des conformères générés converge vers quelques zones conformationnelles soit ici les zones C (correspondant à la formation d'un pseudo-cycle a 7 membres par formation d'une liaison hydrogène intramoléculaire entre les groupements C=O et N-H de la molécule), E, A, D, F, G, A*, F*, C* (les zones notées "*" correspondent à une structure inverse de la zone non étoilée. La zone C correspond à des angles ϕ= 80° et Ψ= -80° alors que la zone C* aura les angles ϕ= -80° et Ψ= 80°). L'intérêt de décrire notre population à l'aide de l'analyse de données est de réduire le nombre de variables significatives dans un premier temps et si nous conservons trois composantes principales, nous ne remplissons pas nos objectifs. Nous allons donc conserver une seule des composantes principales et procéder à l'analyse de regroupement. Il faut souligner ici que ce n'est pas en général le critère retenu pour déterminer le nombre de composantes principales à conserver

et le but de cette étude de l'alanine à partir d'une seule composante principale est démonstratif. En effet, le contenu des composantes principales en information est beaucoup plus élevé que celui des angles ϕ et Ψ. Le but est purement didactique et tend à prouver qu'à l'aide d'une seule variable, nous pouvons décrire aussi bien sinon mieux tous les minima conformationnels qu'avec les deux variables (valeurs des angles ϕ et Ψ) nécessaire pour décrire les structures à l'aide des codes conformationnels et des cartes de Ramachandran.

Si nous examinons les valeurs propres associées à chaque vecteur propre, nous voyons comment sont bâties les composantes principales à partir des distances initiales. Dans la première composante principale, c'est la distance D1 (distance bout-à-bout) qui a le coefficient de combinaison linéaire le plus élevé suivi de D3 (pseudo-cycle à 5 membres) et D2 (pseudo-cycle à 7 membres). La deuxième composante principale est très fortement construite à partir de D4 soit la distance entre la chaîne latérale de l'alanine et le groupement acétyle initial.

1.5.2.2 Méthodes de regroupement

1.5.2.2.1 Les indices statistiques

Le choix du nombre de classes existantes à l'intérieur de l'échantillon est déterminé par l'examen des indices statistiques (voir paragraphe 1.3.4, 1.3.5, 1.5.1). L'examen des graphiques pour CCC pseudo F et pseudo t^2 présente certains problèmes. En effet, pour certaines valeurs du nombre de classes, nous ne pouvons calculer la valeur de ces indices ce qui conduit à des valeurs manquantes sur les graphiques, en particulier pour le pseudo t^2. Ceci vient du fait que pour certains nombre de classes, la procédure de classification est impuissante à trouver une partition correcte pour l'échantillon. En outre, le maximum de classes pour lesquelles nous avons des valeurs pour les indices est de 16. Nous avons par conséquent utilisé pour la prédiction les indices CCC et pseudo F qui conduisent pour les trois méthodes AVERAGE, WARD et CENTROID à une prévision de 16 classes.

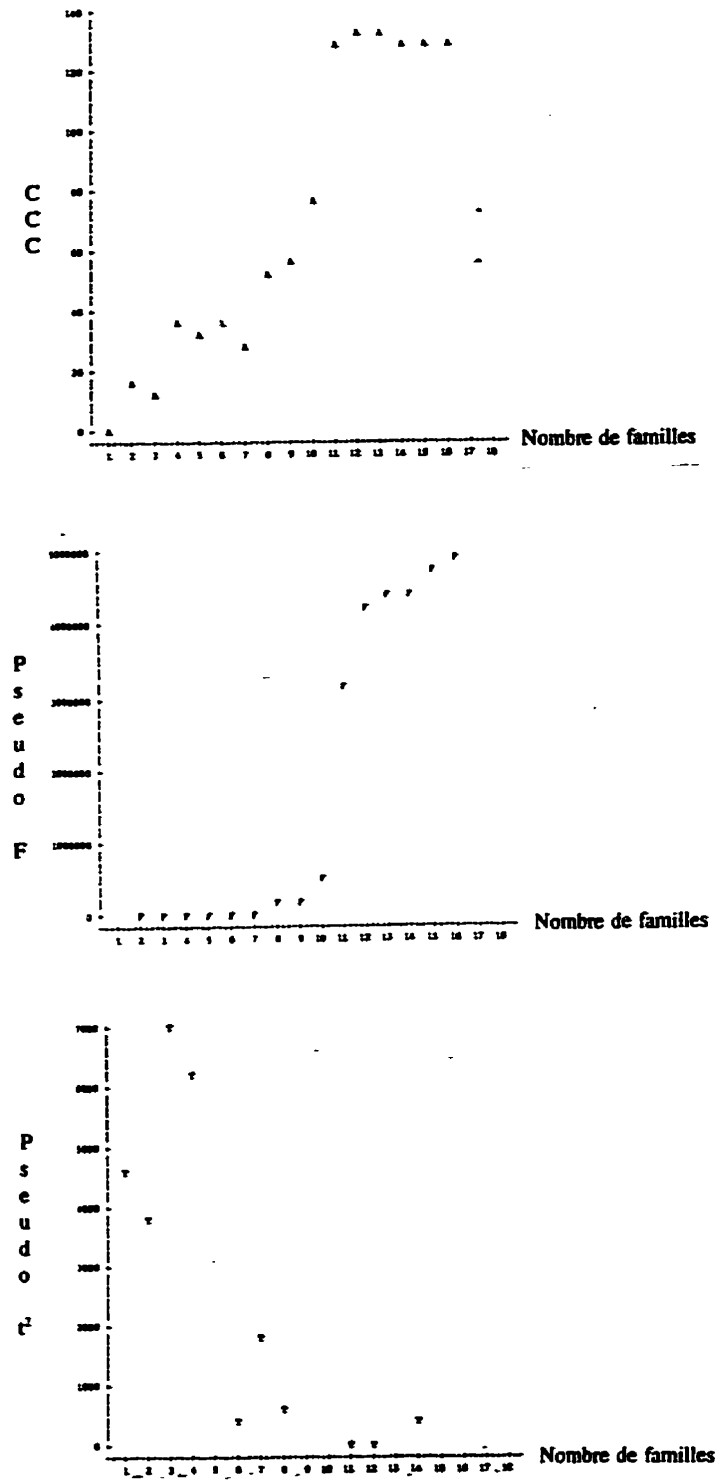


Figure 9. Graphes des trois indices statistiques en fonction du nombre de familles pour la méthode de classification AVERAGE.

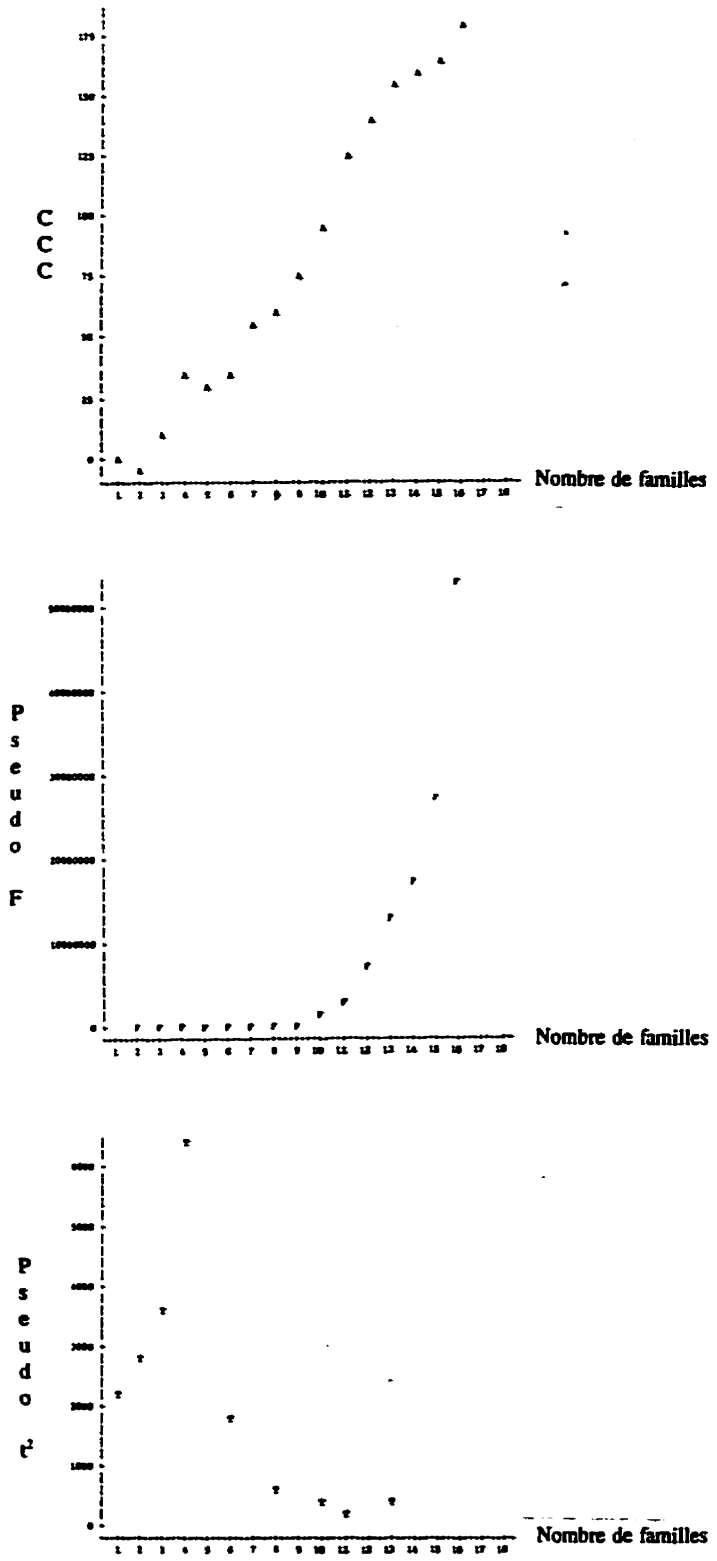


Figure 10. Graphes des trois indices statistiques en fonction du nombre de familles pour la méthode de classification de WARD.

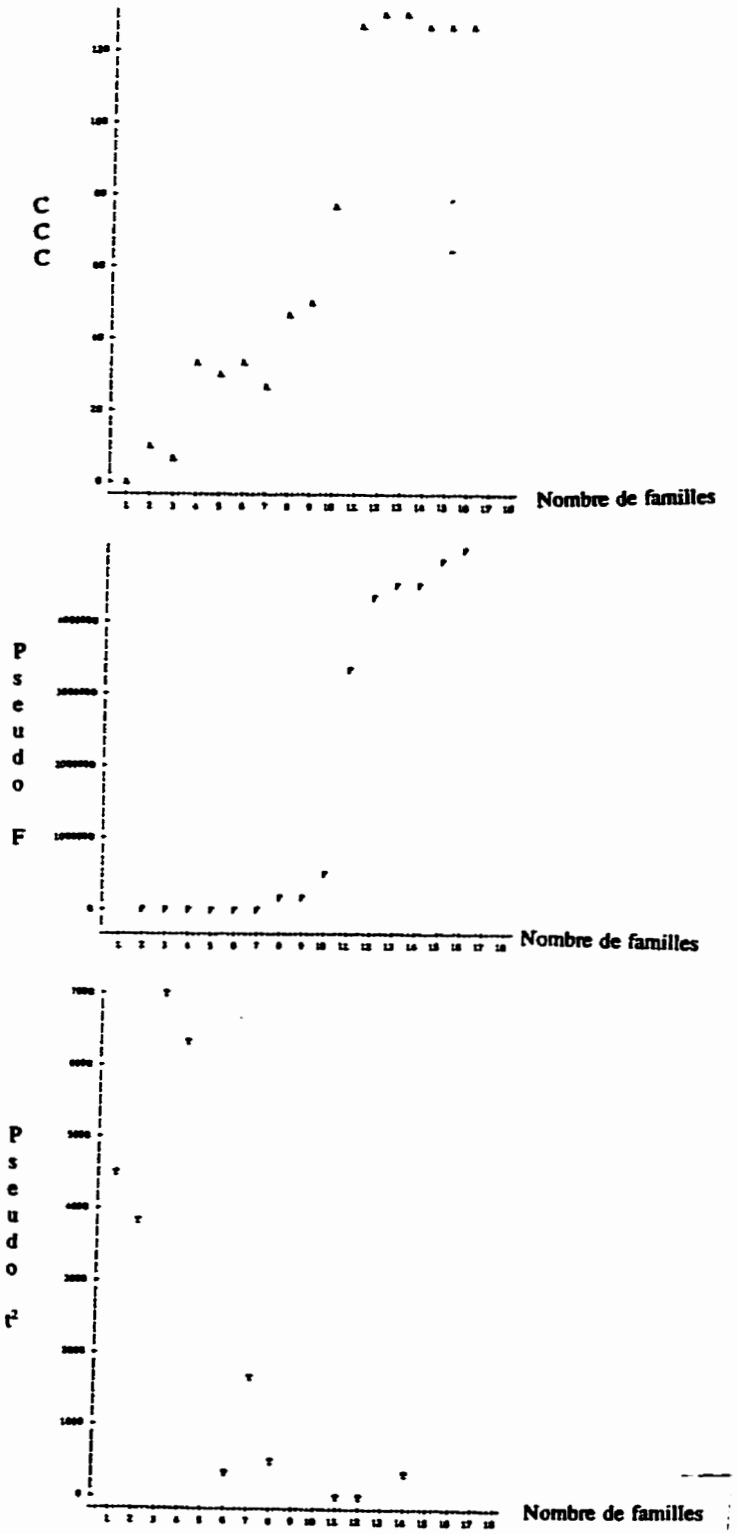


Figure 11. Graphes des trois indices statistiques en fonction du nombre de familles pour la méthode de classification du CENTROID.

1.5.2.2.2 Résultats

Nous ne pouvons présenter de projections des individus sur les variables canoniques puisque notre classification est faite à partir de la seule composante principale cp1. Nous présentons les résultats en donnant pour chaque famille (notée #fam.) trouvée par chaque méthode, le nombre d'individus dans la famille (noté #conf.), la moyenne et l'écart-type des 5 distances (notées respectivement MDn et SDn avec n de 1 à 5) et des angles initiaux ϕ et Ψ (notée MPHI, MPSI et SPHI, SPSI), les codes conformationnels des individus correspondant aux angles ϕ et Ψ (noté code) et la moyenne et l'écart-type de l'énergie conformationnelle absolue (notées MEner. et SEner respectivement). L'homogénéité des classes obtenues est quasiment parfaite. En effet, les distributions des distances et des angles dans chacune des classes permettent de voir la proximité conformationnelle des individus. Ici, les écart-types sur les distances et les angles dièdres phi et psi sont, dans la majorité des classes, égaux à zéro ce qui traduit une classe pour laquelle tous les individus sont identiques, et ce pour les quatre méthodes de classification utilisées. De plus, les "outliers" ont été détectés correctement par trois des méthodes sur quatre. Ils apparaissent sous forme de classes ne contenant qu'un seul individu, cet individu ayant effectivement des caractéristiques de structure unique et non comparables au reste des individus de l'échantillon. Il faut noter que nous avons affaire ici à des "outliers" traduisant une réalité conformationnelle et non conséquent à une erreur de calcul. Ceci est vérifié par les caractéristiques conformationnelle de ces individus qui leur sont propres (valeurs des angles ϕ et ψ). Ce sont des outliers de type chimique. La méthode de classification utilisée doit par conséquent les distinguer en les plaçant dans une classe qui ne contiendra que cet unique "outlier". Les différentes méthodes utilisées sont diversement sensibles à la présence de tel individus. La méthode WARD est la moins performante quant à la détection de ces individus particuliers. En effet, deux des outliers sont placées dans une classe unique (#12) alors que les individus de code F sont placés dans deux classes distinctes (#5 et #6). Ceci confirme l'observation selon laquelle la méthode WARD est moins performante lorsqu'il y a des "outliers" dans l'échantillon à classer. En revanche, la méthode

Tableau 3. Caractéristiques conformationnelles des 16 classes trouvées par les méthodes FASTCLUS et AVERAGE sur la population des 1000 individus de l'alanine.Energies en kcal/mol

fastclus

#fam.	#conf.	MD1	SD1	MD2	SD2	MD3	SD3	MD4	SD4	MD5	SD5	MPHI	SPHI	MPSI	SPSI	Code	MEner.	SEner.
1	114	5.53	0	2.81	0	3.27	0	4.21	0	3.72	0	-79	0	75	0	C	-5.16	0
2	127	7.07	0	4.98	0	2.70	0	3.30	0	3.17	0	-154	0	157	0	E	-4.47	0
3	148	5.33	0	3.55	0	3.51	0	4.24	0	3.23	0	-73	0	-34	0	A	-4.37	0
4	1	5.73	.	2.82	.	3.11	.	4.25	.	3.65	.	-69	.	95	.	C	-4.29	.
5	109	5.96	0	4.48	0	3.46	0	3.37	0	3.73	0	-150	0	45	0	D	-4.07	0
6	73	6.37	0	3.62	0	2.78	0	4.23	0	3.34	0	-75	0	139	0	F	-4.06	0
7	37	5.52	0	2.50	0	3.19	0	2.99	0	2.80	0	59	0	-86	0	C'	2.60	0
8	1	6.32	.	3.55	.	2.81	.	4.23	.	3.37	.	-74	.	135	.	F	-4.05	.
9	1	6.26	.	3.47	.	2.84	.	4.23	.	3.41	.	-74	.	130	.	C	-4.05	.
10	98	6.26	0	4.96	0	3.40	0	3.23	0	3.02	0	-158	0	-57	0	G	-3.44	0
11	1	6.43	.	3.83	.	2.83	.	4.12	.	3.40	.	-93	.	132	.	F	-3.12	.
12	160	5.27	0	3.23	0	3.46	0	3.07	0	3.73	0	54	0	46	0	A'	-2.81	0
13	1	5.35	.	3.34	.	3.43	.	3.06	.	3.74	.	55	.	51	.	A'	-2.48	.
14	87	6.58	0	3.98	0	2.66	0	2.92	0	2.92	0	63	0	-175	0	F'	-0.08	0
15	1	6.03	.	3.97	.	3.10	.	3.04	.	3.64	.	56	.	96	.	H'	1.16	.
16	40	5.37	0	2.66	0	3.35	0	2.71	0	2.96	0	76	0	-64	0	C'	2.10	0

average

#fam.	#conf.	MD1	SD1	MD2	SD2	MD3	SD3	MD4	SD4	MD5	SD5	MPHI	SPHI	MPSI	SPSI	Code	MEner.	SEner.
1	114	5.53	0	2.81	0	3.27	0	4.21	0	3.72	0	-79	0	75	0	C	-5.16	0
2	127	7.07	0	4.98	0	2.70	0	3.30	0	3.17	0	-154	0	157	0	E	-4.47	0
3	148	5.33	0	3.55	0	3.51	0	4.24	0	3.23	0	-73	0	-34	0	A	-4.37	0
4	109	5.96	0	4.48	0	3.46	0	3.37	0	3.73	0	-150	0	45	0	D	-4.07	0
5	73	6.37	0	3.62	0	2.78	0	4.23	0	3.34	0	-75	0	139	0	F	-4.06	0
6	98	6.26	0	4.96	0	3.40	0	3.23	0	3.02	0	-158	0	-57	0	G	-3.44	0
7	160	5.27	0	3.23	0	3.46	0	3.07	0	3.73	0	54	0	46	0	A'	-2.81	0
8	87	6.58	0	3.98	0	2.66	0	2.92	0	2.92	0	63	0	-175	0	F'	-0.08	0
9	40	5.37	0	2.66	0	3.35	0	2.71	0	2.96	0	76	0	-64	0	C'	2.10	0
10	37	5.52	0	2.50	0	3.19	0	2.99	0	2.80	0	59	0	-86	0	C'	2.60	0
11	1	5.73	.	2.82	.	3.11	.	4.25	.	3.65	.	-69	.	95	.	C	-4.29	.
12	1	5.35	.	3.34	.	3.43	.	3.06	.	3.74	.	55	.	51	.	A'	-2.48	.
13	1	6.26	.	3.47	.	2.84	.	4.23	.	3.41	.	-74	.	130	.	C	-4.05	.
14	1	6.03	.	3.97	.	3.10	.	3.04	.	3.64	.	56	.	96	.	H'	1.16	.
15	1	6.43	.	3.83	.	2.83	.	4.12	.	3.40	.	-93	.	132	.	F	-3.12	.
16	1	6.32	.	3.55	.	2.81	.	4.23	.	3.37	.	-74	.	135	.	F	-4.05	.

Tableau 4. Caractéristiques conformationnelles des 16 classes trouvées par les méthodes WARD et CENTROID sur la population des 1000 individus de l'alanine. Energies en kcal/mol

ward

#fam.	#conf.	MD1	SD1	MD2	SD2	MD3	SD3	MD4	SD4	MD5	SD5	MPHI	SPHI	MPSI	SPSI	Code	MEner.	SEner.
1	114	5.53	0	2.81	0	3.27	0	4.21	0	3.72	0	-79	0	75	0	C	-5.16	0
2	127	7.07	0	4.98	0	2.70	0	3.30	0	3.17	0	-154	0	157	0	E	-4.47	0
3	148	5.33	0	3.55	0	3.51	0	4.24	0	3.23	0	-73	0	-34	0	A	-4.37	0
4	109	5.96	0	4.48	0	3.46	0	3.37	0	3.73	0	-150	0	45	0	D	-4.07	0
5	25	6.36	0	3.61	0	2.79	0	4.23	0	3.35	0	-75	0	138	0	F	-4.06	0
6	48	6.37	0	3.63	0	2.78	0	4.23	0	3.34	0	-75	0	139	0	F	-4.06	0
7	98	6.26	0	4.96	0	3.40	0	3.23	0	3.02	0	-158	0	-57	0	G	-3.44	0
8	160	5.27	0	3.23	0	3.46	0	3.07	0	3.73	0	54	0	46	0	A*	-2.81	0
9	87	6.58	0	3.98	0	2.66	0	2.92	0	2.92	0	63	0	-175	0	F*	-0.08	0
10	40	5.37	0	2.66	0	3.35	0	2.71	0	2.96	0	76	0	-64	0	C*	2.10	0
11	37	5.52	0	2.50	0	3.19	0	2.99	0	2.80	0	59	0	-86	0	C*	2.60	0
12	2	6.15	0	3.72	0	2.97	0	3.64	1	3.53	0	-9	92	113	24	C	-1.44	4
13	1	5.73	.	2.82	.	3.11	.	4.25	.	3.65	.	-69	.	95	.	C	-4.29	.
14	1	6.43	.	3.83	.	2.83	.	4.12	.	3.40	.	-93	.	132	.	F	-3.12	.
15	1	5.35	.	3.34	.	3.43	.	3.06	.	3.74	.	55	.	51	.	A*	-2.48	.
16	1	6.32	.	3.55	.	2.81	.	4.23	.	3.37	.	-74	.	135	.	F	-4.05	.

centroid

#fam.	#conf.	MD1	SD1	MD2	SD2	MD3	SD3	MD4	SD4	MD5	SD5	MPHI	SPHI	MPSI	SPSI	Code	MEner.	SEner.
1	114	5.53	0	2.81	0	3.27	0	4.21	0	3.72	0	-79	0	75	0	C	-5.16	0
2	127	7.07	0	4.98	0	2.70	0	3.30	0	3.17	0	-154	0	157	0	E	-4.47	0
3	148	5.33	0	3.55	0	3.51	0	4.24	0	3.23	0	-73	0	-34	0	A	-4.37	0
4	109	5.96	0	4.48	0	3.46	0	3.37	0	3.73	0	-150	0	45	0	D	-4.07	0
5	73	6.37	0	3.62	0	2.78	0	4.23	0	3.34	0	-75	0	139	0	F	-4.06	0
6	98	6.26	0	4.96	0	3.40	0	3.23	0	3.02	0	-158	0	-57	0	G	-3.44	0
7	160	5.27	0	3.23	0	3.46	0	3.07	0	3.73	0	54	0	46	0	A*	-2.81	0
8	87	6.58	0	3.98	0	2.66	0	2.92	0	2.92	0	63	0	-175	0	F*	-0.08	0
9	40	5.37	0	2.66	0	3.35	0	2.71	0	2.96	0	76	0	-64	0	C*	2.10	0
10	37	5.52	0	2.50	0	3.19	0	2.99	0	2.80	0	59	0	-86	0	C*	2.60	0
11	1	5.73	.	2.82	.	3.11	.	4.25	.	3.65	.	-69	.	95	.	C	-4.29	.
12	1	5.35	.	3.34	.	3.43	.	3.06	.	3.74	.	55	.	51	.	A*	-2.48	.
13	1	6.26	.	3.47	.	2.84	.	4.23	.	3.41	.	-74	.	130	.	C	-4.05	.
14	1	6.03	.	3.97	.	3.10	.	3.04	.	3.64	.	56	.	96	.	H*	1.16	.
15	1	6.43	.	3.83	.	2.83	.	4.12	.	3.40	.	-93	.	132	.	F	-3.12	.
16	1	6.32	.	3.55	.	2.81	.	4.23	.	3.37	.	-74	.	135	.	F	-4.05	.

AVERAGE, supposée posséder la même restriction quant aux "outliers" n'est pas affectée par ces derniers dans notre cas. Si nous comparons les résultats que nous obtenons avec la littérature, nous constatons que nous retrouvons toutes les classes prédites pour le résidu Ala, soit 9 conformations différentes traduites par les codes conformationnels C, E, A, D, F, G, A*, F*, C* (codes relatifs aux valeurs prises par les angles ϕ et ψ) identifiés dans (53, 57). De plus, la classe codée C* est séparée en deux classes distinctes, traduisant deux groupes d'angles dièdres caractéristiques pour ϕ et ψ. Ce minimum supplémentaire de conformation codée C* n'a pas été localisé par les études précédentes de ce peptide. Nous pensons que cela vient du choix des angles variables lors de la minimisation. En effet, il a été montré que la restriction de certains angles pouvait empêcher l'accès à certains minima (58). Les études précédentes sur ce peptide ont restreint les angles de l'acétyle initial et du méthyle final à 180° ce qui restreint, par conséquent, le nombre de minima accessibles et gène de processus de minimisation. De plus, 6 individus atypiques ont été échantillonnés qui n'ont jamais étés décrits précédemment. Ces individus sont effectivement peu représentatif des tendances générales de conformations de l'alanine et donc peu importants pour décrire qualitativement la population. En effet, ils sont négligeable relativement à leur poids statistique. Néanmoins, ce sont de véritables minima qui ne correspondent pas à une erreur de calcul et doivent être par conséquent conservé si nous voulons décrire complètement les possibilités conformationnelles de ce peptide. En effet, il n'est pas exclu que dans certaine conditions particulières du milieu biologique, ces conformation ait un rôle à jouer.

La classe contenant les individus codés C présente une moyenne de 2.81Å pour la distance D2 traduisant la présence d'un lien hydrogène C=O:::N-H fermant une pseudo-cycle à 7 membres. Les deux classes C* présentent également la formation de ce lien hydrogène traduit par des distances D2 respectivement de 2.66Å et 2.50Å. Nous présentons la carte de Ramachandran localisant les minima obtenu lors de notre étude par rapport aux angles dièdres ϕ et ψ.

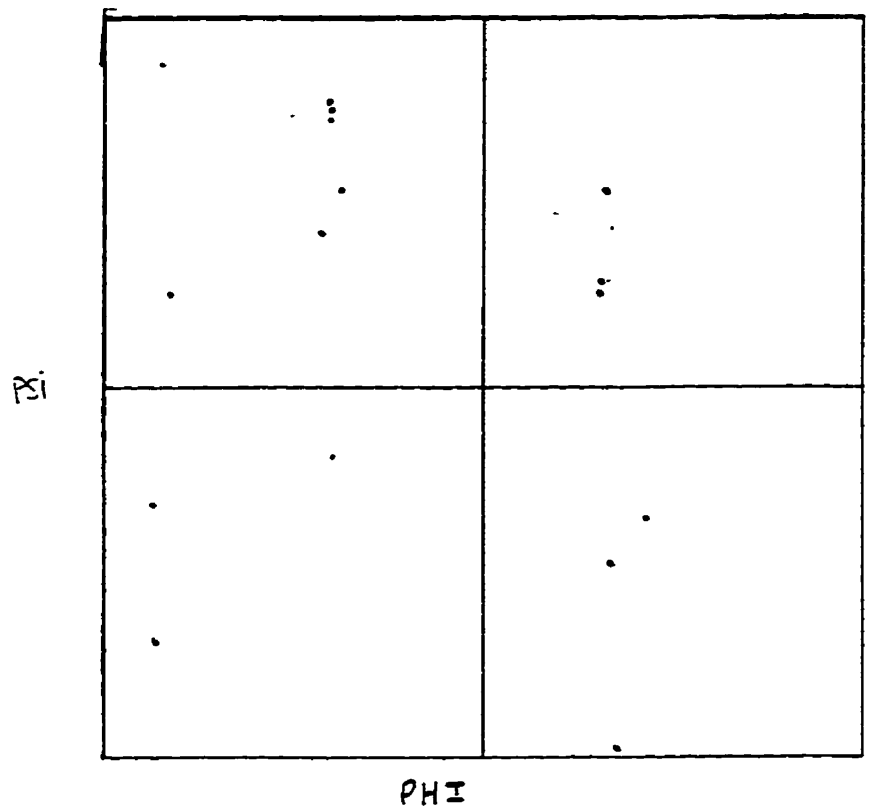


Figure 12. Carte de Ramachandran localisant les individus par rapport à leurs angles dièdres ϕ et ψ.

Il faut souligner que les 999 individus de la population sont tracés sur cette carte mais les minima à l'intérieur d'une classe sont tous exactement identiques quant à leurs angles dièdres ϕ et ψ et donc les 114 individus codés C apparaissent sous forme d'un seul point sur la carte. Cette observation est à comparer avec la carte de la page 33 de la référence (40) où les minima appartenant à un même groupe (A ou C...) n'ont pas tous exactement les mêmes angles ϕ et ψ mais sont dispersés autour d'une moyenne d'angle. Ce fait est à relier encore une fois avec le fait que la restriction de certains angles empêche l'accès à certains minima. Dans ce cas, nous constatons qu'imposer les angles de l'acétyle et du méthyle initial et final a empêché la convergence totale de certains conformères en gênant le processus de minimisation. De plus, nous retrouvons exactement les 16 classes trouvées par les méthodes de classification sous forme de 16 points distincts sur la carte de Ramachandran. Tous les

conformations de départ ont donc minimisé vers 16 types de conformations, chacune d'elle ayant un poids statistique plus ou moins grand relié au nombre de fois où elles ont été échantillonnées.

1.5.2.3 Discussion et conclusion

Nous avons démontré sur un exemple simple les possibilités de l'analyse de données appliquées à l'analyse de l'espace conformationnel peptidique. Les méthodes de classification automatique appliquées après le traitement des données initiales par l'analyse en composantes principales permettent de retrouver toutes les caractéristiques conformationnelles connues pour le peptide. De plus, l'utilisation de ces méthodes permet de décrire toutes les caractéristiques structurales à partir d'un seule variable (cp1) alors que la description classique des peptides utilise au minimum les deux angles dièdres ϕ et ψ pour arriver au même résultat. De plus, notre méthode est habile à distinguer des structures légèrement différentes (à l'intérieur des individus codés C*) habituellement réunies dans le même ensemble (zone C*).

Il reste certains problèmes qui sont importants dans l'utilisation de l'analyse de donnée: le choix des variables initiales, le choix du nombre de variable secondaires (composantes principales) à conserver, et l'interprétation des indices statistiques. En effet, aurions nous obtenu le même succès en utilisant un nombre plus restreint de distances ou en choisissant ces dernières différemment? Nous pouvons le penser puisque notre analyse aboutit avec une variable ne contenant que 53.87% de l'information totale. Il doit donc y avoir des variables distances initiales non nécessaires ou encore l'information totale contenue dans l'échantillon n'est pas limitée à la position des différents minima relativement aux angles dièdres ϕ et ψ. Nous avons donc effectué la même analyse en éliminant une à la fois, les 5 distances initiales. Si nous conservons 4 des 5 distances initiales, nous n'obtenons jamais un classement correct. Selon la distance retirée des variables de départ, le classement sera plus ou moins affecté. Plus la distance retirée participe de façon importante à la construction de cp1, plus le

classement sera un échec. Ainsi, il est préférable d'éliminer des composantes principales plutôt que des distances. En effet, conserver une seule composante principale soit moins de 60% de l'information totale conduit à un classement réussi. De plus, le temps de calcul nécessaire à une analyse en composante principale est négligeable bien que dépendant du nombre de variables initiales. Pour une molécule plus grosse où il est impossible de conserver toutes les distances initiales, il faudra retirer celles qui participent le moins à la construction des premières composantes principales soit celles possédant les coefficients de combinaison linéaire les plus faibles. Nous en déduisons que l'information contenue dans nos distances et par la suite dans les composantes principales est plus importante que la seule position des minima classiques codés par les angles phi et psi de la chaîne principale.

Nous pourrions exploiter l'information supplémentaire contenue dans nos composantes principales et tenter de décrire l'ensemble de la conformation de la molécule soit la position de la chaîne latérale (habituellement décrite par les codes g^+, g^- ou t selon que l'angle χ^1 se trouve entre 0 et 120, 0 et -120 ou 120 et -120 degrés respectivement), de l'acétyle initial et du méthylamide final. Nous avons par conséquent regardé comment trier notre échantillon de manière à placer dans une même classe les individus dont les CH_3 sont orientés de la même manière en plus d'avoir les mêmes phi et psi. Cela augmenterait donc le nombre de classes or lors de notre examen des indices statistiques pour les cinq distances, nous avons constaté que le maximum de classes possibles est de 16. Ce qui signifie que pour avoir les renseignement utiles sur les chaînes latérales, nous devons soit augmenter le nombre de composantes principales utilisées, soit augmenter le nombre de distances, soit changer certaines de ces distances ce qui donne un grand nombre de possibilités à explorer ce qui sort du cadre de l'étude de l'alanine effectuée à titre démonstratif des possibilités de l'analyse de données appliquée au tri de conformations peptidiques.

CHAPITRE 2

APPLICATION DES TECHNIQUES D'ANALYSE DE DONNÉES À UNE POPULATION PEPTIDIQUE APYA

2.1 Introduction

2.1.1 Intérêt structural

Le peptide choisi pour mettre au point et tester l'analyse de population par les techniques d'analyse de données est N-acétyl-Ala-Pro-Tyr-Ala, noté par la suite APYA. Cette molécule présente des caractéristiques structurales connues puisque l'analyse de dichroïsme circulaire a montré une structure en tournant β. D'autre part, ce peptide a servi de modèle pour déterminer les caractéristiques conformationnelles nécessaires à une complexation avec le cation calcium, celui-ci étant impliqué dans de nombreux processus biologiques (59). De taille intermédiaire, ce peptide constitue un bon candidat pour tester l'applicabilité et l'intérêt des méthodes d'analyse de données aux molécules possédant une grande variabilité conformationnelle.

2.1.2 Génération de la population à analyser

Le problème des minima multiples est une des difficultés fondamentales dans l'analyse conformationnelle des molécules très flexibles et possédant de nombreux degrés de libertés. Notre approche est traduite dans le logiciel PEPSEA développé au laboratoire (40) et validé par des comparaisons avec d'autres approches (60), La population résultante est donc un échantillon de minima représentés dans l'espace des angles de torsion variables. Nous savons

d'après Li et Scheraga (2, 7, 13) que le nombre total de conformations possible est une fonction exponentielle du nombre de degrés de libertés de la molécule. La taille de l'échantillon doit donc être représentative de la taille de la population que l'on veut décrire. Nous avons généré 2073 conformations minimisées avec 11 variables initiales qui sont tous les angles phi, psi et khi de la molécule. 2000 conformations ont été retenues pour l'analyse subséquente, les conformations supplémentaires étant celles pour lesquelles l'algorithme de minimisation n'a pu trouver de minimum. Le minimum global de l'échantillon se trouve à -19.74 kcal/mol et le domaine total d'énergie couvert est de 51.26 kcal/mol. La distribution énergétique des conformères, présentée à la figure suivante, est tracée par tranche de 1kcal/mol et relativement au minimum global c-à-d que la valeur -19.74 est retirée à l'énergie de chaque conformère.

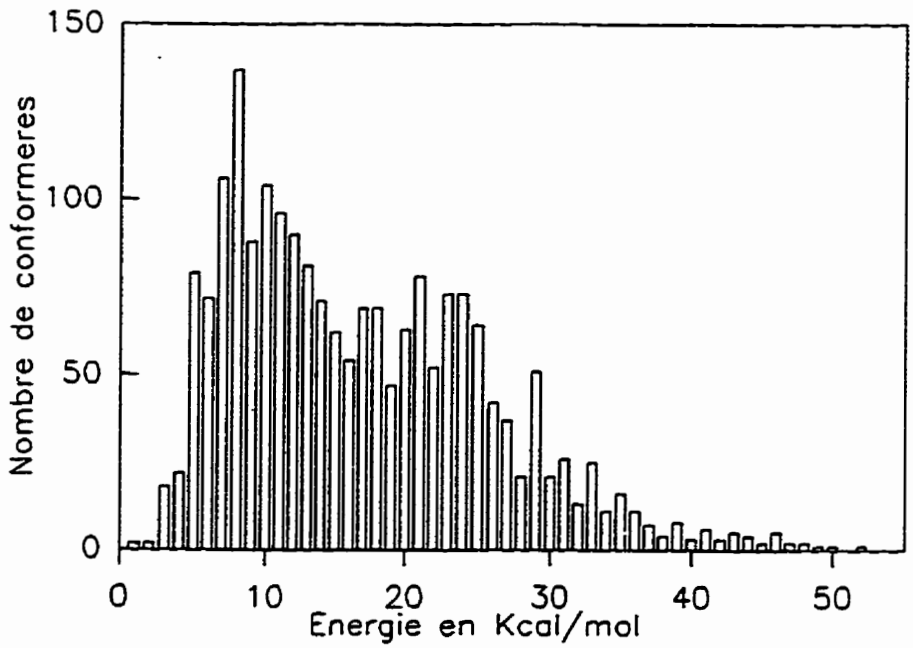


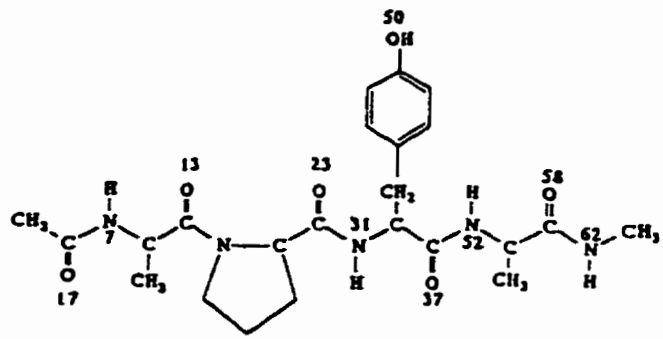
Figure 13. Distribution énergétique des 2000 conformations de l'échantillon.

L'irrégularité de la distribution traduit des contraintes conformationnelles intrinsèques à la molécule. Ces contraintes créent des barrières de rotation conduisant le processus de minimisation d'énergie vers des minima locaux. (voir la référence (58) page 67).

2.2 Analyse des données

2.2.1 Changement de variable

Ainsi qu'il a été souligné, les variables de départ caractérisant les individus pour l'analyse des données ne peuvent être constituées directement par les angles de torsion. Les distances inter-atomiques décrivant la molécule sont les suivantes:



D1	17-31	O_{Ac} N_3	C_{10} AP
D2	17-52	O_{Ac} N_4	C_{13} APY
D3	17-62	O_{Ac} $N_{mét}$	C_{16} APYA_2
D4	13-31	O_1 N_3	C_7 P
D5	13-52	O_1 N_4	C_{10} PY
D6	13-62	O_1 $N_{mét}$	C_{13} PYA_2
D7	23-7	O_2 N_1	C_8 PA
D8	23-52	O_2 N_4	C_7 Y
D9	23-62	O_2 $N_{mét}$	C_{10} YA_2
D10	37-7	O_3 N_1	C_{11} YPA
D11	37-62	O_3 $N_{mét}$	C_7 A_2
D12	50-17	O_{OH} O_{Ac}	C_{OH} Ac
D13	50-13	O_{OH} O_1	C_{OH} A
D14	50-23	O_{OH} O_2	C_{OH} P
D15	50-37	O_{OH} O_3	C_{OH} Y
D16	50-58	O_{OH} O_4	C_{OH} A_2
D17	58-7	O_4 N_1	C_{14} A_2YPA
D18	58-31	O_4 N_3	C_8 A_2Y

Figure 14. Distances inter-atomiques en Å utilisées comme variables pour l'analyse des données.

Les 18 distances retenues sont mesurées entre les atomes d'oxygène et d'azote. Elles peuvent éventuellement correspondre conformationnellement à la formation d'une liaison hydrogène. Si la distance N-O est inférieure à 3.8Å, la liaison est possible. Des statistiques élémentaires telles que la moyenne et l'écart-type ont été calculées pour ces distances (tableau 5).

Tableau 5. Statistiques élémentaires sur les 18 distances.

Variable	N	Moyenne	Ecart-type	Minimum	Maximum
D1	2000	3.06	0.38	2.62	3.65
D2	2000	5.49	0.83	3.96	7.20
D3	2000	7.26	1.43	4.25	10.74
D4	2000	3.43	0.51	2.67	4.34
D5	2000	5.46	1.13	3.18	7.51
D6	2000	7.17	1.75	2.71	10.96
D7	2000	5.79	0.72	4.05	6.72
D8	2000	3.92	0.81	2.36	5.14
D9	2000	5.90	1.53	2.58	8.66
D10	2000	7.42	1.39	3.56	10.12
D11	2000	3.82	0.81	2.47	5.11
D12	2000	8.55	1.82	5.17	11.30
D13	2000	8.46	2.06	3.34	11.53
D14	2000	7.57	1.26	4.47	9.08
D15	2000	7.07	1.16	4.95	8.84
D16	2000	8.76	2.15	3.80	12.18
D17	2000	8.91	2.21	2.67	13.44
D18	2000	5.68	0.72	3.96	7.11

La première étape dans l'étude des distances est de vérifier l'homogénéité de l'échantillon caractérisé par ces distances. Pour cela, le critère de gauss est utilisé: pour chaque distance, tous les individus doivent être compris dans l'intervalle de $\mu-3\sigma$ à $\mu+3\sigma$. Nous constatons que nos 2000 individus remplissent cette condition pour chacune des 18 distances. Cela montre qu'aucun individu ne s'écarte de la tendance moyenne (par suite d'une erreur dans les données ou le calcul par exemple). Cette caractéristique est essentielle puisque les techniques d'analyse de données sont facilement faussées par la présence de données aberrantes. De plus,

si nous regardons le minimum des distances D2, D3, D7, D12, D14, D15 et D18, nous constatons que ce minimum n'est pas compatible avec l'existence de liaisons hydrogène puisque supérieur à 3.8Å pour toutes ces distances. Nous choisirons donc de retirer ces distances des variables caractérisant nos individus de manière à ne conserver que celles pour lesquelles le lien est compatible avec des caractéristiques structurales. Notre échantillon de départ pour l'analyse des données sera donc constitué par 2000 individus caractérisés chacun par 11 distances inter-atomiques.

2.2.2 Analyse en composantes principales

La première étape de l'analyse en composantes principales consiste à calculer les coefficients de corrélation entre les 11 distances initiales. La matrice des corrélations est ensuite diagonalisée pour obtenir les valeurs propres et les vecteurs propres correspondants. Cette diagonalisation permet la construction des nouvelles variables, composantes principales (notées "cp") à partir des variables distances initiales. Le résultat de l'ACP, effectué à l'aide du programme PRINCOMP faisant partie du logiciel SAS est présenté au tableau suivant. Le temps de calcul de cette analyse est de l'ordre de quelques dizaines de secondes.

La première indication sur le résultat de l'analyse en composantes principale nous est fournie par l'examen de la valeur propre associée à chaque composante principale, ces dernières étant classées par ordre de représentativité décroissant. A la première composante principale est associée la variance maximale. Cette variance, exprimée en pourcentage de la variance totale donnera la part d'information portée par la composante principale. Ainsi, la première composante principale porte 28.48% de toute l'information contenue dans les 11 distances de départ. Nous constatons ainsi que les 7 premières composantes principales portent 90.62% de l'information contenue dans l'échantillon, et seront retenues pour l'analyse de regroupement subséquente. En effet, les 4 dernières composantes ne portant qu'une part minime de l'information alourdiraient inutilement le calcul si elles étaient conservées.

Tableau 6. Résultats de l'analyse en composantes principales.

Matrice des corrélations

	D1	D4	D5	D6	D8	D9	D10	D11	D13	D16	D17
D1	1.0000	0.4579	0.3126	0.2149	-.0153	-.0481	0.4994	-.0283	-.0341	0.0091	0.2867
D4	0.4579	1.0000	0.3453	0.1275	0.0952	0.0578	0.2218	0.0072	0.1499	-.0202	0.0278
D5	0.3126	0.3453	1.0000	0.7553	0.2921	0.1635	0.2049	0.0115	-.2556	-.1210	0.6350
D6	0.2149	0.1275	0.7553	1.0000	0.2533	0.4350	0.2090	0.1607	-.2594	-.2235	0.5964
D8	-.0153	0.0952	0.2921	0.2533	1.0000	0.7369	-.1596	0.0825	-.0347	-.1872	0.1230
D9	-.0481	0.0578	0.1635	0.4350	0.7369	1.0000	-.1391	0.3054	-.0167	-.2563	0.0687
D10	0.4994	0.2218	0.2049	0.2090	-.1596	-.1391	1.0000	-.0095	-.2783	0.0487	0.4661
D11	-.0283	0.0072	0.0115	0.1607	0.0825	0.3054	-.0095	1.0000	0.0052	-.0215	-.0290
D13	-.0341	0.1499	-.2556	-.2594	-.0347	-.0167	-.2783	0.0052	1.0000	-.3585	-.3710
D16	0.0091	-.0202	-.1210	-.2235	-.1872	-.2563	0.0487	-.0215	-.3585	1.0000	-.1500
D17	0.2867	0.0278	0.6350	0.5964	0.1230	0.0687	0.4661	-.0290	-.3710	-.1500	1.0000

Valeurs propres

	Valeurs propres	Pourcentage de variance	Pourcentages Cumulés
cp1	3.13243	0.284766	0.28477
cp2	2.10004	0.190913	0.47568
cp3	1.46654	0.133322	0.60900
cp4	1.11075	0.100978	0.70998
cp5	0.96192	0.087447	0.79743
cp6	0.75803	0.068911	0.86634
cp7	0.43881	0.039892	0.90623
cp8	0.37868	0.034425	0.94065
cp9	0.35393	0.032176	0.97283
cp10	0.22068	0.020062	0.99289
cp11	0.07819	0.007108	1.00000

Vecteurs propres

	cp1	cp2	cp3	cp4	cp5	cp6	cp7	cp8	cp9	cp10	cp11
D1	0.276573	-.303562	0.399344	0.211034	-.006844	-.242720	0.656937	-.227448	-.288482	0.042558	-.007349
D4	0.202293	-.104562	0.568975	0.322766	-.235306	0.280922	-.533753	-.054883	-.037632	0.278827	0.137416
D5	0.477777	-.023322	-.009736	-.126399	-.155512	0.464019	0.035793	0.162140	-.107744	-.439911	-.531898
D6	0.474914	0.109071	-.122647	-.135220	0.116186	0.282554	0.153873	-.361878	0.444611	-.060873	0.529038
D8	0.232145	0.473380	-.009622	0.173305	-.390665	-.315712	0.016587	0.388976	-.210072	-.297715	0.394028
D9	0.235782	0.533389	-.026583	0.238146	-.045277	-.284225	0.010814	-.243127	0.318930	0.320822	-.505746
D10	0.264055	-.414467	0.065124	0.088238	0.247475	-.529854	-.329109	0.130726	0.379475	-.362462	-.060779
D11	0.071692	0.220911	-.044627	0.489183	0.754584	0.206530	-.010626	0.180003	-.234700	-.054865	0.045909
D13	-.201214	0.207784	0.593664	-.221817	0.143770	0.105939	0.284070	0.450180	0.445924	-.013508	-.020432
D16	-.117862	-.293377	-.320667	0.588490	-.301253	0.214490	0.262512	0.293698	0.388433	0.089856	0.002408
D17	0.435940	-.161416	-.198388	-.296089	0.100838	-.077025	0.012546	0.489067	-.106837	0.623707	0.031702

L'examen des vecteurs propres permet de voir à partir de quelles distances, et dans quelle proportion, ont étés construites les composantes principales. En effet, chaque composante

principale est la combinaison linéaire des 11 distances initiales et les vecteurs propres sont les coefficients de combinaison linéaire. Les distances pour lesquelles les coefficients de combinaison linéaire seront élevés seront celles qui auront contribuées le plus à la formation de cette composante principale. Nous pouvons ainsi relier directement les composantes principales aux distances importantes, celles qui traduisent les tendances conformationnelles de la molécule.

Les distances D5, D6 et D17 participent fortement à la formation de la première composante principale. Ces distances correspondent à la formation éventuelle de structures secondaires soit un tournant β sur les résidus Pro-Tyr, un cycle à 13 membres (qui est le type de tournant constituant les hélices α lorsque répétés plusieurs fois) sur les résidus Pro-Tyr-Ala et un cycle "inverse" (formation d'un lien hydrogène dans le sens C-terminal/ N-terminal plutôt que dans le sens normal du peptide c-à-d N-terminal/ C-terminal) à 14 membres impliquant les résidus Ala-Tyr-Pro-Ala. Ainsi, la première composante principale, la plus représentative de la variabilité de la molécule, traduit une organisation à longue distance sur la molécule. Plus précisément, les structures secondaires décrites ici traduisent un repliement de la molécule autour des résidus centraux Pro et Tyr.

La deuxième composante principale est formée majoritairement à partir des distances D8, D9, et D10. Ces distances correspondent à la formation éventuelle des structures secondaires suivantes: un pseudo-cycle à 7 membres sur la tyrosine, un tournant β sur les résidus Tyr-Ala et à la formation d'un pseudo-cycle "inverse" à 11 membres centré sur les résidus Tyr-Pro-Ala. Ces structures traduisent une organisation à plus courte distance que dans le cas de la première composante principale et centrée autour du résidu tyrosine.

La troisième composante principale est formée par les distances D1, D4 et D13. Les structures secondaires correspondant à ces distances sont un tournant β sur les résidus Ala-

Pro, un pseudo-cycle à 7 membres sur la proline et une structure impliquant un lien entre la fonction alcool de la tyrosine et le carbonyle du premier résidu d'alanine.

Si nous classons les distances par ordre d'importance, c'est à dire de représentativité, nous aurons: D5, D6, D17, D9, D8, D10, D13, D4, D1. Il est généralement admis que le repliement des peptides et protéines est assuré par la formation de structures secondaires classiques en tournant β, γ et hélice α. Il est intéressant de constater que certaines de ces distances correspondent à la formation de structures secondaires non-classiques dans la description des conformations peptidiques. En effet, nous avons utilisé des distances correspondant à la formation de pseudo-cycles "inverses". Il apparait que ces distances décrivent très adéquatement un échantillon de peptide. De plus, la septième distance qui décrit le mieux l'échantillon implique un repliement de la chaîne latérale de la tyrosine sur la chaîne principale du peptide. Cela suggère qu'une part non-négligeable du processus de repliement des peptides peut être induit par un chaîne latérale.

2.2.3 Méthodes de regroupement

Nous désirons par le biais de ces méthodes déterminer s'il existe des classes ou groupes éventuels à l'intérieur de notre population. Si nous sommes capable d'affecter chaque individu à un groupe, nous pourrons décrire toutes les possibilités conformationnelles de ce peptide à partir de quelques molécules qui seront chacune affectées d'un poids de représentativité. Ce poids étant relatif à la taille, l'homogénéité, la distribution énergétique de la classe à laquelle cette molécule appartient. Le problème majeur de la classification est de décider quel est le nombre de classes existantes dans cet échantillon. Les indices statistiques aident à résoudre généralement ce problème. D'autre part, les différentes méthodes de classification sont toutes biaisées de manière différente (certaines tendent à former des classes de variance faible, d'autres tendent à placer un nombre égal d'individus dans chaque classe...). Une manière de s'affranchir de ce problème est d'utiliser conjointement plusieurs méthodes, de préférence

hiérarchiques et non-hiérarchiques, jusqu'à obtenir des résultats consistants pour plusieurs types de classification.

2.2.3.1 Les indices statistiques

Ces indices n'étant pas dénués de restrictions quant à leur interprétation, il est nécessaire d'interpréter conjointement les résultats des trois tests soit le CCC (Cubic Clustering Criterion), le psF (pseudo F) et le pst^2 (pseudo t^2). Nous rappelons que ces tests suggèrent la présence d'un certain nombre de familles dans l'échantillon lorsqu'un maximum pour CCC et psF coïncide avec un minimum, immédiatement suivi d'un maximum pst^2 (voir paragraphe 1.3.4, 1.3.5, 1.5.1) . Nous présentons ici les tracés de ces trois indices pour les trois méthodes de classification hiérarchique que nous envisageons d'employer soit AVERAGE, WARD et CENTROID telles qu'implantées dans le logiciel SAS sous l'option CLUSTER. Pour la classification non-hiérarchique FASTCLUS, le calcul de l'indice Pseudo t^2 ne s'appliquant pas, nous présentons les résultats pour CCC et Pseudo F.

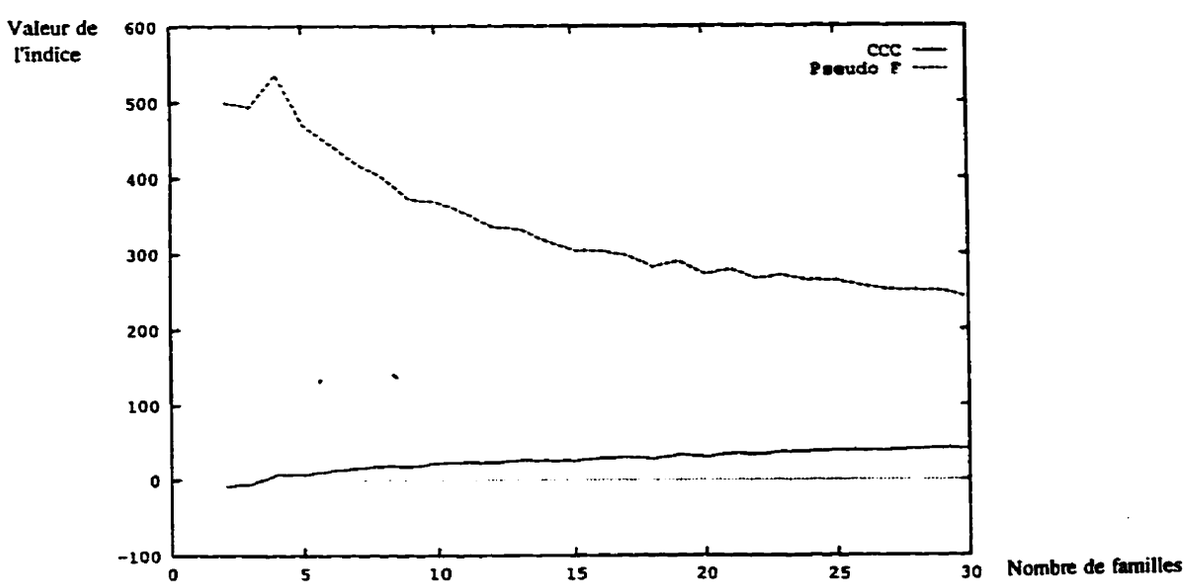


Figure 15. Indices statistiques pour la classification option FASTCLUS.

Les indices statistiques CCC et Pseudo F s'accordent quant au choix de 4 familles.

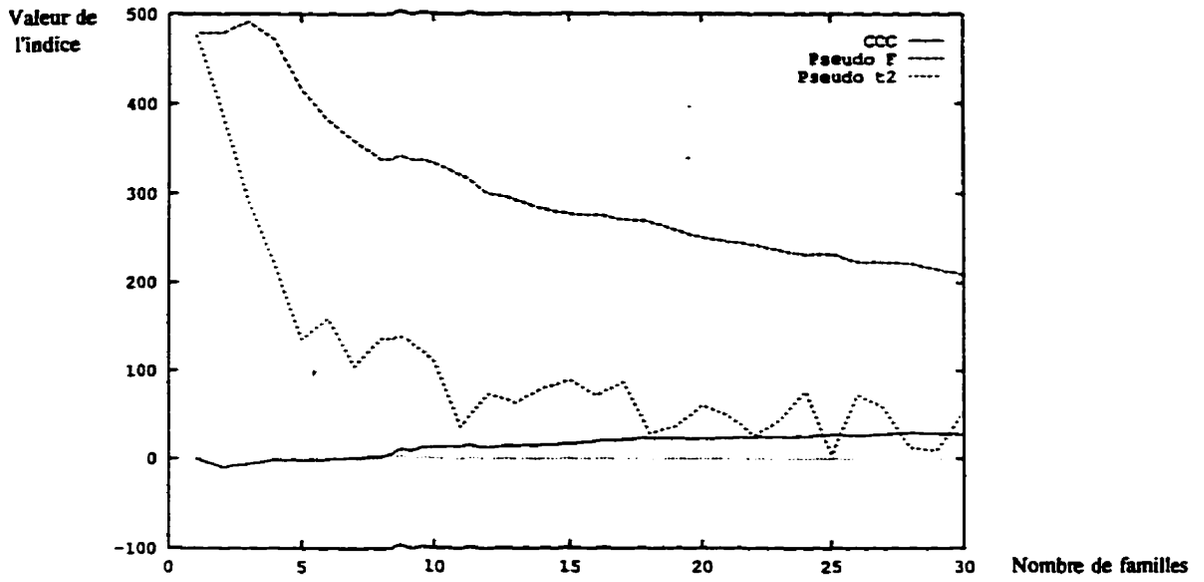


Figure 16. Indices statistiques pour la classification hiérarchique option AVERAGE.

Les indices CCC et pst^2 présentent un accord pour 5 ou 7 classes. Le psF est plus difficile à interpréter.

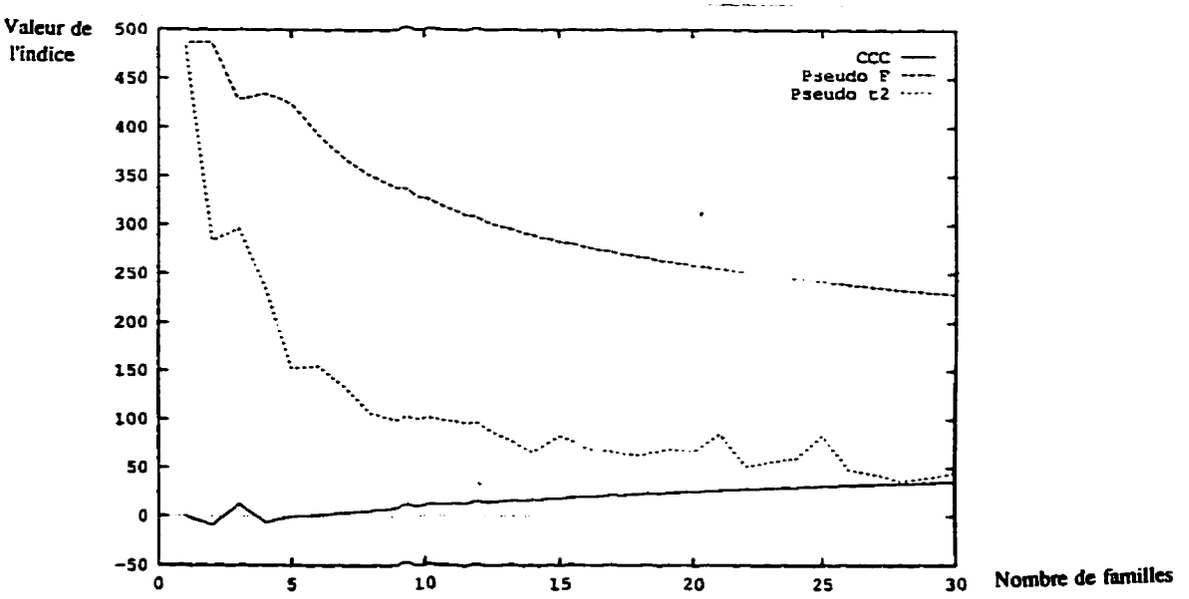


Figure 17. Indices statistiques pour la classification hiérarchique option WARD.

Les indices statistiques présentent un accord quant au choix de 5 familles puisque nous observons un pic pour la valeur de CCC et pseudo F correspondant à 5 familles associé à un minimum pour pseudo t^2.

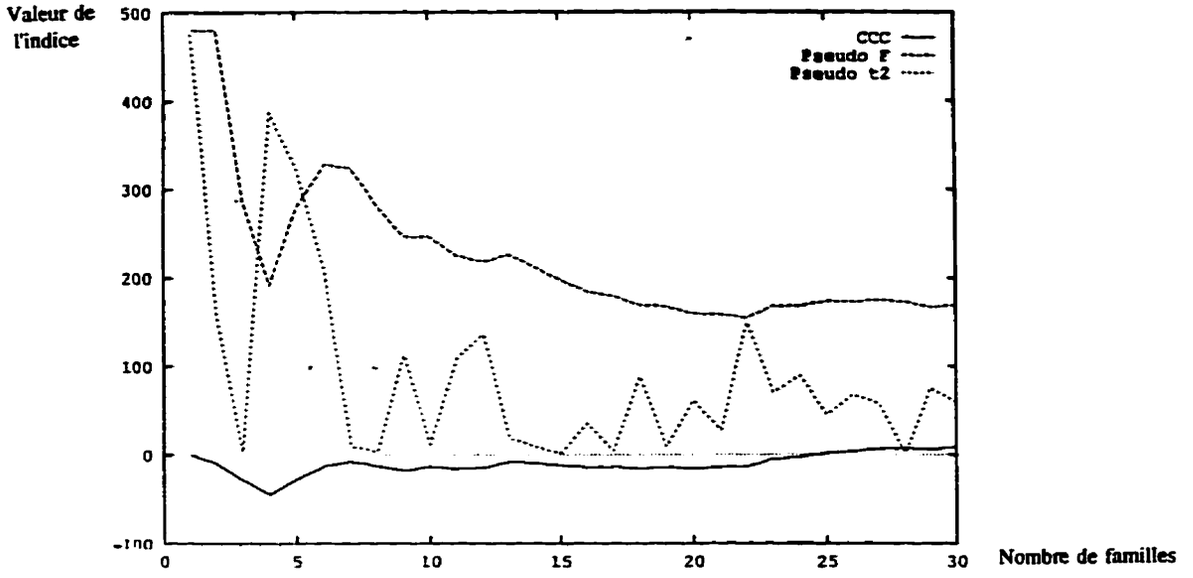


Figure 18. Indices statistiques pour la classification hiérarchique option CENTROID.

Les indices statistiques montrent un accord pour le choix de 4 ou 6 familles dans l'échantillon.

Nous constatons à l'examen de ces courbes que les indices statistiques peuvent être délicats à interpréter c'est pourquoi il est judicieux de les examiner pour différentes méthodes de classification ce qui permet de dégager un consensus pour différentes méthodes. Nous constatons qu'ici, il y a un accord pour les méthodes AVERAGE et WARD pour 5 familles. Pour la méthode CENTROID en revanche, nous avons la possibilité de 4 ou 6 familles. Nous décidons de faire une classification avec 5 familles finales, même pour la méthode CENTROID et de comparer ensuite les résultats.

2.2.3.2 Classification de l'échantillon en familles

2.2.3.2.1 Méthode

Nous avons choisi d'utiliser concurremment une méthode de classification non-hiérarchique (option FASTCLUS du logiciel SAS) et trois méthodes de classification ascendante hiérarchique (méthodes AVERAGE, WARD et CENTROID de l'option CLUSTER du logiciel SAS), ces dernières étant biaisées de différentes façons. La méthode AVERAGE tend à fusionner les classes possédant de faibles variances et produit des classes de même variance. La méthode WARD fusionne les classes possédant le même nombre d'observations et produit des classes ayant à peu près le même nombre d'observations. La méthode CENTROID est moins performante que les deux précédentes. Nous l'avons néanmoins utilisée car elle est moins sensible aux individus isolés ou "outliers" c'est à dire des individus dont les caractéristiques sont très différentes des tendances générales de l'échantillon. L'échantillon à classifier est constitué par 2000 conformères caractérisés par leurs coordonnées dans l'espace des 7 premières composantes principales. Nous rappelons que les variables pour la classification se doivent d'être non-corrélées. Le choix des variables composantes principales permet de remplir cette condition puisqu'elles sont linéairement indépendantes par construction, ce qui n'est pas le cas des variables distances.

2.2.3.2.2 Résultats

Les résultats obtenus pour les quatre méthodes doivent être validés mathématiquement et chimiquement. Pour montrer que ces méthodes sont capables de constituer des familles d'individus à l'intérieur de notre échantillon, nous présentons la projection des individus dans l'espace des variables canoniques. Les individus doivent être regroupés en autant de nuages distincts qu'il existe de familles à l'intérieur de l'échantillon. Pour montrer que les individus classés dans une même famille présentent des caractéristiques conformationnelles identiques, nous utilisons les descripteurs habituels aux molécules peptidiques soit l'étude des codes conformationnels de chaque résidu et la présentation de superpositions graphiques des

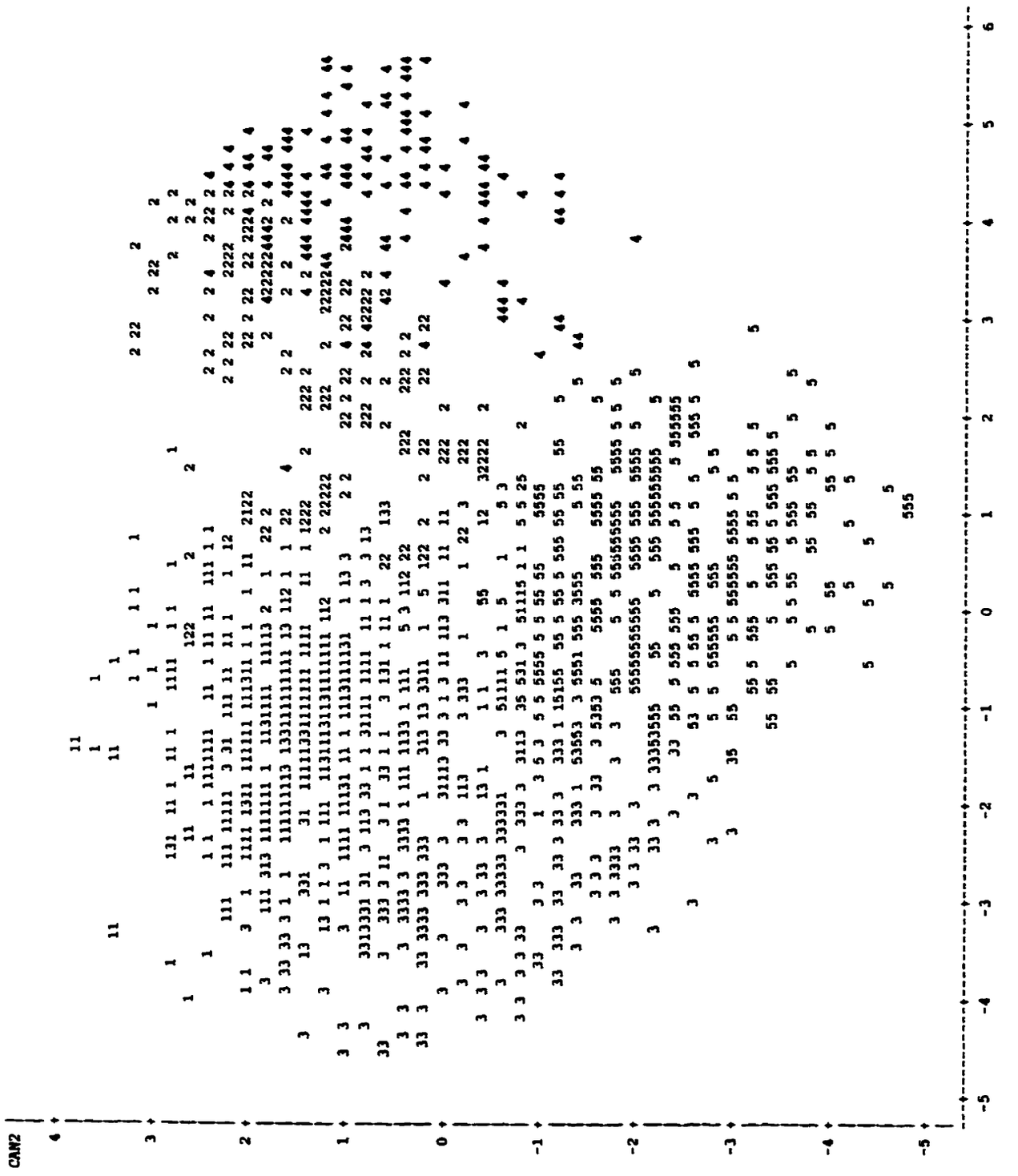


Figure 19. Projection des individus repérés par le numéro de leur famille dans l'espace des variables CAN1 et CAN2 canoniques pour la classification par FASTCLUS.

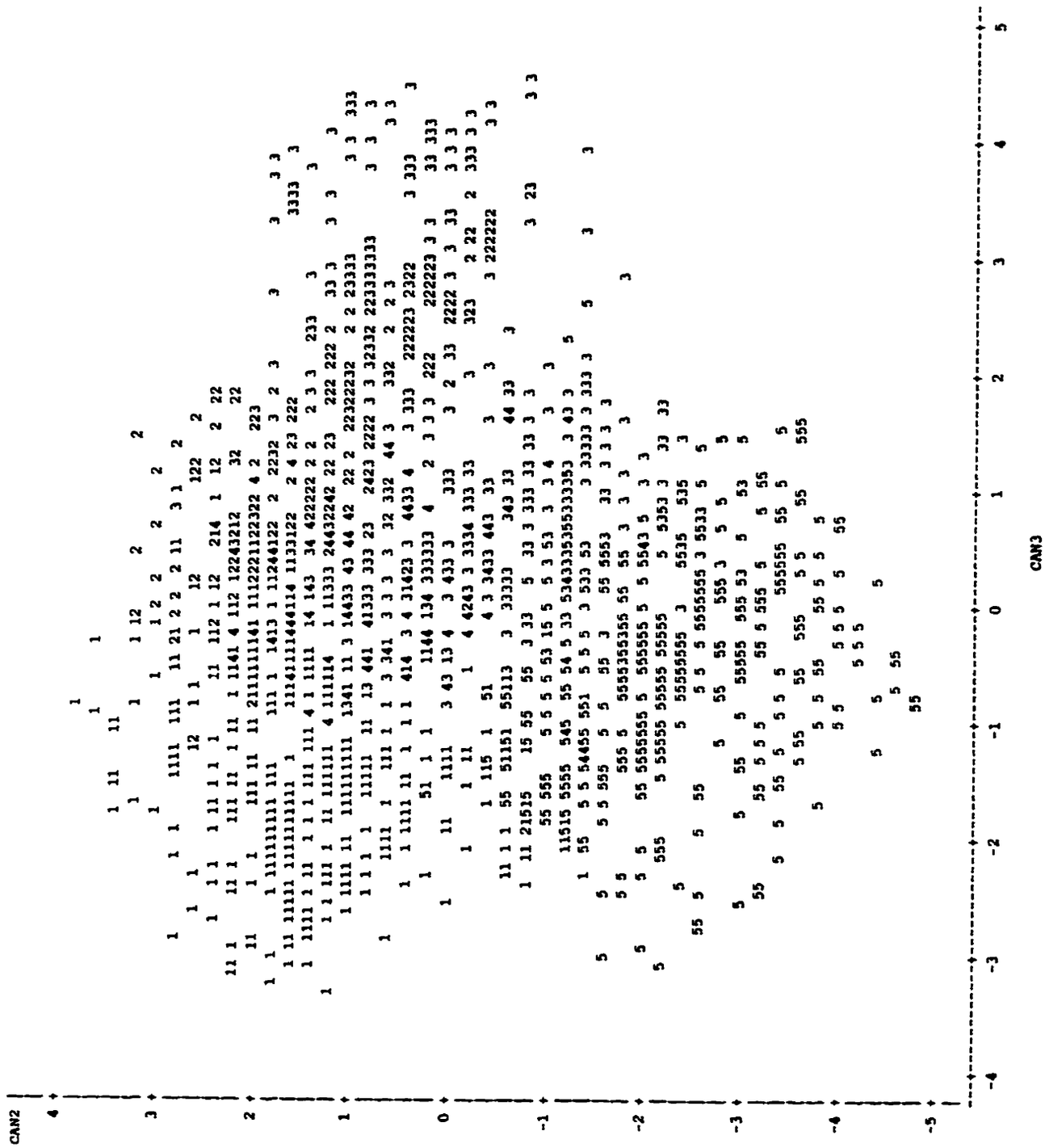


Figure 20. Projection des individus repérés par le numéro de leur famille dans l'espace des variables CAN2 et CAN3 canoniques pour la classification par FASTCLUS.

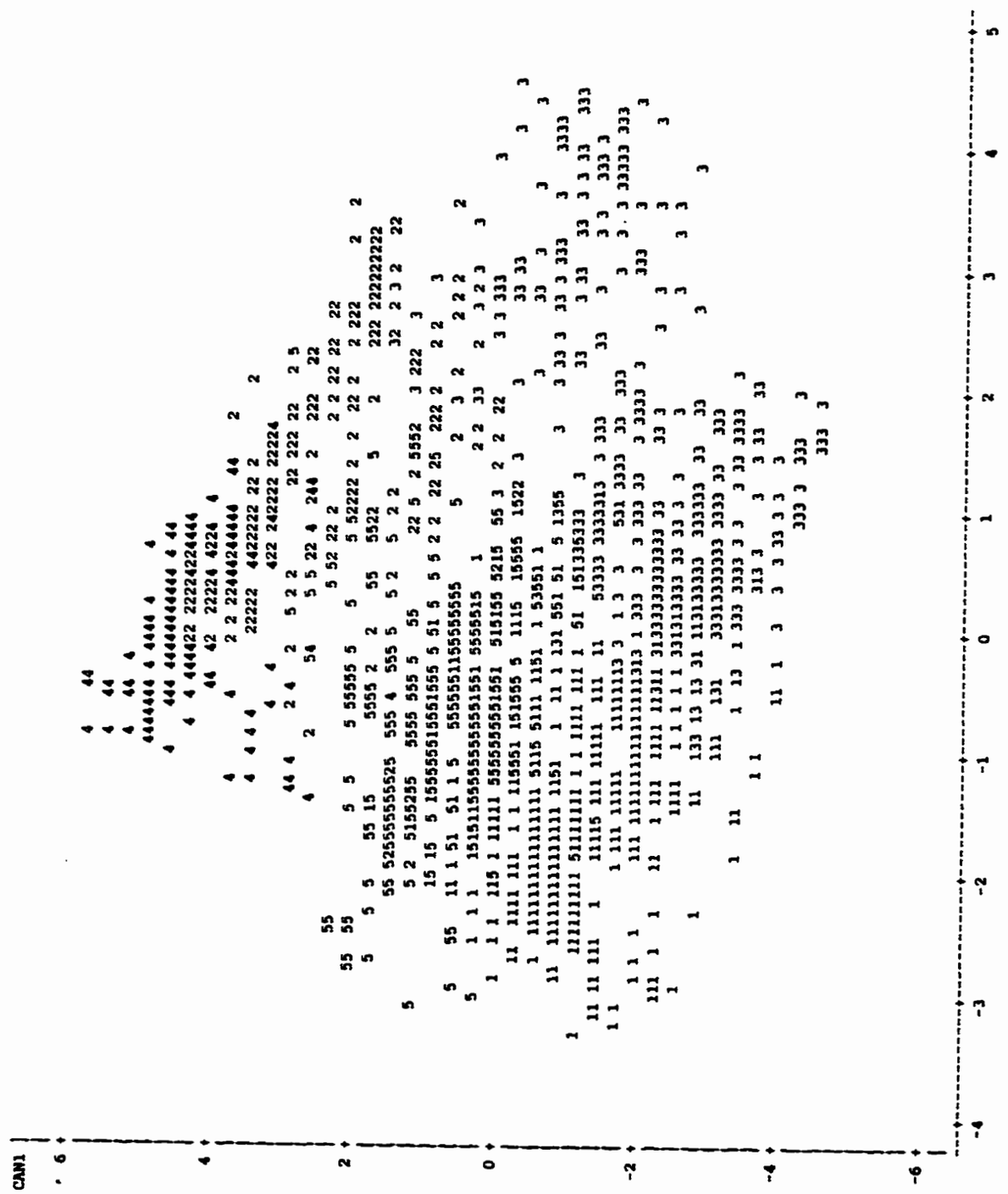


Figure 21. Projection des individus repérés par le numéro de leur famille dans l'espace des variables CAN1 et CAN3 canoniques pour la classification par FASTCLUS.

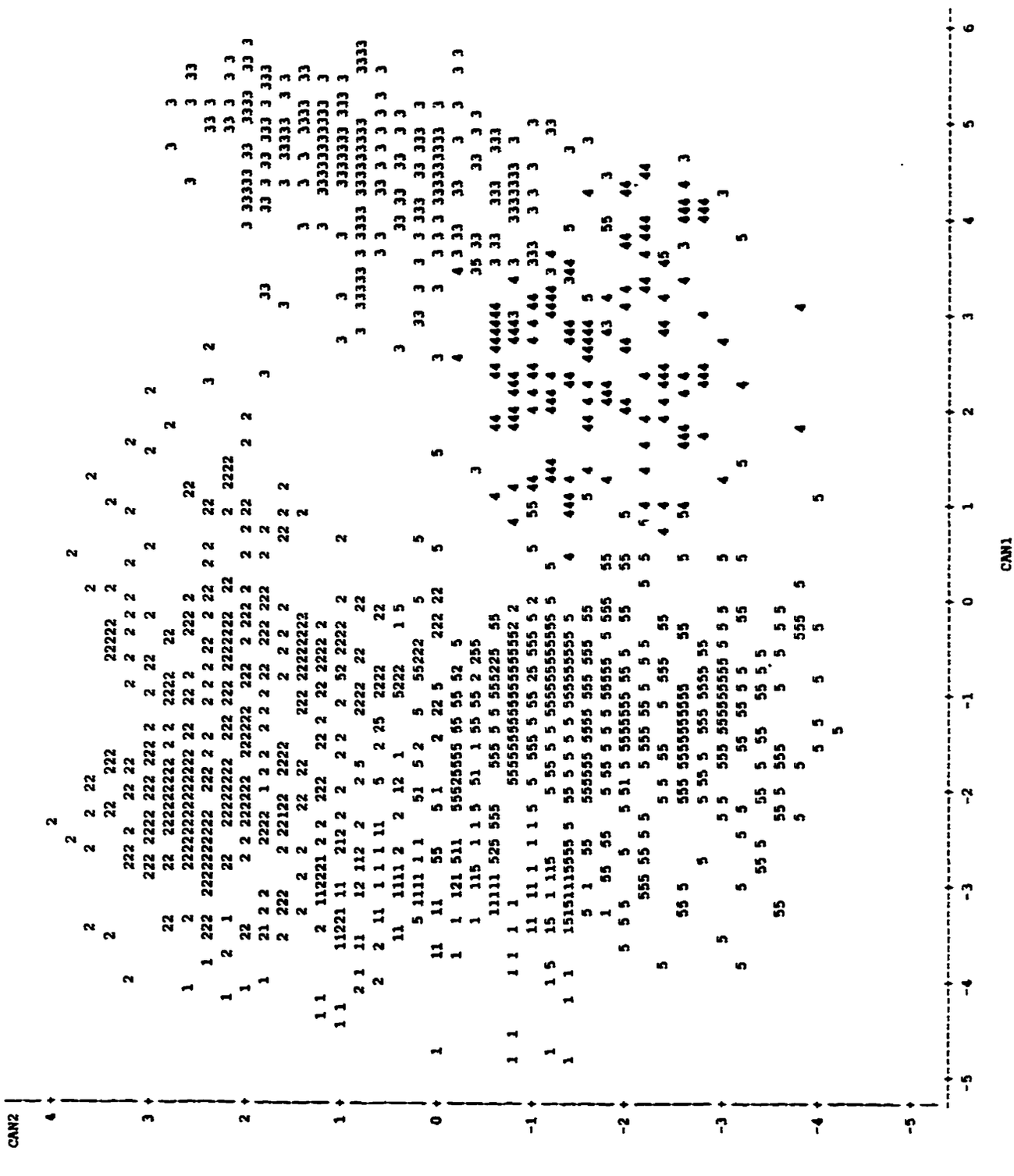


Figure 22. Projection des individus repérés par le numéro de leur famille dans l'espace des variables CAN1 et CAN2 canoniques pour la classification par AVERAGE.

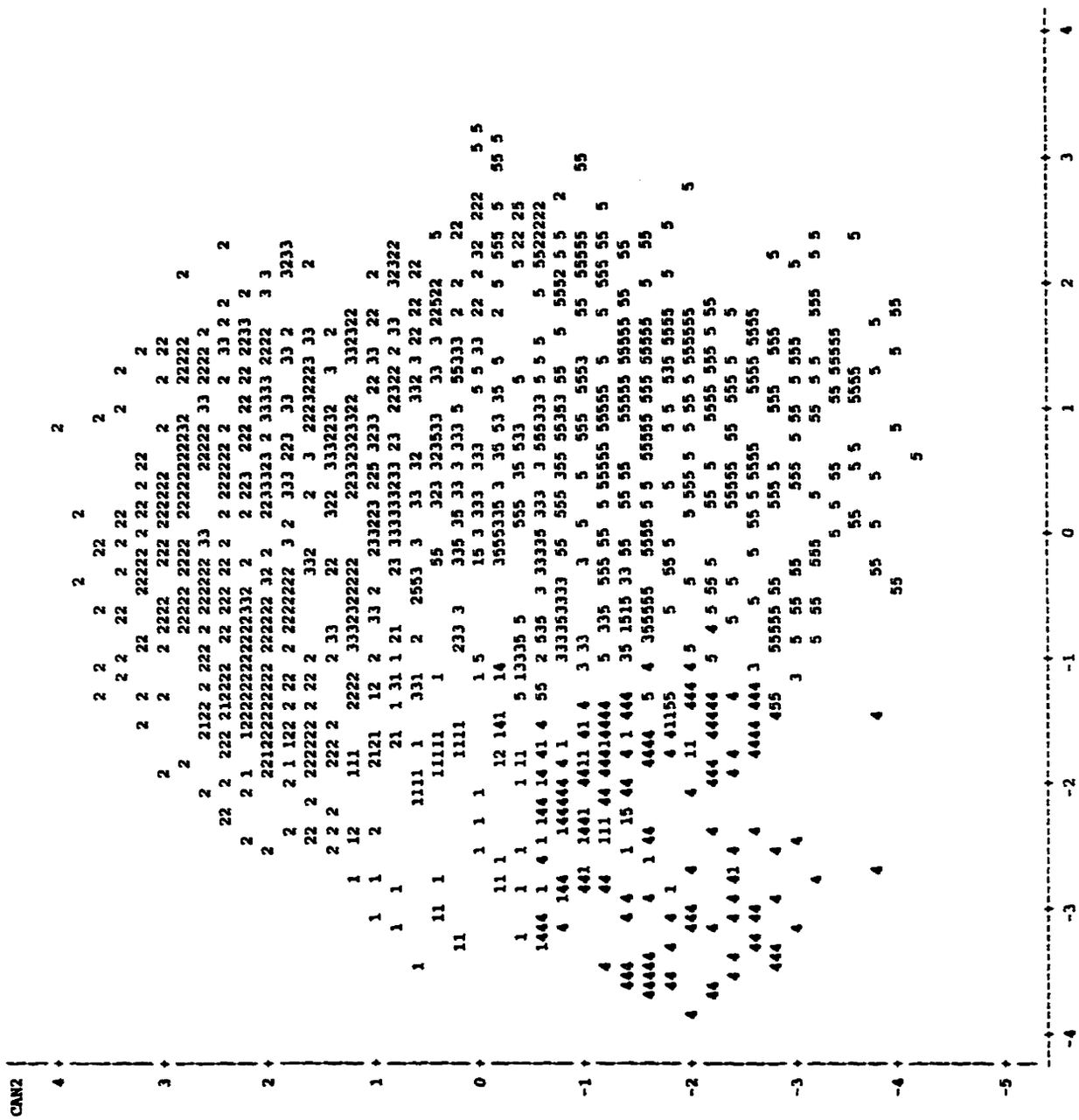


Figure 23. Projection des individus repérés par le numéro de leur famille dans l'espace des variables CAN2 et CAN3 canoniques pour la classification par AVERAGE.

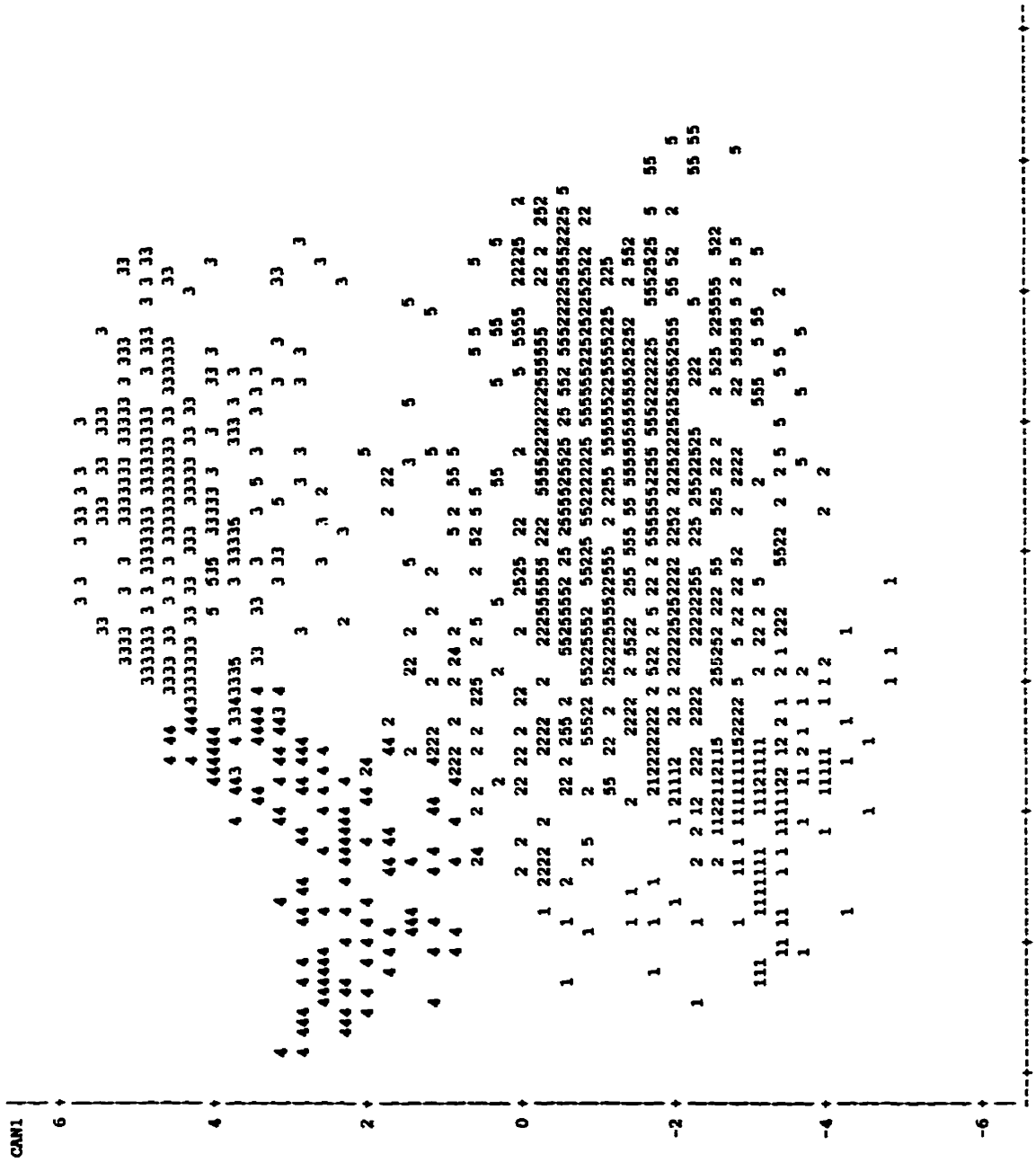


Figure 24. Projection des individus repérés par le numéro de leur famille dans l'espace des variables CAN1 et CAN3 canoniques pour la classification par AVERAGE.

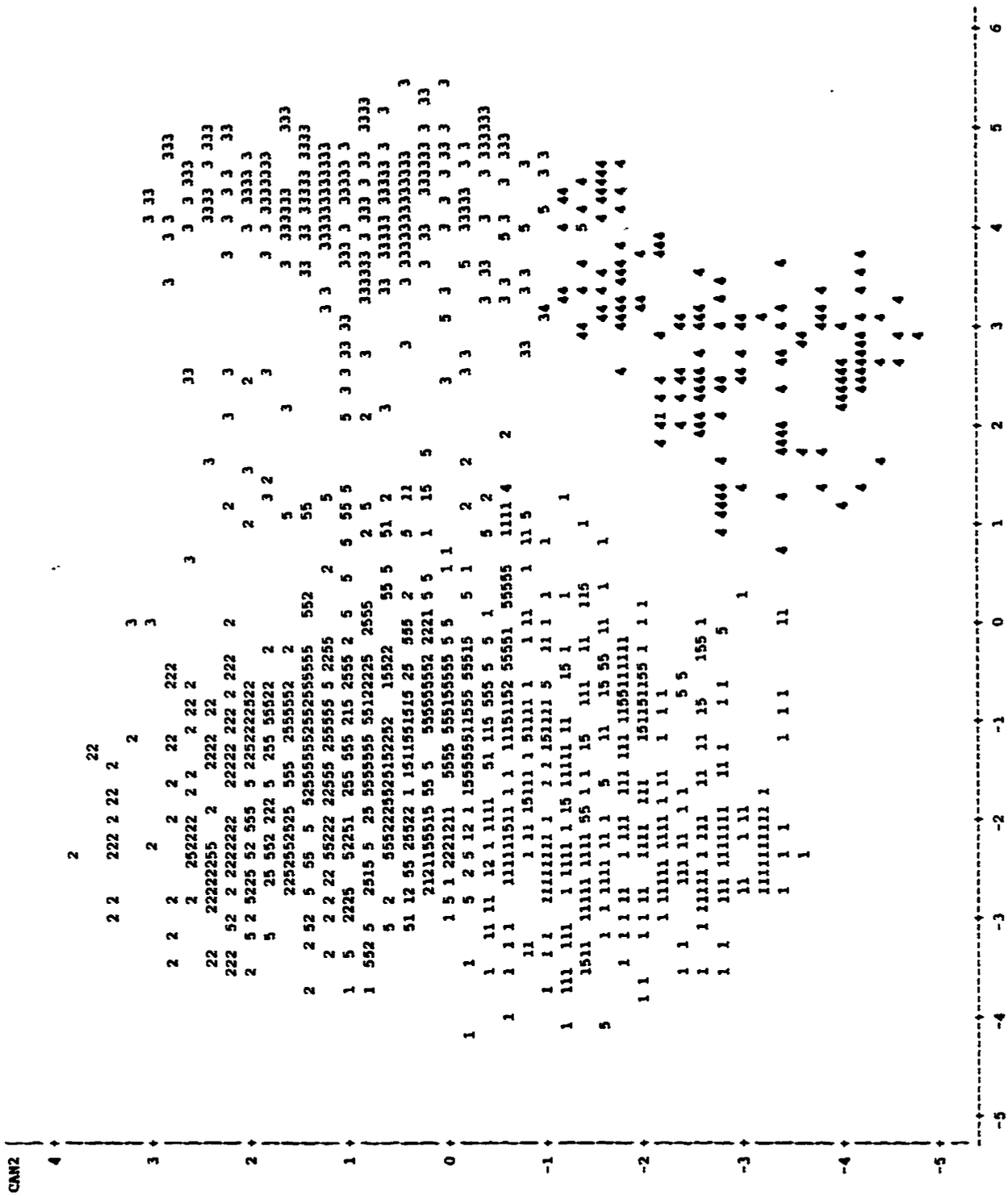


Figure 25. Projection des individus repérés par le numéro de leur famille dans l'espace des variables CAN1 et CAN2 canoniques pour la classification par WARD.

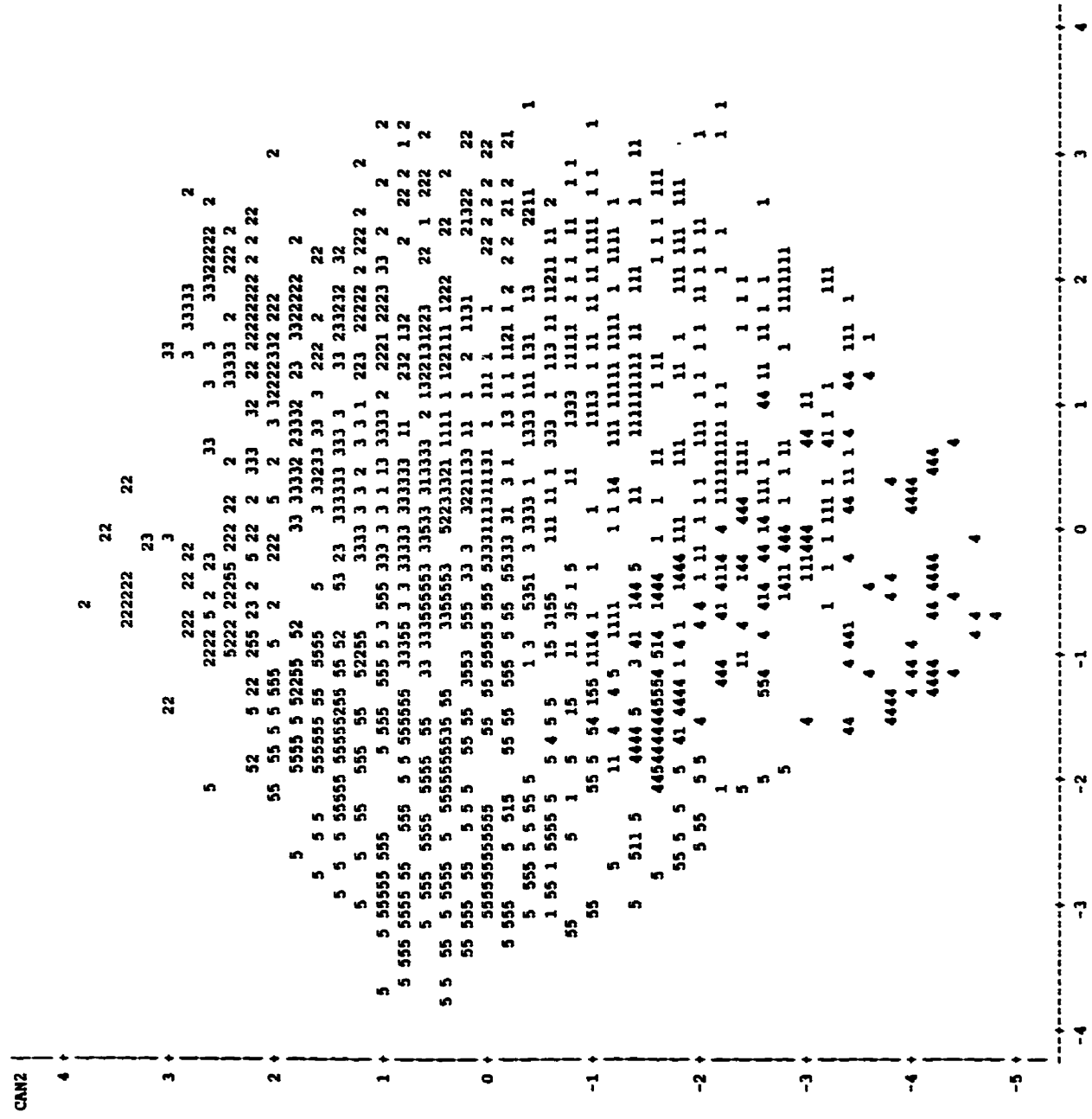


Figure 26. Projection des individus repérés par le numéro de leur famille dans l'espace des variables canoniques CAN2 et CAN3 pour la classification par WARD.

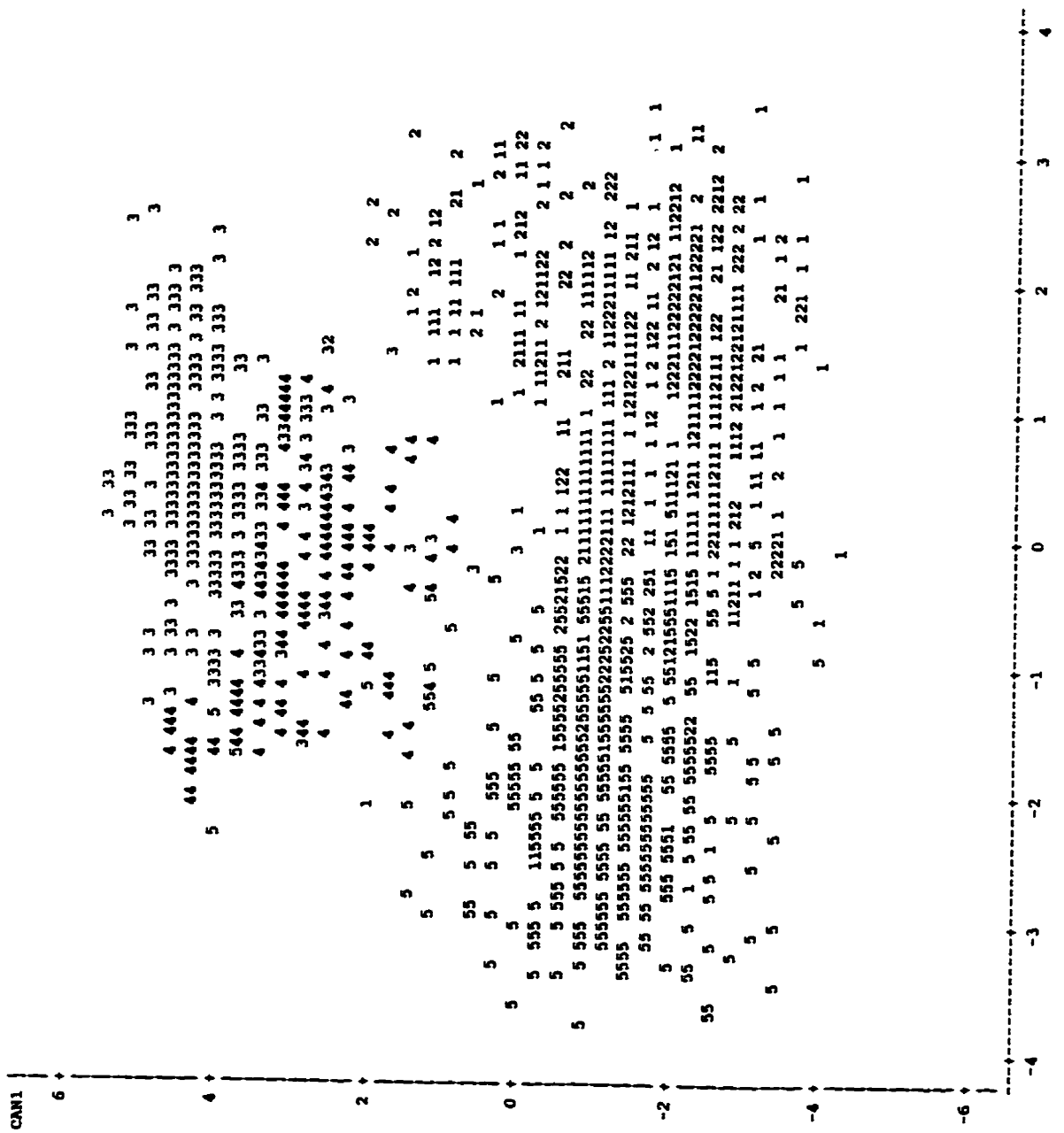


Figure 27. Projection des individus repérés par le numéro de leur famille dans l'espace des variables canoniques CAN1 et CAN3 pour la classification par WARD.

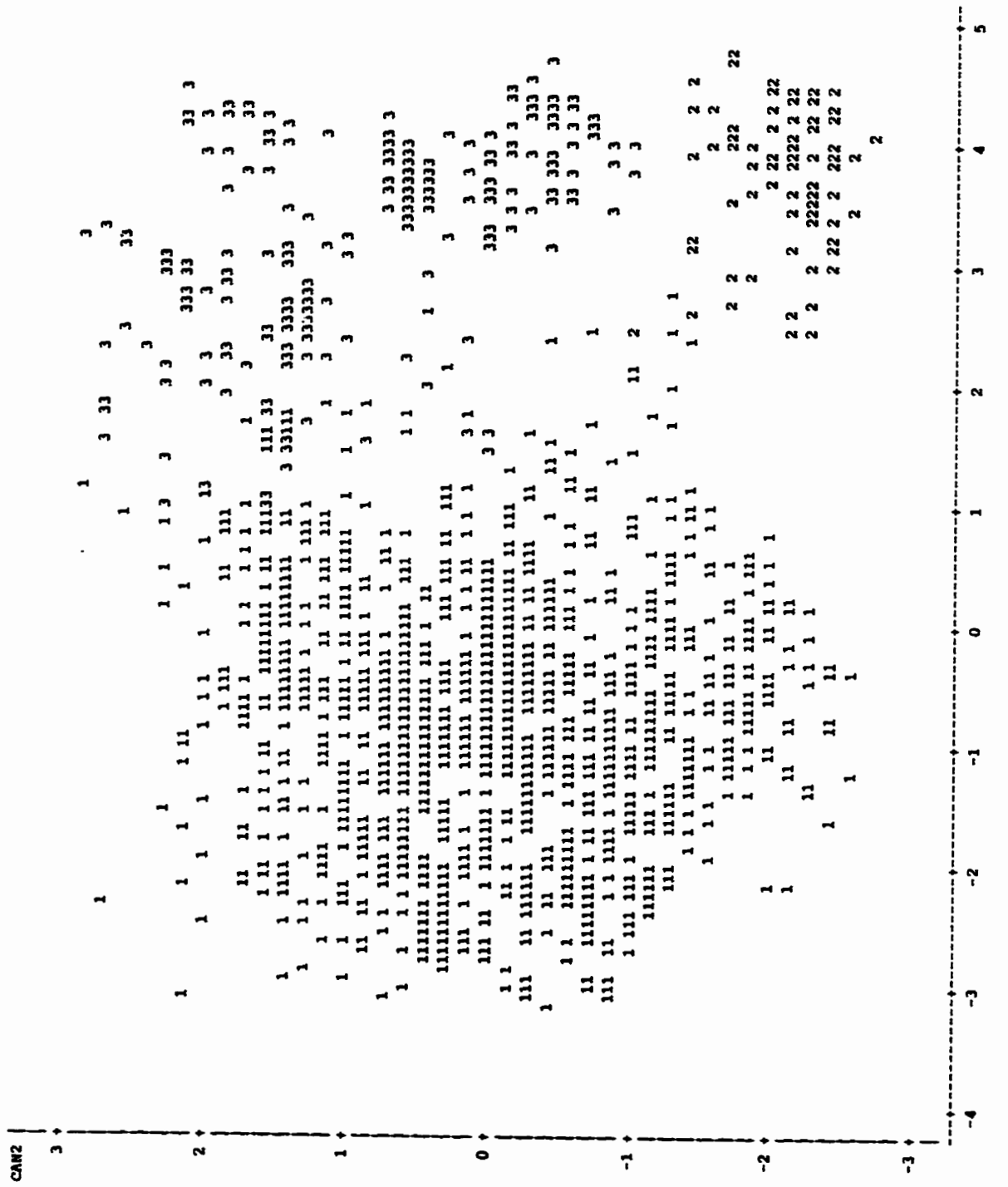


Figure 28. Projection des individus repérés par le numéro de leur famille dans l'espace des variables canoniques CAN1 et CAN2 pour la classification par CENTROID.

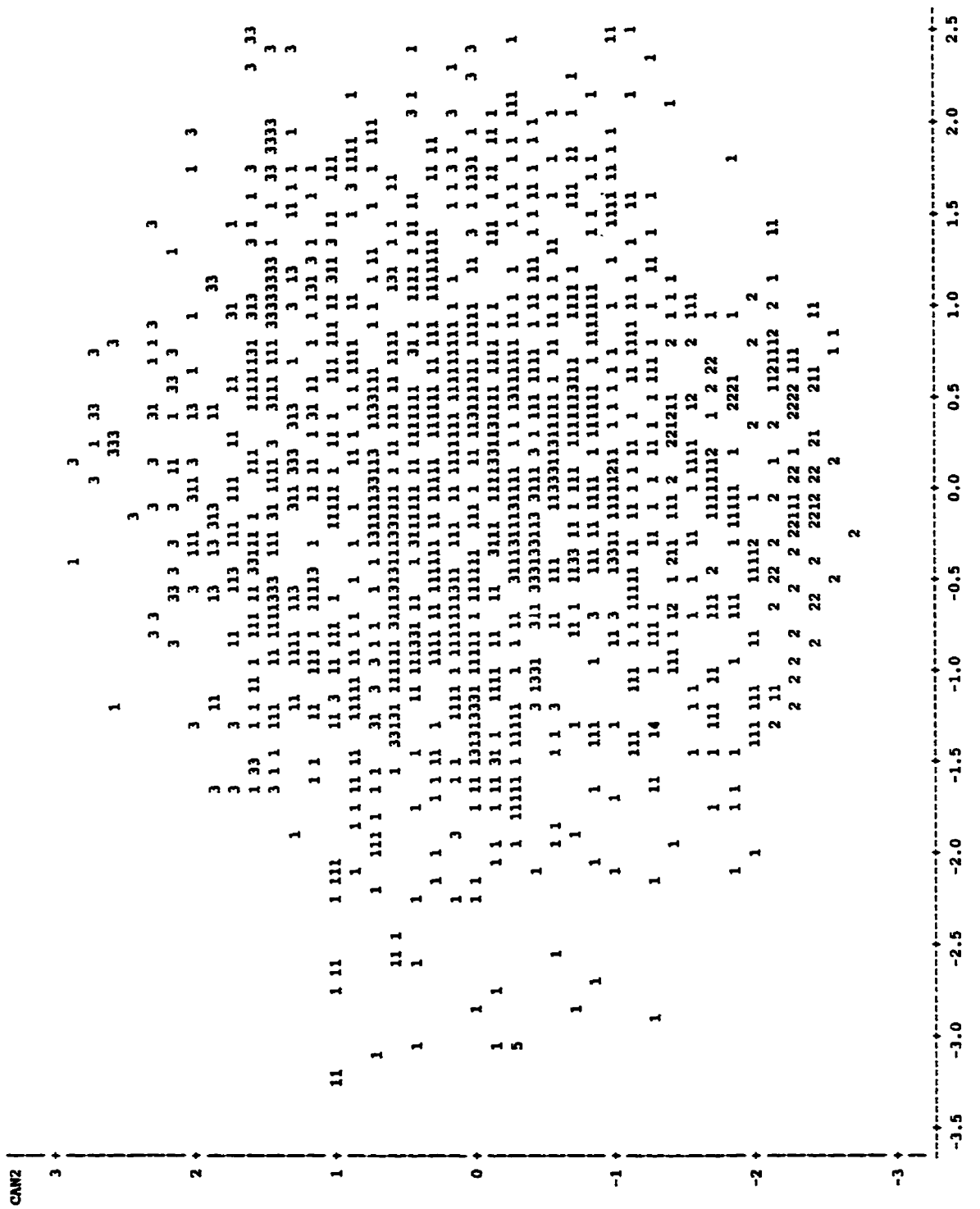


Figure 29. Projection des individus repérés par le numéro de leur famille dans l'espace des variables canoniques CAN2 et CAN3 pour la classification par CENTROID.

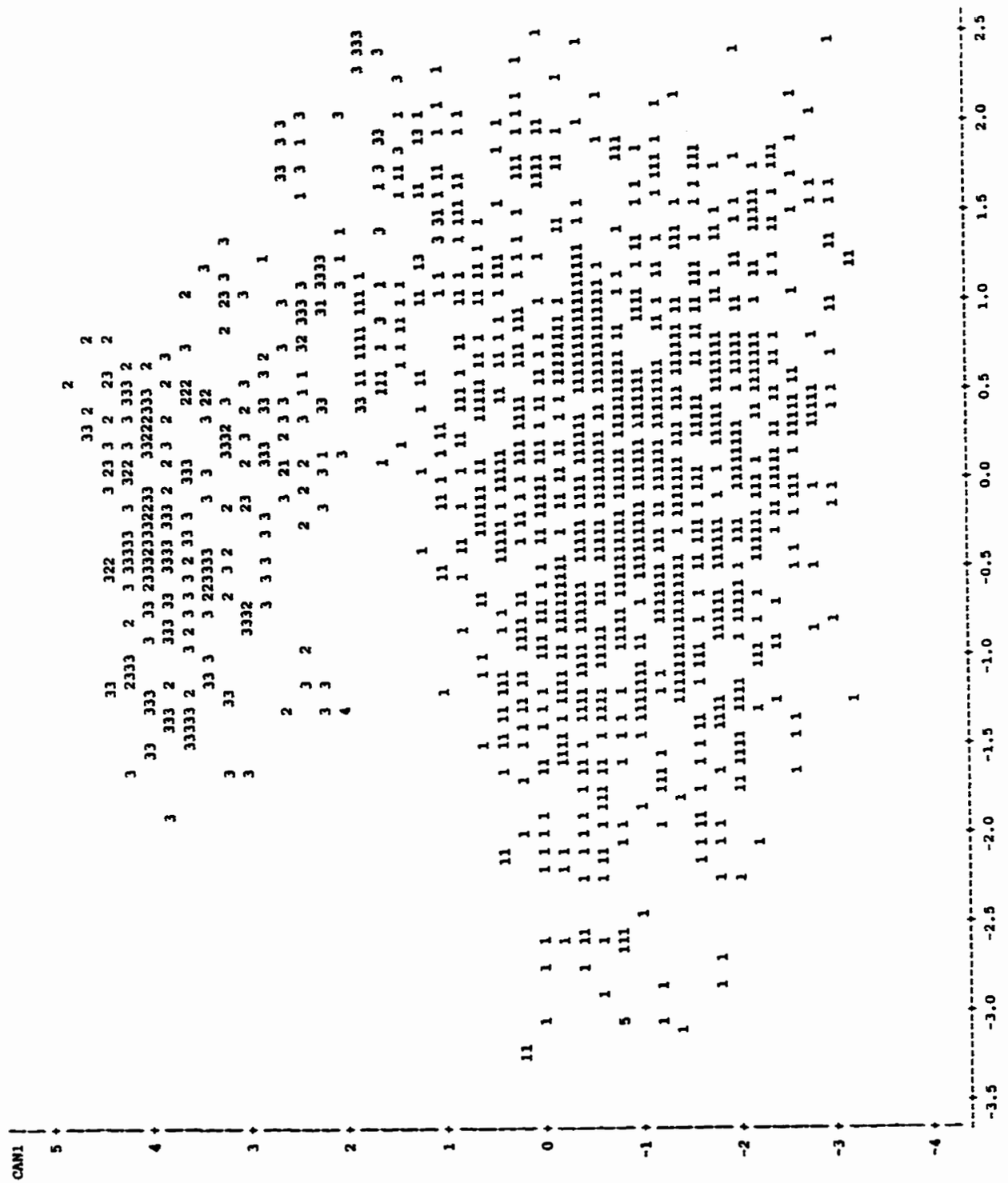


Figure 30. Projection des individus repérés par le numéro de leur famille dans l'espace des variables canoniques CAN1 et CAN3 pour la classification par CENTROID.

individus dans chaque famille. Les figures précedentes de 19 à 30 présentent la projection des individus dans les trois premières variables canonique (sur 7 variables canoniques au total) pour chacune des méthodes de classification utilisées. Nous constatons sur ces graphiques que les familles trouvées par les quatre méthodes de classification apparaissent bien séparées lorsqu'elles sont projetées dans l'espace des variables canoniques. La meilleure séparation est obtenue par la projection sur les variables canoniques #1 et #2 (CAN1 et CAN2), la seconde meilleure représentation en projetant sur CAN1 et CAN3, et la troisième meilleure représentation des familles est obtenue en projetant les individus sur CAN2 et CAN3. Les méthodes AVERAGE (Figure 22, 23 et 24) et WARD (Figure 25, 26 et 27) sont particulièrement performantes puisque nous observons des nuages d'individus distincts pour chaque famille. Les méthodes FASTCLUS (Figure 19, 20 et 21) et CENTROID (Figure 28, 29 et 30) séparent un peu moins bien les 5 familles ce qui apparaît sur les figures par un mélange d'individus appartenant à différentes familles et ce plus particulièrement à la périphérie des nuages d'individus. Il faut néanmoins noter que nous projetons ici des nuages de points dans deux dimensions alors qu'ils sont en réalité placés dans un hyperespace à 7 dimensions. Cela entraîne des déformations et superpositions éventuelles des nuages de points, c'est pourquoi certaines familles peuvent apparaître imbriquées selon les axes de projection choisis, alors qu'elles ne le sont pas en réalité.

Après avoir classé nos individus en famille, nous avons recherché les caractéristiques conformationnelles initiales de ces individus c'est à dire l'énergie conformationnelle, les angles de torsion et les codes conformationnels correspondants (voir description des codes conformationnels page 31). Nous présentons ces données pour les 10 premiers individus de chaque famille pour les quatre méthodes de classification dans les tableaux suivants. Nous présentons également les moyennes des distances calculées pour chaque famille. Nous pourrons ainsi corréler les observations sur les codes conformationnels avec les distances caractéristiques dans chaque famille.

Tableau 7. Caractéristiques conformationnelles des 10 premiers individus des 5 familles trouvées par la méthode FASTCLUS. Energies en kcal/mol.

			Ala			Pro		Tyr				Ala			
#fam.	#conf.	Energie	PHI	PSI	KHI	PHI	PSI	PHI	PSI	KHI1	KHI2	PHI	PSI	KHI	Code
1	5	-18.19	-152	75	57	-75	73	-146	162	-59	108	-79	75	61	D C E C
1	7	-17.79	-152	76	177	-75	74	-156	153	175	74	-79	77	-59	D C E C
1	12	-17.59	-152	76	178	-75	-17	-155	153	175	74	-79	77	60	D A E C
1	16	-17.46	-153	75	177	-75	74	-146	162	-60	-71	-72	-35	61	D C E A
1	19	-17.38	-152	74	57	-75	74	-146	160	-60	109	-154	159	59	D C E E
1	28	-16.97	-153	74	-62	-75	74	-146	161	-59	108	-149	45	61	D C E D
1	33	-16.78	-152	76	-62	-75	-12	-153	175	78	102	-68	95	-60	D A E C
1	34	-16.74	-152	77	-62	-75	-17	-155	158	174	79	-146	154	-179	D A E E
1	35	-16.65	-152	76	-62	-75	-17	-154	153	177	70	-151	83	-62	D A E D
1	42	-16.42	-151	76	179	-75	-9	-139	155	61	-93	-76	91	-59	D B E C
2	36	-16.58	-154	76	57	-75	165	-156	154	176	-104	-78	77	60	D F E C
2	39	-16.45	-154	76	177	-75	162	-156	153	176	74	-79	78	60	D F E C
2	77	-15.74	-154	76	58	-75	74	-76	157	180	-99	-147	156	179	D C F E
2	93	-15.55	-153	76	-62	-75	162	-156	152	177	72	-150	85	57	D F E D
2	97	-15.53	-152	76	-62	-75	161	-156	152	178	69	-150	81	-61	D F E D
2	105	-15.43	-153	77	-62	-75	163	-157	158	177	-101	-148	156	179	D F E E
2	113	-15.35	-152	76	178	-75	169	-64	-32	179	80	-70	-31	-178	D F A A
2	135	-15.10	-152	76	178	-75	168	-61	-40	177	78	-84	71	-58	D F A C
2	155	-14.84	-152	76	179	-75	176	-62	-26	70	85	-72	-31	-177	D F A A
2	161	-14.71	-151	76	57	-75	177	-62	-26	70	-95	-72	-31	61	D F A A
3	1	-19.74	-153	76	178	-75	-20	-76	-22	70	-97	-87	-42	-56	D A A A
3	2	-19.52	-151	77	59	-75	-17	-77	-32	-56	-67	-87	-41	-177	D A A A
3	3	-18.71	-152	76	179	-75	-21	-71	-36	-180	-97	-85	-40	63	D A A A
3	4	-18.48	-151	76	58	-75	-14	-86	-10	-53	-67	-145	42	-58	D A B D
3	8	-17.74	53	74	64	-75	-19	-78	-22	74	84	-92	-45	-56	A*A A A
3	9	-17.64	-151	77	-59	-75	9	-100	23	-50	112	-155	-56	178	D B B G
3	10	-17.63	-151	77	-178	-75	-9	-99	21	-50	112	-152	-56	-61	D B B G
3	11	-17.63	53	74	64	-75	-16	-78	-30	-56	-68	-89	-42	-57	A*A A A
3	13	-17.54	52	74	-176	-75	-19	-77	-22	73	-95	-93	-45	-178	A*A A A
3	14	-17.53	52	74	-175	-75	-19	-78	-22	72	-95	-92	-46	-178	A*A A A
4	40	-16.44	-152	76	57	-75	164	-156	162	57	93	-77	82	-179	D F E C
4	65	-15.86	-152	77	58	-75	164	-143	159	-60	113	-154	159	-60	D F E E
4	66	-15.85	-153	76	-61	-75	165	-143	160	-59	113	-154	159	59	D F E E
4	73	-15.78	-153	77	57	-75	156	-135	29	-55	116	-157	158	59	D F D E
4	80	-15.73	-153	76	57	-75	163	-142	159	-59	-65	-155	159	59	D F E E
4	117	-15.31	-152	76	177	-75	168	-158	157	51	-93	-79	-41	-58	D F E A
4	121	-15.28	-152	76	177	-75	166	-156	157	52	87	-80	-42	-178	D F E A
4	126	-15.20	-152	76	178	-75	159	-135	29	-54	-64	-152	45	-59	D F D D
4	144	-15.00	-152	76	-62	-75	163	-159	165	60	-89	-154	158	-59	D F E E
4	151	-14.89	-152	77	-61	-75	165	-143	158	-60	112	-157	-57	-65	D F E G
5	6	-17.94	-153	74	177	-75	71	-145	32	-57	110	-80	75	-179	D C D C
5	17	-17.42	-152	77	180	-75	83	-64	-28	-62	115	-69	-31	-58	D C A A
5	18	-17.41	-152	77	179	-75	82	-64	-28	-62	-64	-69	-31	-179	D C A A
5	30	-16.88	-152	77	180	-75	82	-64	-31	-62	-64	-143	46	-58	D C A D
5	38	-16.56	-154	74	58	-75	70	-143	31	-58	-69	-158	-57	-65	D C D G
5	50	-16.06	-152	77	59	-75	77	-58	110	-64	115	55	39	-52	D C C A*
5	53	-15.96	-153	76	-61	-75	73	-68	-35	-180	-99	-82	72	60	D C A C
5	55	-15.95	-152	76	58	-75	73	-66	-37	178	-98	-81	73	-59	D C A C
5	64	-15.86	-153	76	57	-75	75	-76	103	-180	70	-73	-33	-178	D C C A
5	69	-15.84	-157	155	164	-75	77	-68	-27	-58	-62	-71	-30	61	E C A A

Tableau 8. Statistiques élémentaires sur les 18 distances à l'intérieur de chaque famille pour la classification par la méthode hiérarchique FASTCLUS.

Cluster=1

Variable	N	Mean	Std Dev	Minim.	Maximum
D1	555	2.81	0.17	2.62	3.12
D4	555	3.10	0.45	2.67	4.31
D5	555	5.38	0.80	3.28	6.92
D6	555	7.31	1.23	3.84	10.35
D8	555	4.69	0.40	2.36	5.14
D9	555	7.23	0.81	4.09	8.66
D10	555	6.13	1.13	3.56	8.70
D11	555	3.93	0.81	2.48	5.11
D13	555	9.08	1.37	3.84	11.28
D16	555	7.94	2.32	3.80	12.08
D17	555	8.17	1.92	3.53	12.22

Cluster=2

Variable	N	Mean	Std Dev	Minimum	Maximum
D1	230	3.59	0.17	2.65	3.65
D4	230	4.02	0.15	2.76	4.26
D5	230	6.54	0.50	6.65	7.46
D6	230	8.32	1.34	5.36	10.79
D8	230	3.97	0.68	2.36	5.00
D9	230	5.87	1.46	2.68	8.57
D10	230	8.39	1.12	5.25	10.07
D11	230	3.91	0.83	2.67	5.30
D13	230	10.22	0.86	6.58	11.36
D16	230	7.30	1.88	3.81	11.58
D17	230	10.25	1.75	5.73	13.27

Cluster=3

Variable	N	Mean	Std Dev	Minimum	Maximum
D1	473	1.85	0.38	2.62	3.65
D4	473	3.46	0.46	2.67	4.34
D5	473	4.84	0.66	3.28	6.31
D6	473	5.02	1.31	2.71	7.83
D8	473	3.39	0.60	2.36	4.97
D9	473	5.09	1.22	2.60	7.77
D10	473	7.21	0.91	4.08	8.75
D11	473	3.80	0.79	2.67	5.05
D13	473	9.49	1.39	6.06	11.53
D16	473	9.53	1.70	4.01	12.18
D17	473	8.72	1.45	2.67	9.72

Cluster=4

Variable	N	Mean	Std Dev	Minimum	Maximum
D1	204	3.59	0.10	2.91	3.65
D4	204	3.95	0.16	3.20	4.28
D5	204	6.89	0.47	4.97	7.51
D6	204	9.02	0.90	6.75	10.96
D8	204	4.66	0.39	2.87	5.05
D9	204	6.95	0.69	4.30	8.62
D10	204	9.21	0.57	7.85	10.12
D11	204	3.84	0.84	2.50	5.02
D13	204	5.52	1.34	3.34	9.84
D16	204	9.42	1.91	4.49	12.15
D17	204	11.45	1.06	7.87	13.44

Cluster=5

Variable	N	Mean	Std Dev	Minimum	Maximum
D1	537	2.88	0.19	2.62	3.64
D4	537	3.10	0.41	2.67	4.28
D5	537	5.78	0.61	3.86	6.90
D6	537	7.70	1.02	4.98	9.97
D8	537	3.28	0.50	2.36	4.96
D9	537	4.85	1.25	2.58	7.30
D10	537	7.85	0.99	4.42	9.27
D11	537	3.69	0.78	2.67	5.02
D13	537	7.26	1.93	3.51	10.38
D16	537	9.30	1.89	4.06	12.17
D17	537	10.06	1.21	5.90	12.44

Dans chacune des cinq familles trouvées par la méthode FASTCLUS, nous pouvons décrire à l'aide des tableaux 7 (où nous présentons les angles ϕ, ψ et χ de quelques molécules ainsi que les codes conformationnels pour la chaîne principale) et 8 (où nous présentons la moyenne de chaque distance caractéristique calculée sur l'ensemble des conformations de la famille) les particularités structurelles de chaque famille. Pour chacune des 5 familles, nous observons une caractéristique conformationnelle typique soit dans la première famille, les séquences -CE- et -AE- soit un tournant γ sur le résidu Pro suivi d'un résidu Tyr étendu. La deuxième famille présente les séquences -FE- et -FA- soit une conformation étendue pour les résidus centraux. Dans le cas de la troisième famille, la séquence est -AA-, ce qui traduit la présence d'un tournant β sur les résidus centraux. La quatrième famille est caractérisée par la séquence DFE- soit une molécule pour laquelle trois des quatre résidus ont une conformation étendue. La dernière famille est caractérisée par une séquence -CA- soit un tournant γ sur le résidu Pro. L'examen de ces séquences caractéristiques montre que l'organisation conformationnelle se fait autour des deux résidus centraux soit Pro et Tyr en particulier le résidu Pro, contraint conformationnellement et qui induit les structures particulières en tournant β. Si nous regardons les moyennes de distances calculées à l'intérieur de chaque famille, nous pouvons recouper les observations structurales faites sur les codes. En effet, dans la première famille caractérisée par un tournant γ sur le résidu Pro, la distance correspondante D4 est courte (moyenne 3.30Å) de même pour la cinquième famille caractérisée par ce même tournant et donc cette même courte distance (moyenne 3.10Å). Ces deux familles diffèrent par les moyennes des distances D8 et D9 qui indiquent une conformation des résidus Tyr et Ala final étendue dans la première famille (moyenne 4.70Å et 7.23Å) et repliée dans la cinquième famille avec des moyennes de 3.28Å et 4.85Å. Les familles deux et quatre présentent une distance D5 longue ce qui traduit la conformation étendue des résidus Pro et Tyr. La quatrième famille se distingue par la présence d'une moyenne de distance courte pour D13, ce qui traduit une interaction de la chaîne latérale de la Tyr avec le premier résidu Ala alors que dans toutes les autres familles cette chaîne latérale ne présente pas d'interaction avec la chaîne principale du peptide. De manière générale, la

quatrième famille présente des moyennes élevées pour toutes ses autres distances ce qui traduit une conformation très étendue du peptide dans cette famille. La troisième famille présente une distance courte pour D5 ce qui traduit le tournant β sur les résidus centraux Pro et Tyr et toutes les autres distances, relativement courtes traduisent une structure plutôt compacte pour l'ensemble de la molécule à l'exception de la chaîne latérale du résidu Tyr.

En observant les résultats obtenus pour les 5 familles trouvées dans l'échantillon par la classification par l'option AVERAGE présentés dans les tableaux 10 et 11, nous pouvons, de la même manière que pour les résultats obtenus par l'option FASTCLUS, retrouver les caractéristiques structurelles propres à chaque famille. Nous constatons alors des recoupements avec les résultats de l'option FASTCLUS. En effet, les familles obtenues sont globalement les mêmes et donc les caractéristiques structurales propres à chaque famille sont décrites par les mêmes codes et les mêmes distances caractéristiques (voir description des 5 familles page 73). La même observation peut être faite en ce qui concerne les résultats de l'option WARD (tableaux 12 et 13), qui confirment la classification obtenue par les méthodes FASTCLUS et AVERAGE. La différence entre les résultats de la classification est le numéro attribué à chaque famille donc des familles identiques conformationnellement peuvent être classées dans des familles de numéros différents selon l'option de classification utilisée c'est pourquoi nous présentons le tableau 9 où sont noté les familles identiques.

Tableau 9. Equivalence entre les familles trouvées par chaque option de classification.

FASTCLUS	AVERAGE	WARD	CENTROID
1	2	2/1	
2	4/3	4/3	3
3	1	1	1
4	3	3/2	2
5	5	5	

Tableau 10. Caractéristiques conformationnelles des 10 premiers individus des 5 familles trouvées par la méthode AVERAGE. Energies en kcal/mol.

			Ala			Pro		Tyr				Ala			
#fam.	#conf.	Energie	PHI	PSI	KHI	PHI	PSI	PHI	PSI	KHI1	KHI2	PHI	PSI	KHI	Code
1	1	-19.74	-153	76	178	-75	-20	-76	-22	70	-97	-87	-42	-56	D A A A
1	2	-19.52	-151	77	59	-75	-17	-77	-32	-56	-67	-87	-41	-177	D A A A
1	3	-18.71	-152	76	179	-75	-21	-71	-36	-180	-97	-85	-40	63	D A A A
1	4	-18.48	-151	76	58	-75	-14	-86	-10	-53	-67	-145	42	-58	D A B D
1	8	-17.74	53	74	64	-75	-19	-78	-22	74	84	-92	-45	-56	A*A A A
1	9	-17.64	-151	77	-59	-75	-9	-100	23	-50	112	-155	-56	178	D B B G
1	10	-17.63	-151	77	-178	-75	-9	-99	21	-50	112	-152	-56	-61	D B B G
1	11	-17.63	53	74	64	-75	-16	-78	-30	-56	-68	-89	-42	-57	A*A A A
1	13	-17.54	52	74	-176	-75	-19	-77	-22	73	-95	-93	-45	-178	A*A A A
1	14	-17.53	52	74	-175	-75	-19	-78	-22	72	-95	-92	-46	-178	A*A A A
2	5	-18.19	-152	75	57	-75	73	-146	162	-59	108	-79	75	61	D C E C
2	6	-17.94	-153	74	177	-75	71	-145	32	-57	110	-80	75	-179	D C D C
2	7	-17.79	-152	76	177	-75	74	-156	153	175	74	-79	77	-59	D C E C
2	12	-17.59	-152	76	178	-75	-17	-155	153	175	74	-79	77	60	D A E C
2	16	-17.46	-153	75	177	-75	74	-146	162	-60	-71	-72	-35	61	D C E A
2	19	-17.38	-152	74	57	-75	74	-146	160	-60	109	-154	159	59	D C E E
2	26	-17.05	-153	76	178	-75	68	-152	-53	-65	105	-152	37	61	D C G D
2	28	-16.97	-153	74	-62	-75	74	-146	161	-59	108	-149	45	61	D C E D
2	29	-16.90	-153	76	-63	-75	70	-166	-52	167	66	-153	38	-178	D C G D
2	33	-16.78	-152	76	-62	-75	-12	-153	175	78	102	-68	95	-60	D A E C
3	36	-16.58	-154	76	57	-75	165	-156	154	176	-104	-78	77	60	D F E C
3	39	-16.45	-154	76	177	-75	162	-156	153	176	74	-79	78	60	D F E C
3	40	-16.44	-152	76	57	-75	164	-156	162	57	93	-77	82	-179	D F E C
3	65	-15.86	-152	77	58	-75	164	-143	159	-60	113	-154	159	-60	D F E E
3	66	-15.85	-153	76	-61	-75	165	-143	160	-59	113	-154	159	59	D F E E
3	73	-15.78	-153	77	57	-75	156	-135	29	-55	116	-157	158	59	D F D E
3	80	-15.73	-153	76	57	-75	163	-142	159	-59	-65	-155	159	59	D F E E
3	93	-15.55	-153	76	-62	-75	162	-156	152	177	72	-150	85	57	D F E D
3	97	-15.53	-152	76	-62	-75	161	-156	152	178	69	-150	81	-61	D F E D
3	105	-15.43	-153	77	-62	-75	163	-157	158	177	-101	-148	156	179	D F E E
4	113	-15.35	-152	76	178	-75	169	-64	-32	179	80	-70	-31	-178	D F A A
4	135	-15.10	-152	76	178	-75	168	-61	-40	177	78	-84	71	-58	D F A C
4	155	-14.84	-152	76	179	-75	176	-62	-26	70	85	-72	-31	-177	D F A A
4	161	-14.71	-151	76	57	-75	177	-62	-26	70	-95	-72	-31	61	D F A A
4	169	-14.55	-152	76	58	-75	166	-59	122	178	78	55	37	68	D F C A*
4	186	-14.34	52	74	-177	-75	147	49	40	-48	120	-160	-57	53	A*F A*G
4	187	-14.34	52	74	63	-75	147	49	39	-49	119	-159	-56	-65	A*F A*G
4	301	-13.25	52	73	62	-75	163	-65	-33	179	-97	-70	-32	-179	A*F A A
4	318	-13.06	52	73	-56	-75	163	-62	-39	177	78	-83	71	-58	A*F A C
4	337	-12.93	52	74	-56	-75	146	49	40	-165	-114	-160	-56	-65	A*F A*G
5	17	-17.42	-152	77	180	-75	83	-64	-28	-62	115	-69	-31	-58	D C A A
5	18	-17.41	-152	77	179	-75	82	-64	-28	-62	-64	-69	-31	-179	D C A A
5	30	-16.88	-152	77	180	-75	82	-64	-31	-62	-64	-143	46	-58	D C A D
5	38	-16.56	-154	74	58	-75	70	-143	31	-58	-69	-158	-57	-65	D C D G
5	50	-16.06	-152	77	59	-75	77	-58	110	-64	115	55	39	-52	D C C A*
5	53	-15.96	-153	76	-61	-75	73	-68	-35	-180	-99	-82	72	60	D C A C
5	55	-15.95	-152	76	58	-75	73	-66	-37	178	-98	-81	73	-59	D C A C
5	64	-15.86	-153	76	57	-75	75	-76	103	-180	70	-73	-33	-178	D C C A
5	69	-15.84	-157	155	164	-75	77	-68	-27	-58	-62	-71	-30	61	E C A A
5	75	-15.75	52	72	-177	-75	71	-143	32	-57	-68	-154	158	59	A*C D E

Tableau 11. Statistiques élémentaires sur les 18 distances à l'intérieur de chaque famille pour la classification par la méthode non-hiérarchique AVERAGE.

CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
D1	162	2.67	0.07	2.62	2.93
D4	162	3.66	0.21	3.05	4.07
D5	162	3.64	0.58	3.18	5.72
D6	162	4.27	1.17	2.71	6.66
D8	162	3.60	0.33	2.74	4.64
D9	162	5.26	1.15	2.92	7.80
D10	162	7.11	0.91	4.08	8.32
D11	162	3.76	0.73	2.48	6.97
D13	162	9.19	1.67	6.36	11.24
D16	162	8.97	2.17	3.83	12.08
D17	162	6.56	1.39	2.88	8.45

CLUSTER=2

Variable	N	Mean	Std Dev	Minimum	Maximum
D1	628	2.81	0.17	2.62	3.34
D4	628	3.31	0.46	2.67	4.31
D5	628	5.34	0.88	3.25	6.92
D6	628	7.19	1.50	2.89	10.35
D8	628	4.67	0.40	3.00	5.14
D9	628	7.17	0.77	4.97	8.66
D10	628	6.38	1.18	3.56	8.75
D11	628	3.98	0.81	2.48	5.11
D13	628	9.01	1.36	3.84	11.28
D16	628	8.13	2.30	3.80	12.08
D17	628	8.22	2.07	2.67	12.44

CLUSTER=3

Variable	N	Mean	Std Dev	Minimum	Maximum
D1	366	3.61	0.04	3.05	3.65
D4	366	3.98	0.13	3.26	4.26
D5	366	6.78	0.49	5.54	7.51
D6	366	8.91	0.95	6.54	10.96
D8	366	4.43	0.58	2.73	5.05
D9	366	6.74	0.91	3.64	8.62
D10	366	8.91	0.95	5.25	10.12
D11	366	3.94	0.83	2.48	5.02
D13	366	7.73	2.60	3.34	11.34
D16	366	8.29	2.20	3.81	12.15
D17	366	11.10	1.46	5.71	13.44

CLUSTER=4

Variable	N	Mean	Std Dev	Minimum	Maximum
D1	188	3.59	0.08	3.05	3.65
D4	188	3.96	0.23	3.15	4.34
D5	188	6.62	1.09	3.51	8.60
D6	188	9.41	1.21	2.81	7.34
D8	188	3.08	0.38	2.36	4.08
D9	188	4.41	1.18	2.60	6.25
D10	188	7.54	0.78	5.94	9.09
D11	188	3.93	0.72	2.47	5.01
D13	188	10.22	1.24	6.34	11.53
D16	188	9.32	1.67	3.99	11.98
D17	188	7.46	1.51	3.57	9.92

CLUSTER=5

Variable	N	Mean	Std Dev	Minimum	Maximum
D1	655	2.92	0.21	2.62	3.64
D4	655	3.08	0.41	2.67	4.28
D5	655	5.52	0.77	3.72	6.90
D6	655	7.38	1.11	3.97	9.97
D8	655	3.24	0.49	2.36	4.60
D9	655	4.80	1.21	2.58	7.53
D10	655	7.63	1.16	4.42	9.51
D11	655	3.71	0.79	2.47	5.03
D13	655	7.64	2.02	3.51	11.07
D16	655	9.41	1.81	4.04	12.18
D17	655	9.34	1.79	2.92	12.33

Tableau 12. Caractéristiques conformationnelles des 10 premiers individus des 5 familles trouvées par la méthode WARD. Energies en kcal/mol.

			Ala			Pro		Tyr				Ala			
#fam.	#conf.	Energie	PHI	PSI	KHI	PHI	PSI	PHI	PSI	KHI1	KHI2	PHI	PSI	KHI	Code
1	1	-19.74	-153	76	178	-75	-20	-76	-22	70	-97	-87	-42	-56	D A A A
1	2	-19.52	-151	77	59	-75	-17	-77	-32	-56	-67	-87	-41	-177	D A A A
1	3	-18.71	-152	76	179	-75	-21	-71	-36	-180	-97	-85	-40	63	D A A A
1	4	-18.48	-151	76	58	-75	-14	-86	-10	-53	-67	-145	42	-58	D A B D
1	5	-18.19	-152	75	57	-75	73	-146	162	-59	108	-79	75	61	D C E C
1	6	-17.94	-153	74	177	-75	71	-145	32	-57	110	-80	75	-179	D C D C
1	7	-17.79	-152	76	177	-75	74	-156	153	175	74	-79	77	-59	D C E C
1	8	-17.74	53	74	64	-75	-19	-78	-22	74	84	-92	-45	-56	A*A A A
1	9	-17.64	-151	77	-59	-75	-9	-100	23	-50	112	-155	-56	178	D B B G
1	10	-17.63	-151	77	-178	-75	-9	-99	21	-50	112	-152	-56	-61	D B B G
2	12	-17.59	-152	76	178	-75	-17	-155	153	175	74	-79	77	60	D A E C
2	19	-17.38	-152	74	57	-75	74	-146	160	-60	109	-154	159	59	D C E E
2	28	-16.97	-153	74	-62	-75	74	-146	161	-59	108	-149	45	61	D C E D
2	33	-16.78	-152	76	-62	-75	-12	-153	175	78	102	-68	95	-60	D A E C
2	34	-16.74	-152	77	-62	-75	-17	-155	158	174	79	-146	154	-179	D A E E
2	35	-16.65	-152	76	-62	-75	-17	-154	153	177	70	-151	83	-62	D A E D
2	42	-16.42	-151	76	179	-75	-9	-139	155	61	-93	-76	91	-59	D B E C
2	44	-16.32	-152	75	-62	-75	73	-146	158	-60	109	-158	-57	53	D C E G
2	45	-16.20	-152	76	58	-75	-21	-99	150	178	74	-79	78	60	D A F C
2	51	-15.98	-153	76	177	-75	-17	-154	159	177	-100	-147	-58	55	D A E G
3	36	-16.58	-154	76	57	-75	165	-156	154	176	-104	-78	77	60	D F E C
3	39	-16.45	-154	76	177	-75	162	-156	153	176	74	-79	78	60	D F E C
3	40	-16.44	-152	76	57	-75	164	-156	162	57	93	-77	82	-179	D F E C
3	65	-15.86	-152	77	58	-75	164	-143	159	-60	113	-154	159	-60	D F E E
3	66	-15.85	-153	76	-61	-75	165	-143	160	-59	113	-154	159	59	D F E E
3	73	-15.78	-153	77	57	-75	156	-135	29	-55	116	-157	158	59	D F D E
3	80	-15.73	-153	76	57	-75	163	-142	159	-59	-65	-155	159	59	D F E E
3	93	-15.55	-153	76	-62	-75	162	-156	152	177	72	-150	85	57	D F E D
3	97	-15.53	-152	76	-62	-75	161	-156	152	178	69	-150	81	-61	D F E D
3	105	-15.43	-153	77	-62	-75	163	-157	158	177	-101	-148	156	179	D F E E
4	113	-15.35	-152	76	178	-75	169	-64	-32	179	80	-70	-31	-178	D F A A
4	135	-15.10	-152	76	178	-75	168	-61	-40	177	78	-84	71	-58	D F A C
4	155	-14.84	-152	76	179	-75	176	-62	-26	70	85	-72	-31	-177	D F A A
4	161	-14.71	-151	76	57	-75	177	-62	-26	70	-95	-72	-31	61	D F A A
4	169	-14.55	-152	76	58	-75	166	-59	122	178	78	55	37	68	D F C A*
4	186	-14.34	52	74	-177	-75	147	49	40	-48	120	-160	-57	53	A*F A*G
4	187	-14.34	52	74	63	-75	147	49	39	-49	119	-159	-56	-65	A*F A*G
4	301	-13.25	52	73	62	-75	163	-65	-33	179	-97	-70	-32	-179	A*F A A
4	318	-13.06	52	73	-56	-75	163	-62	-39	177	78	-83	71	-58	A*F A C
4	337	-12.93	52	74	-56	-75	146	49	40	-165	-114	-160	-56	-65	A*F A*G
5	17	-17.42	-152	77	180	-75	83	-64	-28	-62	115	-69	-31	-58	D C A A
5	18	-17.41	-152	77	179	-75	82	-64	-28	-62	-64	-69	-31	-179	D C A A
5	30	-16.88	-152	77	180	-75	82	-64	-31	-62	-64	-143	46	-58	D C A D
5	50	-16.06	-152	77	59	-75	77	-58	110	-64	115	55	39	-52	D C C A*
5	53	-15.96	-153	76	-61	-75	73	-68	-35	-180	-99	-82	72	60	D C A C
5	55	-15.95	-152	76	58	-75	73	-66	-37	178	-98	-81	73	-59	D C A C
5	64	-15.86	-153	76	57	-75	75	-76	103	-180	70	-73	-33	-178	D C C A
5	69	-15.84	-157	155	164	-75	77	-68	-27	-58	-62	-71	-30	61	E C A A
5	79	-15.73	-153	77	58	-75	73	-65	-30	179	75	-141	43	61	D C A D
5	83	-15.72	-152	76	-62	-75	73	-66	-29	179	75	-141	42	-58	D C A D

Tableau 13. Statistiques élémentaires sur les 18 distances à l'intérieur de chaque famille pour la classification par la méthode non-hiérarchique WARD.

CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
D1	536	2.82	0.18	2.62	3.54
D4	536	3.27	0.44	2.67	4.28
D5	536	4.39	0.92	3.18	6.90
D6	536	5.51	1.37	2.71	9.21
D8	536	4.13	0.75	2.39	5.14
D9	536	6.19	1.26	2.70	8.41
D10	536	6.97	1.04	3.59	8.75
D11	536	3.81	0.77	2.48	5.11
D13	536	8.98	1.36	3.84	11.25
D16	536	9.14	2.07	1.83	12.18
D17	536	7.03	1.73	2.67	10.97

CLUSTER=2

Variable	N	Mean	Std Dev	Minimum	Maximum
D1	351	2.77	0.17	2.62	3.15
D4	351	3.35	0.40	2.68	4.31
D5	351	5.87	0.48	4.16	6.92
D6	351	8.28	0.90	5.73	10.35
D8	351	4.60	0.48	3.36	5.13
D9	351	7.30	0.78	5.30	8.66
D10	351	6.04	1.09	3.56	8.58
D11	351	4.31	0.70	2.48	5.04
D13	351	8.79	1.62	3.65	11.20
D16	351	7.77	2.16	3.80	12.30
D17	351	9.13	2.01	3.67	12.44

CLUSTER=3

Variable	N	Mean	Std Dev	Minimum	Maximum
D1	174	3.60	0.08	2.91	3.65
D4	174	3.98	0.13	3.26	4.26
D5	174	6.77	0.49	5.54	7.51
D6	174	8.91	0.94	6.54	10.96
D8	174	4.62	0.59	2.73	5.05
D9	174	6.73	0.91	3.64	8.62
D10	174	8.89	0.95	5.25	10.12
D11	174	3.96	0.83	2.48	5.02
D13	174	7.69	2.59	3.34	11.34
D16	174	8.32	2.20	3.81	12.15
D17	174	11.08	1.47	5.73	13.44

CLUSTER=4

Variable	N	Mean	Std Dev	Minimum	Maximum
D1	187	3.59	0.08	3.05	3.65
D4	187	3.96	0.23	3.15	4.34
D5	187	6.62	1.09	3.51	8.66
D6	187	5.41	1.21	2.81	7.34
D8	187	3.08	0.38	2.36	4.08
D9	187	4.42	1.18	2.60	6.25
D10	187	7.54	0.78	5.94	9.06
D11	187	3.52	0.72	2.47	5.01
D13	187	10.22	1.25	6.34	11.53
D16	187	9.31	1.67	3.99	11.96
D17	187	7.45	1.51	3.57	9.92

CLUSTER=5

Variable	N	Mean	Std Dev	Minimum	Maximum
D1	551	2.91	0.20	2.62	3.64
D4	551	3.08	0.41	2.67	4.28
D5	551	5.62	0.67	3.44	6.90
D6	551	7.48	1.30	4.26	9.64
D8	551	3.26	0.48	2.36	4.12
D9	551	4.67	1.20	2.50	6.84
D10	551	7.70	1.20	4.42	9.31
D11	551	3.34	0.75	2.47	5.02
D13	551	7.65	1.90	3.51	11.07
D16	551	9.13	1.91	4.04	12.17
D17	551	9.61	1.48	4.05	12.13

Quant aux résultats obtenus par l'option de classification CENTROID, ils sont moins concluants. En effet, si nous imposons un nombre de famille final de 5, le résultat de la classification donne deux familles qui ne contiennent chacune qu'un seul individu. Le choix de 4 familles finales pourtant proposé par l'étude des indices statistiques Pseudo F et Pseudo t^2 conduit à un résultat similaire c-à-d qu'une famille sur les quatre contient un seul individu. C'est pourquoi nous présentons les résultats pour les trois familles réellement trouvées par cette option de classification dans les tableaux 14 et 15.

Tableau 14. Caractéristiques conformationnelles des 10 premiers individus des 5 familles trouvées par la méthode CENTROID. Energies en kcal/mol.

			Ala			Pro		Tyr				Ala			
#fam.	#conf.	Energie	PHI	PSI	KHI	PHI	PSI	PHI	PSI	KHI1	KHI2	PHI	PSI	KHI	Code
1	1	-19.74	-153	76	178	-75	-20	-76	-22	70	-97	-87	-42	-56	D A A A
1	2	-19.52	-151	77	59	-75	-17	-77	-32	-56	-67	-87	-41	-177	D A A A
1	3	-18.71	-152	76	179	-75	-21	-71	-36	-180	-97	-85	-40	63	D A A A
1	4	-18.48	-151	76	58	-75	-14	-86	-10	-53	-67	-145	42	-58	D A B D
1	5	-18.19	-152	75	57	-75	73	-146	162	-59	108	-79	75	61	D C E C
1	6	-17.94	-153	74	177	-75	71	-145	32	-57	110	-80	75	-179	D C D C
1	7	-17.79	-152	76	177	-75	74	-156	153	175	74	-79	77	-59	D C E C
1	8	-17.74	53	74	64	-75	-19	-78	-22	74	84	-92	-45	-56	A*A A A
1	9	-17.64	-151	77	-59	-75	-9	-100	23	-50	112	-155	-56	178	D B B G
1	10	-17.63	-151	77	-178	-75	-9	-99	21	-50	112	-152	-56	-61	D B B G
2	65	-15.86	-152	77	58	-75	164	-143	159	-60	113	-154	159	-60	D F E E
2	66	-15.85	-153	76	-61	-75	165	-143	160	-59	113	-154	159	59	D F E E
2	73	-15.78	-153	77	57	-75	156	-135	29	-55	116	-157	158	59	D F D E
2	80	-15.73	-153	76	57	-75	163	-142	159	-59	-65	-155	159	59	D F E E
2	126	-15.20	-152	76	178	-75	159	-135	29	-54	-64	-152	45	-59	D F D D
2	151	-14.89	-152	77	-61	-75	165	-143	158	-60	112	-157	-57	-65	D F E G
2	152	-14.89	-153	76	-62	-75	165	-143	158	-60	113	-157	-57	174	D F E G
2	160	-14.72	-152	77	58	-75	159	-150	-54	-69	-67	-80	75	61	D F G C
2	175	-14.46	-152	76	178	-75	164	-143	157	-60	113	54	46	-53	D F E A*
2	194	-14.29	-153	77	-62	-75	164	-141	158	-58	-65	54	47	-174	D F E A*
3	36	-16.58	-154	76	57	-75	165	-156	154	176	-104	-78	77	60	D F E C
3	39	-16.45	-154	76	177	-75	162	-156	153	176	74	-79	78	60	D F E C
3	40	-16.44	-152	76	57	-75	164	-156	162	57	93	-77	82	-179	D F E C
3	93	-15.55	-153	76	-62	-75	162	-156	152	177	72	-150	85	57	D F E D
3	97	-15.53	-152	76	-62	-75	161	-156	152	178	69	-150	81	-61	D F E D
3	105	-15.43	-153	77	-62	-75	163	-157	158	177	-101	-148	156	179	D F E E
3	117	-15.31	-152	76	177	-75	168	-158	157	51	-93	-79	-41	-58	D F E A
3	121	-15.28	-152	76	177	-75	166	-156	157	52	87	-80	-42	-178	D F E A
3	144	-15.00	-152	76	-62	-75	163	-159	165	60	-89	-154	158	-59	D F E E
3	162	-14.63	-152	76	177	-75	171	-67	-33	178	77	-153	156	-61	D F A E

Tableau 15. Statistiques élémentaires sur les 18 distances à l'intérieur de chaque famille pour la classification par la méthode non-hiérarchique CENTROID.

CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
D1	1632	2.93	0.10	2.62	3.65
D4	1632	3.31	0.48	2.67	4.34
D5	1632	5.16	1.02	3.18	6.92
D6	1632	6.77	1.65	2.71	10.35
D8	1632	3.80	0.81	2.36	5.14
D9	1632	5.71	1.57	2.58	8.66
D10	1632	7.09	1.25	3.56	9.51
D11	1632	3.80	0.80	2.47	5.11
D13	1632	8.62	1.87	3.51	11.53
D16	1632	8.86	2.12	3.80	12.18
D17	1632	8.41	2.05	2.67	12.44

CLUSTER=2

Variable	N	Mean	Std Dev	Minimum	Maximum
D1	94	3.59	0.05	3.37	3.65
D4	94	3.68	0.19	3.26	4.13
D5	94	6.87	0.50	5.54	7.51
D6	94	9.13	0.85	6.75	10.96
D8	94	4.67	0.15	3.76	5.00
D9	94	6.95	0.89	4.67	8.50
D10	94	9.21	0.52	7.96	10.03
D11	94	3.77	0.88	2.50	5.01
D13	94	6.21	0.65	3.16	6.45
D16	94	11.00	0.69	9.55	12.15
D17	94	11.62	1.01	9.44	13.44

CLUSTER=3

Variable	N	Mean	Std Dev	Minimum	Maximum
D1	272	3.62	0.03	3.05	3.65
D4	272	4.01	0.08	3.55	4.26
D5	272	6.75	0.49	5.77	7.49
D6	272	8.83	0.97	6.54	10.79
D8	272	4.35	0.63	2.73	5.05
D9	272	6.66	0.91	3.64	8.62
D10	272	8.81	1.04	5.25	10.12
D11	272	4.00	0.80	2.48	5.02
D13	272	8.95	1.78	6.43	11.34
D16	272	7.16	1.72	3.81	11.69
D17	272	10.93	1.56	5.73	13.27

En observant les premiers individus placés respectivement dans les familles 2 et 3, pour la classification par l'option CENTROID, la discrimination entre les structures caractéristiques pour ces deux familles n'apparaît pas évidente. En effet, les codes conformationnels des 10 premiers individus, présentés dans le tableau 14 sont peu différents. En revanche, l'examen des moyennes de distances par famille présentées dans le tableau 15 permet de constater que ces familles se distinguent par les interactions entre la chaîne latérale de Tyr et la chaîne principale du peptide. Il est donc normal que cette caractéristique n'apparaisse pas à l'étude

des codes conformationnels puisque ces derniers codent uniquement la structure de la chaîne principale du peptide. Pour la deuxième famille, la chaîne latérale est repliée sur le début de la chaîne principale du peptide c'est à dire le résidu Ala (distance D13 courte: moyenne de 4.22Å et D16 longue: moyenne de 11.00Å) alors que pour la famille trois, c'est l'inverse: la distance D16 est courte (moyenne de 7.36Å) par rapport à la distance D13 longue (dans une moindre mesure: moyenne 8.95Å) indique un repliement de la chaîne latérale du peptide sur la fin de la chaîne principale du peptide c'est à dire le résidu terminal Ala.

Nous présentons maintenant les résultats graphiques: nous avons superposé les 10 premiers individus de chaque famille pour la classification par l'option WARD. Sur chaque superposition, nous voyons apparaître les éléments structuraux caractéristiques de chaque famille.

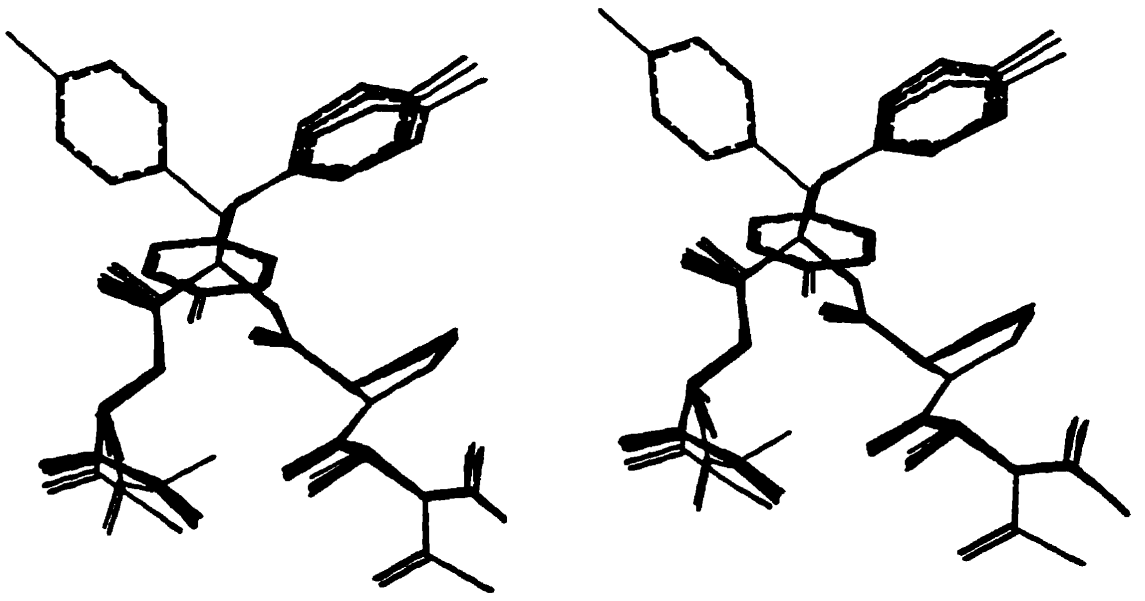


Figure 31. Superposition des 10 premiers individus de la famille #1 pour la classification par l'option WARD.

Figure 32. Superposition des 10 premiers individus de la famille #2 pour la classification par l'option WARD.

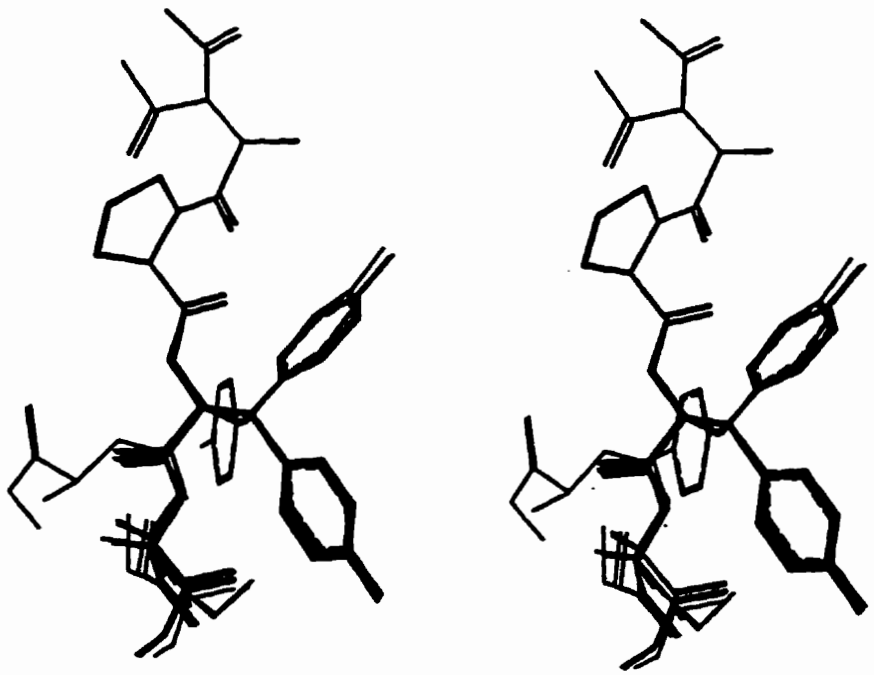


Figure 33. Superposition des 10 premiers individus de la famille #3 pour la classification par l'option WARD.

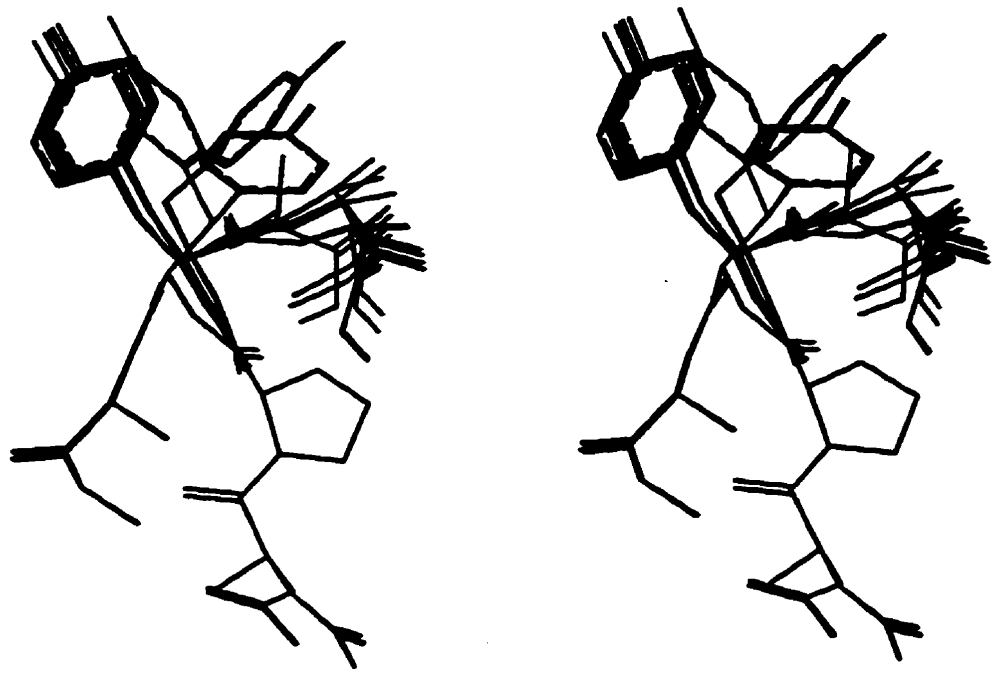


Figure 34. Superposition des 10 premiers individus de la famille #4 pour la classification par l'option WARD.

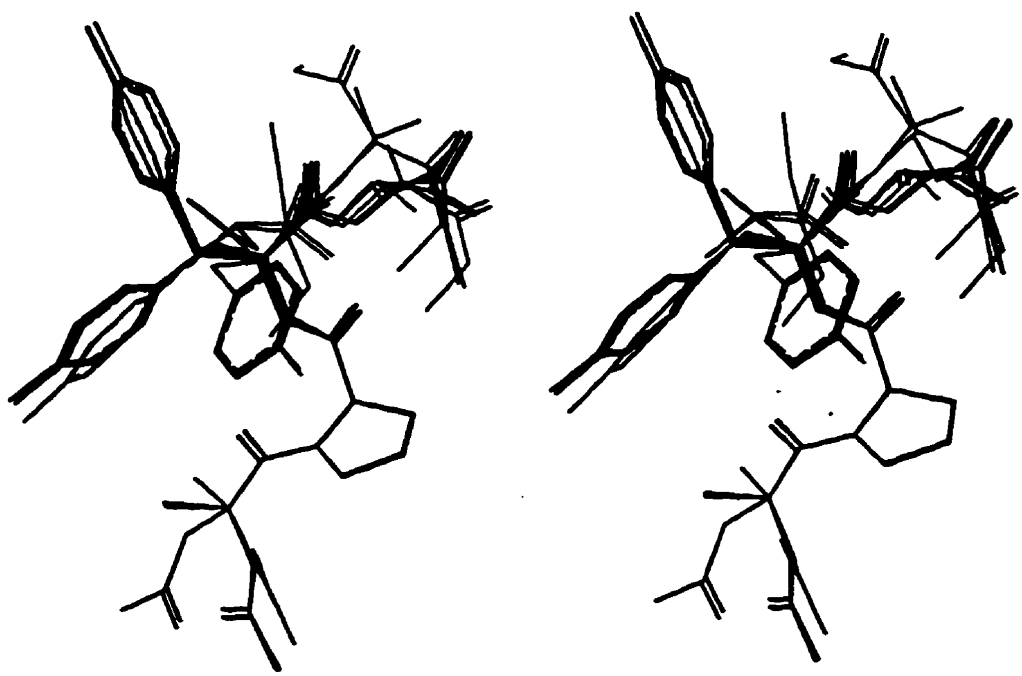


Figure 35. Superposition des 10 premiers individus de la famille #5 pour la classification par l'option WARD.

2.2.3.2.3 Conclusion sur la classification

Nous désirions savoir si les méthodes de classification étaient capables de trier correctement un échantillon de conformation et déterminer ainsi les méthodes les plus utiles à cette fin. Globalement, le résultat est utilisable: les familles trouvées par la classification ont un sens au point de vue structural et le temps de CPU exigé par cette classification est raisonnable ce qui permet d'essayer plusieurs options et différents niveaux de classement. L'étape la plus longue est de définir les caractéristiques structurales qui permettent de faire le lien individu mathématique/molécule (moyennes de distances, superpositions graphiques, codes conformationnels).

Néanmoins, nous avons rencontré un certain nombre de difficultés inhérentes aux méthodes de classification. Dans l'interprétation des indices statistiques qui ne proposent pas le même choix pour le nombre de familles selon les options de classification. De plus, ainsi que nous l'avons constaté pour l'option CENTROID, le nombre de famille trouvé par les indices statistiques ne cadre pas forcément avec le résultat de la classification subséquente.
Tel que prévu par la théorie, les performances des différentes méthodes sont inégales mais consistantes en ce qui concerne les résultats. Pour les méthodes hiérarchiques, tel que prévu, l'option CENTROID est moins performante que les options WARD et AVERAGE puisqu'elle ne découvre que 3 familles dans notre échantillon. Les autres méthodes découvrent sensiblement les mêmes familles mais avec un nombre d'individus affecté à chaque famille qui diffère selon la méthode. Cela montre l'importance d'utiliser conjointement plusieurs méthodes de classification.

Les moyennes des distances à l'intérieur de chaque famille conduisent à deux observations: les résultats de la classification dépendent étroitement du choix de ces distances. Nous avons rejeté au départ les distances pour lesquelles nous n'avons observé aucune valeur compatible avec l'existence d'une liaison hydrogène. Néanmoins, certaines de nos familles sont

constituées par des conformations majoritairement étendues. Ces familles sont celles pour lesquelles les structures sont les plus difficiles à décrire, dû peut-être à l'absence de distances adéquates tel que par exemple D2 qui pourrait caractériser la famille 4 trouvée lors de la classification par FASTCLUS. L'élimination de trop nombreuses distances pourrait donc être un handicap pour la classification de l'échantillon, d'autant plus que l'augmentation du temps de calcul sera imperceptible dans la phase ACP. D'autre part, les frontières entre les familles sont floues ce qui se traduit par un nombre d'individus différent affecté à chaque famille par les différentes options. Plus nous avons de variables pour la classification (7 c.p.) et plus le problème de déterminer les frontières réelles entre chaque famille sera grand. Néanmoins, réduire le nombre de distances initiales n'est pas une solution intéressante. Nous avons en effet déterminé par des études préalables qu'il existe un nombre de variables minimal pour décrire une population. Pour éclaircir ce point, nous avons fait plusieurs ACP successives en réduisant le nombre de distances initiales tout en prenant soin d'éliminer celles qui avaient les coefficients de corrélation les plus élevés avec celles que nous conservions. Nous avons observé que si nous réduisons le nombre de distances initiales, le nombre de composantes principales qui permet de conserver 90% de la variation totale de l'échantillon n'est pas réduit. De plus, si nous réduisons le nombre de distances initiales, nous pouvons perdre une partie de l'information sur les conformères. Nous avons donc observé que le nombre de composantes principales dépend de la complexité intrinsèque de l'échantillon et non du nombre de distances initiales.

En conclusion, suite à cette étude, nous préconisons de ne pas réduire d'emblée le nombre de variables initiales pour l'analyse de données. En revanche, il est nécessaire de réduire le nombre de variables composantes principales utilisées pour la classification subséquente puisque le temps de classification augmente de manière drastique avec le nombre de variables. Néanmoins, il n'est pas possible de réduire le nombre de composantes principales sous un certain seuil puisque ce nombre dépend de la complexité de l'échantillon. Ce travail à été publié dans (61).

CHAPITRE 3

APPLICATION AUX PEPTIDES INHIBITEURS DE LA CHOLÉCYSTOKININE

3.1 Introduction

La cholécystokinine est une hormone présente dans le système gastro-intestinal ainsi que dans le système nerveux central communément noté CNS. Elle agit sur les sécrétions pancréatiques, la vésicule biliaire, la mobilité des intestins, la satiété, dans le CNS, les fragments CCK contrôlent les réponses endocrines, moteurs et comportementales en agissant sur diverses fonctions telles que la croissance, la reproduction, la digestion, le métabolisme et la dynamique du système cardio-vasculaire. Une attention toute particulière est dévolue à cette hormone ainsi que tous les fragments dérivés depuis de nombreuses années spécialement de la part des biologistes et des pharmacologues et de nombreuses revues ont été publiées aussi bien sur le rôle, la dégradation, le répartition de CCK dans le CNS (62), que sur la chimie de la molécule, son rôle en tant qu'hormone et neurotransmetteur, son évolution (63 , 64 , 65 , 66 , 67). Cet intérêt pour CCK n'est pas étonnant lorsqu'on sait que CCK constitue le neuropeptide le plus abondant dans le CNS. De plus, les rôles multiples remplis par CCK ainsi que les différents récepteurs connus pour cette molécule justifient l'abondante littérature à ce sujet. En revanche, si le côté biologique et pharmacologique est relativement bien connu, le mécanisme d'action et les structures actives des fragments dérivés de CCK et, à plus forte raison, les caractéristiques structurales des différents récepteurs sont moins étudiés et n'ont reçu de l'attention que récemment avec l'utilisation croissante des techniques QSAR (Quantitative Structure Activity Relationships).

3.1.1 Intérêt biologique

L'hormone cholécystokinine contenant initialement 95 résidus d'acides aminés est convertie en fragments plus courts: CCK-58, -39, -33, -8, -7, -5 et -4. Il a été démontré que CCK est synthétisée dans le cerveau et ensuite convertie par un enzyme en fragments plus courts. Les CCK-8 et -7 sont en partie sulfatées ce qui augmente leur résistance envers la dégradation enzymatique par les aminopeptidase. Les fragments courts de CCK apparaissent en quantité plus ou moins élevée dans les différentes zones du cerveau. On retrouve le fragment CCK-5, sujet de notre étude en quantité particulièrement importante dans le cortex cérébral, siège de la médiation des processus sensoriels, moteurs et associatifs (68).

Lorsqu'on parle de l'activité biologique des molécules, il faut définir certains termes couramment utilisé tels qu'agoniste, antagoniste et récepteur. Le récepteur est la structure endogène qui "reconnaît" la molécule active ce qui permet l'expression du rôle de la molécule. Les parties de la molécule active qui sont spécifiquement reconnues par le récepteur et nécessaires à l'expression de sa fonction sont appelées pharmacophores. Un agoniste est une molécule qui est capable de se lier au même récepteur que la molécule active et provoque la même réponse biologique. Un antagoniste est une molécule capable également de se lier au même récepteur que la molécule active mais qui, cette fois, inhibe la réponse biologique en bloquant le site de reconnaissance du récepteur. On parlera d'activité pour définir la quantité de substance (molécule active, agoniste ou antagoniste) nécessaire à l'expression ou à l'inhibition de la réponse biologique. On peut ainsi classer les molécules par leur qualité agoniste ou antagoniste envers un récepteur particulier.

Dans le cas de CCK, de nombreuses études pharmacologiques ont conduit au développement d'agonistes et d'antagonistes puissants. Les caractéristiques structurales des récepteurs sont aussi progressivement mieux connues à mesure que l'on étudie la structure des agonistes, antagonistes et fragments CCK actifs. Ces études ont conduit à distinguer au moins trois types de récepteurs pour les fragments CCK: le récepteur CCK-A appelé récepteur périphérique et présent dans le pancréas, les intestins et le colon ainsi qu'en petite quantité en des régions très

localisées du cerveau (région du cerveau avec une barrière sang/cerveau relativement poreuse); le récepteur CCK-B présent avec une large répartition dans le cerveau et imperméable aux fragments CCK et analogues ou antagonistes circulant périphériquement; le récepteur stomacal noté "type-gastrine". Ces récepteurs sont de nature peptidique. Les molécules présentent tantôt une sélectivité à CCK-A ou CCK-B c'est à dire qu'une molécule reconnue par l'un des récepteurs n'aura peu ou pas d'activité sur l'autre tantôt aucune sélectivité c-à-d qu'elle peut être reconnue par n'importe quel récepteur. On note une affinité élevée de CCK-B pour les formes non-sulfatées alors que CCK-A reconnaît spécifiquement les formes sulfatées (69). Quant au récepteur "type-gastrine", les molécules reconnues par ce récepteur le sont aussi par CCK-B in vitro (70) ce qui suggère une relation structurale étroite entre les récepteurs CCK-B et "type-gastrine", la distinction étant faite in vivo par la localisation dans le CNS pour le premier et dans l'estomac pour le second. L'antagoniste le plus puissant est un composé de la famille des benzodiazépines appelé mk-329 ou L-364,718. L'activité de cet antagoniste est comparable à celle du peptide naturel CCK-8 (71). La structure par diffraction de rayons X de ce composé a été résolue (72). Ce composé est sélectif au récepteur CCK-A, ce qui signifie qu'il est spécifiquement reconnu par ce récepteur, et présente des similarités conformationnelles avec le fragment CCK-4 (73) dont la structure par diffraction de rayons X montre une conformation étendue (72, 75). Cinq familles importantes de composés chimiques ont servis de base pour produire des antagonistes de CCK-B: les benzodiazépines, les pyrazolidinones, les quinazolinones (70), les peptoïdes et les amides α-amino acides (74). Parmi ces familles, il est intéressant de constater qu'un antagoniste sélectif à CCK-B et au "type-gastrine" est le stéréo-isomère R de mk-329 que l'on note L-365-260 (71). Les agonistes sélectifs à CCK-B sont de nature peptidique. L'un est appelé BC264 soit l'enchaînement (Boc-Tyr(SO_3H)-gNle-mGly-Trp-(NMe)Nle-Asp-Phe-NH_2) est très actif et l'autre, noté BDNL, est simplement un dérivé de CCK-7: (Boc diNle CCK-7). Ces agonistes provoquent les mêmes réponses que le peptide naturel (75).

3.1.2 Intérêt structural

Nous savons que le fragment CCK-5: Gly-Trp-Met-Asp-Phe-NH_2 est présent en quantité importante dans certaines parties du cerveau. De par sa localisation, il doit donc être un agoniste du récepteur CCK-B. Néanmoins, les fragments peptidiques étant extrêmement flexibles, les études de structures ayant une activité sur les récepteurs ont surtout porté sur des molécules non-peptidiques plus rigides et donc plus faciles à étudier. Le plus petit fragment nécessaire à l'activité est CCK-4 dont la structure par diffraction de rayons X est connue (72, 75) :

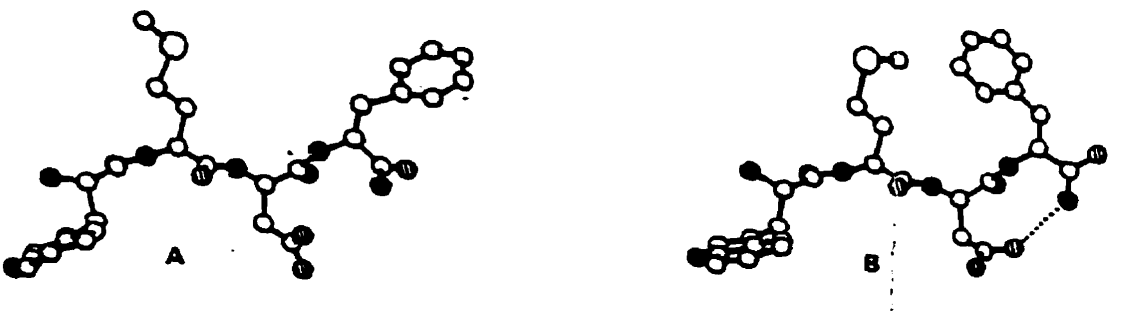


Figure 36. Structure de CCK-4 obtenue par diffraction de rayons X (2 molécules peptidiques de conformations différentes par unité assymétrique).

et qui présente une structure étendue. Ce peptide est actif sur les récepteurs CCK-A. Quant à CCK-5, il est connu que ce peptide adopte une structure repliée même en solution (76 , 77) ce qui en fait un candidat idéal pour une étude par modélisation moléculaire où les conformations de la molécule sont calculées dans le vide ce qui favorise les structures repliées. De plus, cette caractéristique conformationnelle distingue ce peptide de la structure de CCK-4 et laisse supposer un mode de liaison avec le ou les récepteurs éventuels différent pour ces deux fragments et donc une action différente. Cette observation renforce l'hypothèse

que CCK-5 se lie à CCK-B plutôt qu'à CCK-A. En étudiant ce peptide, nous avons plusieurs objectifs. Nous désirons bien entendu explorer l'espace conformationnel de la molécule. Celle-ci étant complexe, avec des chaînes latérales particulièrement encombrantes conduisant à des interactions possibles intéressantes à étudier comme les interactions entre les cycles aromatiques et les interactions entre le soufre et les autres parties de la molécule. Nous pourrons tester l'applicabilité des méthodes d'analyse de données au tri d'un échantillon complexe et de grande taille. Nous avons choisi d'étudier en parallèle la population de l'isomère D-Trp de CCK-5 ce dernier étant quasiment inactif biologiquement. Nous pourrons ainsi déterminer les caractéristiques conformationnelles nécessaires à l'activité. Ces deux isomères étant désignés ordinairement sous le nom de glmap (initiale des résidus qui le composent soit gly, l-tryptophane, méthionine, acide aspartique, phénylalanine) et gdmap (où le tryptophane est remplacé par son isomère d) dans la littérature, cette désignation sera conservée dans la suite du travail. De plus, connaissant quelques unes des caractéristiques conformationnelles des récepteurs, nous désirons utiliser les méthodes d'analyse de données pour trier l'échantillon en ayant des contraintes ce qui permettra de voir si et dans quelle mesure ce peptide rencontre les spécificités conformationnelles des récepteurs.

3.2 Analyse de données

3.2.1 Population

La population initialement générée par le programme PEPSEA comportait 20000 conformères dont 15877 ont convergé vers un minimum inférieur à 100kcal/mol. Les variables décrivant les conformères sont tous les angles ϕ, ψ et ω de la molécule soit 21 angles de torsion variables. Générer un échantillon représentatif pour une telle molécule pose le problème de la taille de l'échantillon. Nous savons que lorsque la taille de l'échantillon de n conformères tend vers la taille N de la population, la moyenne et l'écart-type de l'échantillon tendent respectivement vers la moyenne et l'écart-type de la population. Nous avons tracé moyenne

et écart-type en fonction de la taille croissante de l'échantillon:

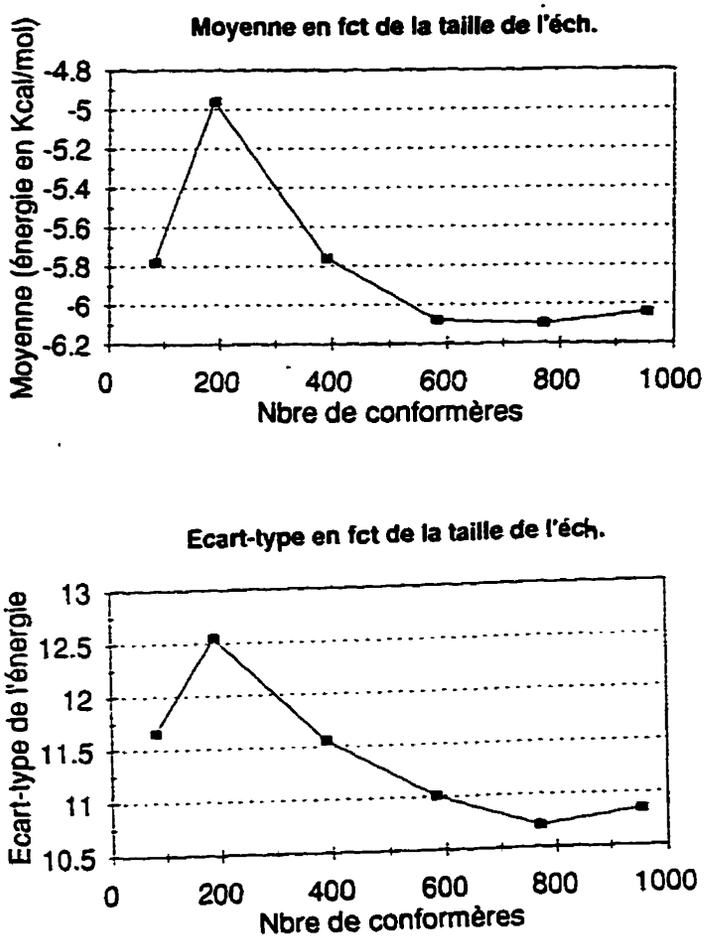


Figure 37. Moyenne et écart-type de l'échantillon en fonction du nombre d'individus.

Nous constatons qu'à partir de 1000 individus, la moyenne et l'écart-type tendent vers une valeur constante. La moyenne et l'écart-type de l'échantillon total de 15877 conformères étant respectivement de -6.09 kcal/mol et 10.73 kcal/mol en valeur non relative par rapport au minimum global. Si nous utilisons le critère de la stabilité de la moyenne et de l'écart-type,

nous constatons qu'un échantillon de 1000 conformères suffit pour représenter statistiquement la population totale.

Or nous savons par ailleurs qu'un échantillon statistiquement représentatif doit comporter 2 à 3% du nombre d'individus de la population. Cela représente dans notre cas un échantillon énorme et très supérieur à 1000 conformères. Comment dans ce cas décider de la taille correcte de l'échantillon? Nous avons décidé de générer un échantillon assez grand pour que nous puissions obtenir des "répliquants" c'est à dire des individus identiques. Cela nous indique que nous avons échantillonné la surface d'énergie correctement et même à plusieurs reprises. De plus, une étude précédente de cette molécule (78) effectuée à partir d'un échantillon de 10000 conformères avait conclu à une insuffisance possible de la taille de l'échantillon. Nous conservons un échantillon de 15000 conformères pour l'analyse subséquente, réparti entre -31.11kcal/mol et 13.09kcal/mol.

L'analogue D-Trp du peptide est étudié en parallèle. Nous avons donc généré une population de 20000 conformères avec les mêmes angles variables. Les distance définissant cet analogue sont donc identiques puisque la seule différence est l'orientation de la chaîne latérale du tryptophane. Nous conservons finalement un échantillon de 15000 minima métastables répartit entre -33.584 et 13.537kcal/mol.

3.2.2 Changement de variables

Pour traduire toutes les caractéristiques conformationnelles de cette molécule, nous utilisons un ensemble de distances inter-atomiques comportant deux types de distances: celles décrivant la chaîne principale du peptide et celle décrivant les interactions entre les chaînes latérales. La molécule numérotée est présentée sur la figure suivante et les distances sont décrites dans le tableau 16.

Tableau 16. Distances inter-atomiques en Å utilisées comme variables pour l'analyse de données.

D1	8-33	C_7 Trp
D2	8-50	C_{10} Trp-Met
D3	8-63	C_{13} Trp-Met-Asp
D4	8-83	C_{16} Trp-Met-Asp-Phe
D5	15-50	C_7 Met
D6	15-63	C_{10} Met-Asp
D7	15-83	C_{13} Met-Asp-Phe
D8	15-2	C_8 Trp-Gly
D9	39-63	C_7 Asp
D10	39-83	C_{10} Asp-Phe
D11	39-9	C_8 Met-Trp
D12	39-2	C_{11} Met-Trp-Gly
D13	56-83	C_7 Phe
D14	56-33	C_8 Asp-Met
D15	56-9	C_{11} Asp-Met-Trp
D16	56-2	C_{14} Asp-Met-Trp-Gly
D17	69-50	C_8 Phe-Asp
D18	69-33	C_{11} Phe-Asp-Met
D19	69-9	C_{14} Phe-Asp-Met-Trp
D20	69-2	C_{17} Phe-Asp-Met-Trp-Gly
D21	23-45	c.l. Trp- c.l. Met
D22	23-61	c.l. Trp- c.l. Asp
D23	23-80	c.l. Trp- c.l Phe
D24	23-46	c.l. Trp- c.l. Met
D25	23-60	c.l. Trp- c.l. Asp
D26	23-73	c.l. Trp-c.l. Phe
D27	26-46	c.l. Trp- c.l. Met
D28	26-45	c.l. Trp- c.l. Met
D29	26-60	c.l. Trp- c.l. Asp
D30	26-61	c.l. Trp- c.l. Asp
D31	26-80	c.l. Trp- c.l. Phe
D32	26-73	c.l. Trp- c.l. Phe
D33	46-60	c.l. Met- c.l. Asp
D34	46-61	c.l. Met- c.l. Asp
D35	46-80	c.l. Met- c.l. Phe
D36	46-73	c.l. Met- c.l. Phe
D37	45-60	c.l. Met- c.l. Asp
D38	45-61	c.l. Met- c.l. Asp
D39	45-80	c.l. Met- c.l. Phe
D40	45-73	c.l. Met- c.l. Phe
D41	60-80	c.l. Asp- c.l. Phe
D42	60-73	c.l. Asp- c.l. Phe
D43	61-80	c.l. Asp- c.l. Phe
D44	61-73	c.l. Asp- c.l. Phe

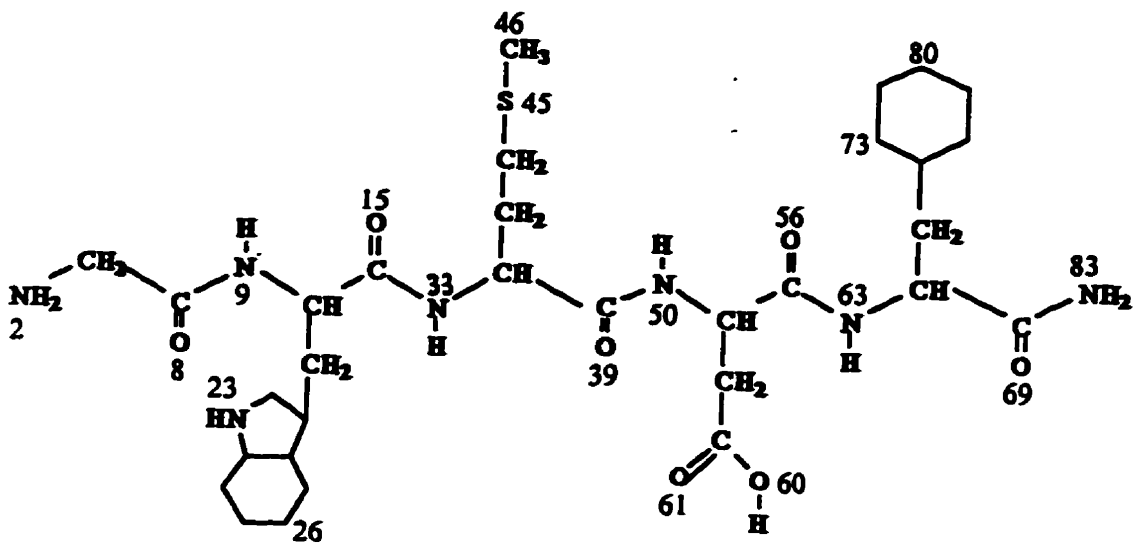


Figure 38. Schéma de la molécule CCK-5.

Suite aux conclusions du chapitre 3, nous avons proposé qu'une démarche correcte ne nécessite pas l'élimination de distances sur la base des longueurs de lien H. Nous conservons pour l'analyse en composantes principales les 44 distances retenues au départ pour lesquelles nous obtenons les statistiques élémentaires suivantes présentées dans le tableau suivant.

Si nous comparons les écarts-type mesurés sur ces distances avec ceux obtenus pour la molécule d'APYA, nous constatons qu'ils traduisent une distribution des distances beaucoup plus large, indice d'une plus grande diversité conformationnelle.

Tableau 17. Statistiqes élémentaires sur les distances pour glmap

VARIABLE	N	MOYENNE	ECART-TYPE	MINIMUM	MAXIMUM
D1	14994	8.45	1.95	3.26	11.96
D2	14994	9.22	2.59	2.69	15.28
D3	14994	11.16	3.54	3.11	20.07
D4	14994	9.10	2.17	3.25	13.45
D5	14994	9.29	2.59	2.63	15.28
D6	14994	10.41	3.09	3.12	18.43
D7	14994	9.48	2.37	3.42	14.49
D8	14994	8.91	2.12	3.36	13.23
D9	14994	9.65	2.69	2.95	15.81
D10	14994	9.59	2.70	3.05	15.97
D11	14994	11.39	3.75	3.20	20.83
D12	14994	10.68	3.29	3.19	19.05
D13	14994	8.37	1.76	3.15	12.44
D14	14994	8.30	1.81	3.12	12.46
D15	14994	10.52	3.02	3.37	17.25
D16	14994	9.67	2.46	3.53	15.43
D17	14994	7.69	1.56	3.01	11.25
D18	14994	7.61	1.62	3.06	11.22
D19	14994	10.01	2.79	3.33	16.21
D20	14994	9.10	2.21	3.37	14.58
D21	14994	8.40	2.11	2.97	12.38
D22	14994	7.23	1.60	2.75	10.78
D23	14994	8.31	2.13	2.91	12.39
D24	14994	7.12	1.66	2.71	10.65
D25	14994	3.93	0.76	2.39	5.18
D26	14994	5.96	1.37	2.51	8.70
D27	14994	7.67	1.88	2.65	12.16
D28	14994	9.18	2.52	2.56	15.71
D29	14994	3.81	0.79	2.34	5.14
D30	14994	5.86	1.37	2.56	8.66
D31	14994	7.69	1.96	2.56	12.09
D32	14994	5.55	0.70	3.91	7.13
D33	14994	3.90	0.81	2.34	5.17
D34	14994	6.06	1.45	2.57	8.72
D35	14994	5.67	0.71	3.95	7.14
D36	14994	7.43	1.42	3.17	10.72
D37	14994	3.98	0.79	2.39	5.15
D38	14994	5.66	0.73	3.96	7.14
D39	14994	7.45	1.41	3.31	10.67
D40	14994	8.79	2.33	2.67	14.08
D41	14994	5.66	0.77	3.99	7.15
D42	14994	7.51	1.42	3.23	10.74
D43	14994	8.93	2.21	2.67	14.26
D44	14994	10.06	3.02	2.71	17.77

Les distances pour l'analogue D-Trp du peptide sont définies par les mêmes atomes que pour le peptide de base et nous obtenons les statistiques élémentaires suivantes calculées sur la population de l'analogue D-Trp:

Tableau 18. Statistiques élémentaires sur les distances pour gdmap

VARIABLE	N	MOYENNE	ECART-TYPE	MINIMUM	MAXIMUM
D1	14998	3.90	0.76	2.40	5.18
D2	14998	5.87	1.42	2.63	8.75
D3	14998	7.55	1.93	2.55	12.29
D4	14998	9.06	2.60	2.59	15.81
D5	14998	3.81	0.79	2.36	5.15
D6	14998	5.85	1.37	2.57	8.67
D7	14998	7.68	1.98	2.56	12.08
D8	14998	5.56	0.69	3.91	7.13
D9	14998	3.90	0.81	2.33	5.17
D10	14998	6.05	1.43	2.53	8.72
D11	14998	5.62	0.65	3.95	7.15
D12	14998	7.33	1.31	3.11	10.74
D13	14998	3.98	0.78	2.37	5.14
D14	14998	5.66	0.73	3.96	7.14
D15	14998	7.35	1.43	3.19	10.62
D16	14998	8.59	2.34	2.67	13.91
D17	14998	5.66	0.77	3.99	7.15
D18	14998	7.51	1.42	3.23	10.71
D19	14998	8.81	2.27	2.61	14.11
D20	14998	9.79	3.10	2.69	17.24
D21	14998	8.46	1.98	3.26	12.02
D22	14998	9.30	2.55	2.75	15.24
D23	14998	11.19	3.51	3.10	19.96
D24	14998	9.06	2.22	3.30	13.47
D25	14998	9.37	2.56	2.63	15.32
D26	14998	10.46	3.02	3.12	18.12
D27	14998	9.46	2.40	3.40	14.27
D28	14998	8.89	2.16	3.47	12.98
D29	14998	9.73	2.62	2.91	15.98
D30	14998	9.66	2.61	2.96	15.94
D31	14998	11.40	3.73	3.24	20.76
D32	14998	10.70	3.24	3.20	19.24
D33	14998	8.39	1.77	3.18	12.45
D34	14998	8.33	1.81	3.11	12.46
D35	14998	10.49	3.03	3.30	17.32
D36	14998	9.66	2.45	3.39	15.51
D37	14998	7.71	1.56	3.09	11.23
D38	14998	7.64	1.62	3.07	11.22
D39	14998	9.97	2.82	3.35	16.30
D40	14998	9.08	2.21	3.34	14.59
D41	14998	8.39	2.10	2.94	12.37
D42	14998	7.23	1.59	2.75	10.73
D43	14998	8.31	2.12	2.93	12.39
D44	14998	7.12	1.66	2.70	10.59

3.2.3 Analyse en composantes principales

3.2.3.1 Analyse de gimap

Dans un premier temps, les corrélations entre chacune des 44 distances sont calculées. La matrice des corrélations obtenue est alors diagonalisée, conduisant aux valeurs propres et aux vecteurs propres correspondants. Les nouvelles variables composantes principales sont construites par combinaison linéaire des variables distances initiales. L'ACP pour cette molécule prend quelques minutes de CPU. La matrice des corrélations ainsi que les valeurs et les vecteurs propres correspondants sont présentés en annexe A.

La matrice des corrélations montre un fait intéressant: nous n'observons pas de corrélations importantes (à partir de .7 de coefficient de corrélation) entre les distances décrivant la chaîne principale (les 20 premières) et les distances décrivant la chaîne latérale (les 24 suivantes). En revanche, entre distances de même type, des corrélations importantes apparaissent. De plus les distances D32 et D37 ne sont corrélées, de manière significative, avec aucune autre. D32 est la distance entre la chaîne latérale du Trp et celle de la Phe et D37, la distance entre la chaîne latérale de la Mét et celle de l'Asp.

L'examen des valeurs propres associées à chaque composante principale donne le pourcentage de variance associé à chaque composante principale soit une mesure de la représentativité de cette dernière. La première composante principale porte 16.40% de l'information initialement contenue dans les 44 distances initiales. Pour conserver 90% de l'information initiale, nous devons conserver 15 composantes principales. Ces 15 premières composantes principales seront utilisées comme variables pour l'analyse de regroupement.

Les vecteurs propres sont les coefficients de combinaison linéaire qui permettent de construire les composantes principales à partir des distances initiales. Les distances dont le coefficient de combinaison linéaire sera le plus important seront celles qui décriront le mieux l'échantillon de conformères spécialement dans la construction des premières composantes principales. La première composante principale est constituée en majorité de distances relatives aux chaînes latérales de la molécule. Les distances D28 (c.l. Trp- c.l. Met), D43 (c.l.

Asp- c.l. Phe) et D44 (c.l. Asp- c.l. Phe) ont les coefficients les plus élevés. La deuxième composante principale en revanche est construite majoritairement par les distances relatives à la chaîne principale avec une grande participation des distances D3 (C_{13} sur Trp-Met-Asp), D6 (C_{10} sur Met-Asp), D11 (C_8 sur Met-Trp) et D12 (C_{11} sur Met-Trp-Gly). Ces distances traduisent une organisation de la chaîne principale autour des résidus initiaux du peptide sans participation de la chaîne latérale de la Phe. La troisième composante principale est constituée majoritairement par les distances D21, D22, D23, D24 qui décrivent toutes les interactions de l'azote de la chaîne latérale du résidu Trp avec les chaînes latérales des autres résidus du peptide. La quatrième composante principale décrit les interactions sur la chaîne principale impliquant surtout les résidus Asp et Phe (distances D14 à D20). La cinquième composante principale est composée surtout des distances D4, D7 et D8 qui traduisent l'organisation à longue distance du peptide puisqu'elles représentent respectivement les pseudo-cycles à 16 membres sur Trp-Met-Asp-Phe, à 13 membres sur Met-Asp-Phe et à 8 membres sur Trp-Gly. La sixième composante principale est composée de distances impliquant les deux types d'interactions: D13, D14, D17 et D18 pour la chaîne principale et D21 à D24 pour les chaînes latérales c'est à dire les résidus Phe, Asp et Met et leurs chaînes latérales respectives. Néanmoins, aucun des coefficients de combinaison linéaire pour ces distances n'est très élevé et par conséquent, aucune distance n'est prépondérante. De manière générale, les composantes principales sont constituées de distances soit décrivant la chaîne principale, soit décrivant les chaînes latérales mais pas les deux à la fois, excepté pour la sixième composante principale. Sur les 15 composantes principales conservées pour l'analyse de regroupement, cinq décrivent la chaîne principale (cp2, cp4, cp5, cp8, cp15) et neuf les chaînes latérales (cp1, cp3, cp7, cp9, cp10, cp11, cp12, cp13, cp14). Si on additionne le pourcentage de variance de chaque composante principale pour chaque type d'information, on obtient que 34.48% de l'information est reliée à la chaîne principale pour 5 composantes principales et 48.59% de l'information est reliée aux chaînes latérales pour 9 composantes principales sur un total de 83.07% d'information (soit la somme des quinze premières composantes principales dont on retire la variance de la sixième composante principale qui décrit les deux types d'interaction). En

conclusion, on montre ici que la diversité conformationnelle du peptide est due, en grande partie, à la présence des chaînes latérales du peptide occasionnant de nombreuses possibilités d'interactions. De plus, ainsi que tend à la prouver l'étude des corrélations entre les distances, les caractéristiques conformationnelles de la chaîne principale seraient indépendantes de celles de la chaîne latérale.

3.2.3.2 Population de gdmap

L'examen de la matrice des corrélations entraîne les même observations que pour glmap. En effet, il n'y a pas de corrélation significative entre les distances décrivant la chaîne principale et celles décrivant les chaînes latérales. De plus, la distance D8 correspondant à l'interaction C_8 Trp-Gly n'est corrélée avec aucune autre distance alors qu'elle l'était de manière importante avec D7 (C_{13} Met-Asp-Phe) pour glmap.

L'examen des valeurs propres fournit le pourcentage de variance associé à chaque composante principale. Pour 15 composantes principales, nous avons 89.91% d'information qui est la valeur la plus proche des 90.06% d'information que nous avons conservé pour gdmap.
Les vecteurs propres sont les coefficients de combinaison linéaire des distances dans les composantes principales. Là encore, nous constatons que les composantes principales sont majoritairement construites à partir des distances décrivant la chaîne principale ou à partir des distances décrivant la chaîne latérale excepté pour la sixième composante principale construite à partir des deux types de distances comme pour glmap. Néanmoins apparaissent ici des différences entre les deux populations: la première composante principale pour gdmap, c-à-d la plus importante en terme de représentativité, est construite par les distances reliées à la chaîne principale: D4, D19 et D20, les deux dernières distances étant assez peu importantes dans la description de gdmap (elles n'apparaissent pas de manière significative dans la construction des 15 premières composantes principales). La deuxième composante principale est reliée aux distances des chaînes latérales: D23, D26, D31, D32 les deux dernières n'étant

pas, encore une fois, des distances essentielles à la description de glmap. La troisième composante principale est constituée par les distance D41, D42, D43, D44 décrivant toutes les interactions de la chaîne latérale de l'Asp avec la chaîne latérale de la Phe. La quatrième composante principale est construite essentiellement des distances D35 à D40 qui traduisent les interaction de la chaîne latérale de la Met avec les chaînes latérales des résidus Phe et Asp. La cinquième composante principale est constituée par D21, D24, D27 et D28 qui décrivent toute l'interaction de la chaîne latérale du Trp avec celle de la Met. La sixième composante principale est constitué par les deux distances D6 et D9 reliées à la chaîne principale soit C_{10} Met-Asp et C_7 Asp, et par les deux distances D33 et D37 reliées à la description des chaînes latérales soit l'interaction des chaînes latérales de la Met et de l'Asp. Si on additionne le pourcentage de variance de chaque composante principale pour chaque type d'information, on obtient que 40.65% de l'information est reliée à la chaîne principale avec 8 composantes principale et 42.26% est reliée aux chaînes latérales avec 6 composantes principales sur un total de 82.90% d'information (soit la somme des variances des quinze premières composantes principales dont on retire la variance de la sixième composante principale puisqu'elle décrit les deux types d'interaction). On se trouve dans le cas inverse de ce qui se passe pour glmap pour laquelle la majorité des composantes principales sont construites à partir des distances qui décrivent les chaînes latérales. Ici, 8 des 14 composantes décrivent la chaîne principale et 6 décrivent la chaîne latérale. Le pourcentage de variance est à peu près équivalent pour les deux types de composantes principales. Nous constatons une différence importante entre les deux molécules: les conformations de glmap sont fortement influencées par la présence des chaînes latérales. Pour gdmap, ceci est beaucoup moins apparent dû sans doute au fait que la position défavorable de la chaîne latérale du D-Trp contraint la chaîne principale et limite les possibilités de conformations et d'interactions.

3.2.4 Regroupement

Le regroupement d'un échantillon de cette taille et avec cette diversité conformationnelle

élevée pose de nombreux problèmes de classification à la fois théoriques et pratiques. De plus, la littérature est pauvre en exemple de traitement de très grands échantillons. Au point de vue théorique, la détermination du nombre de familles dans l'échantillon devient encore plus problématique que pour un petit échantillon. Pour les méthodes de regroupement hiérarchique, il est nécessaire de choisir une méthode non biaisée. Au point de vue pratique, le regroupement hiérarchique d'un tel échantillon nécessite un stockage de données énorme ainsi qu'un temps de CPU élevé. C'est pourquoi, en pratique, on procède en général à un pré-regroupement non-hiérarchique pour réduire le nombre d'individus à regrouper ensuite de façon hiérarchique. Nous avions dans un premier temps tenté cette approche: nous avons pré-regroupé notre échantillon en utilisant la méthode FASTCLUS. Nous nous sommes alors heurté à un nouveau problème: à quel niveau devons nous arrêter le pré-regroupement? Nous avons décidé de suivre les indications fournies par les indices statistiques en fonction du niveau de pré-regroupement. Nous avons observé que la réponse sur le nombre de familles finales révélé par l'étude des indices statistiques variait énormément en fonction du nombre de pré-familles et rendait impossible l'exploitation des résultats. Nous avons donc décidé de procéder à un regroupement direct avec FASTCLUS qui étant une méthode de classification non-hiérarchique est relativement rapide.

3.2.4.1 Indices statistiques pour gimap

Les indices statistiques CCC et Pseudo F on été tracés pour un nombre de familles finales de 2 à 21.

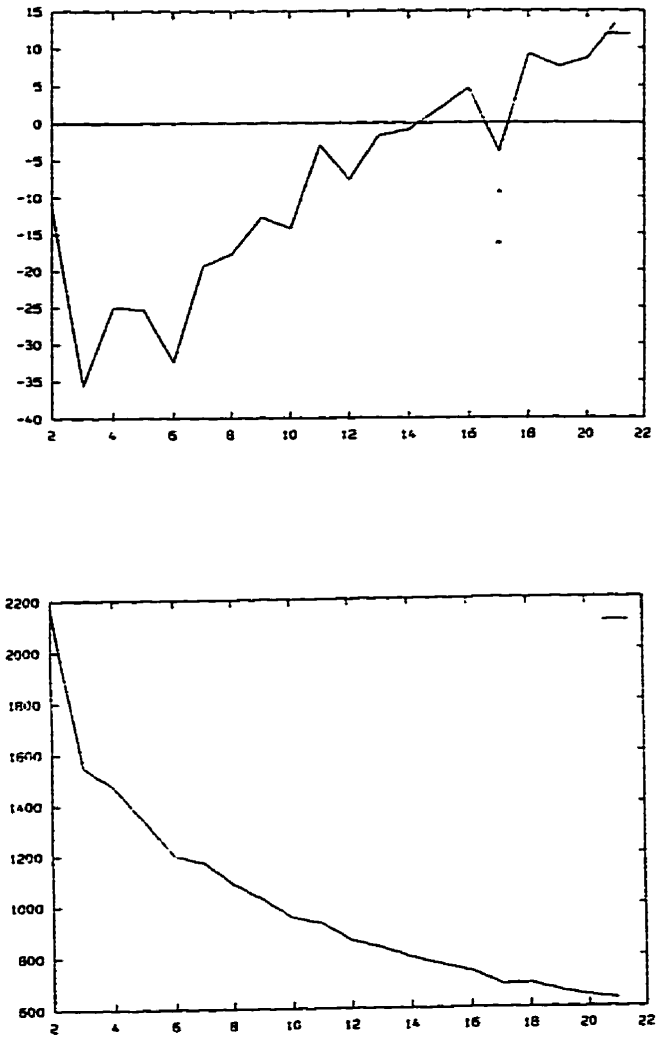


Figure 39. Indices statistiques CCC (haut) et PseudoF (bas) en fonction du nombre de familles.

Le graphique pour CCC suggère 9, 11, 16 ou 18 familles finales. Pour le Pseudo F, moins facile à interpréter, il est suggéré 7, 11, 16 ou 18 familles finales. Nous décidons après avoir regardé les résultats pour ces différents nombre de familles de présenter le regroupement pour 16 familles finales.

3.2.4.2 Indices statistiques pour gdmap

La même procédure de recherche du nombre de famille optimal a été menée pour gdmap. Si nous voulons pouvoir comparer les résultats de la classification entre les deux populations, nous devons opter pour un nombre de familles final comparable sinon identique. En effet, il n'y aurait aucun sens a comparer les résultats de classification pour 11 familles finales sur la population de gdmap si ce nombre n'a pas été suggéré par les l'étude des indices statistiques puisque la classification obtenue serait mauvaise. Nous chercherons donc le nombre optimal pour gdmap qui se rapproche le plus de 16.

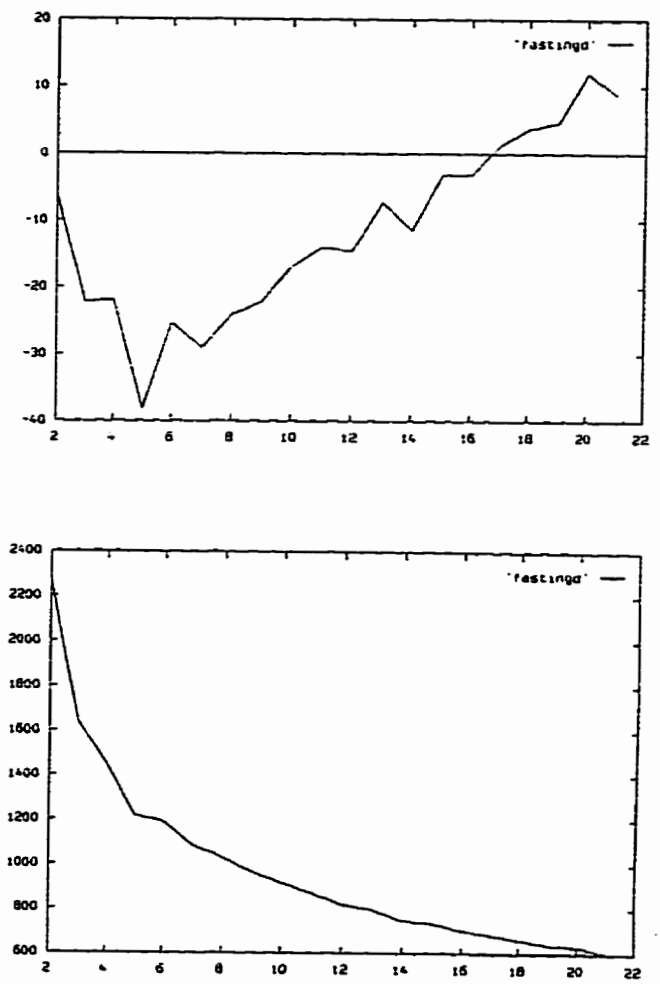


Figure 40. Indices statistiques CCC (haut) et PseudoF (bas) en fonction du nombre de familles.

L'étude des graphiques pour CCC et PseudoF suggère 13 familles. De manière générale, nous n'obtenons pas du tout les mêmes nombre de familles optimaux que pour glmap puisqu'ici, CCC suggère 6, 11, 13, 20 et PseudoF 6, 13, 20 comme nombre optimal de familles dans l'échantillon.

3.2.4.3 Résultats de la classification pour glmap

Nous avons suite à l'analyse en composantes principales constaté qu'une grande partie des caractéristiques conformationnelles de cette molécule était liée aux positions des chaînes latérales. C'est pourquoi dans ce cas, l'étude des codes conformationnels à l'intérieur de chaque famille ne nous sera pas d'un grand secours. En effet les codes conformationnels repèrent les domaines conformationnels pour des angles Φ et Ψ caractéristiques. Ces codes sont donc utiles pour décrire la chaîne principale d'un peptide.

Dans le cas présent, nous allons plutôt interpréter directement les résultats en regardant les moyennes et écart-type pour les distances à l'intérieur de chaque famille. Nous désirons trouver quelles sont les distances caractéristiques propres à chaque famille. Les distances les plus caractéristiques seront celles pour lesquelles la distribution dans l'échantillon sera la moins large possible soit celles possédant l'écart-type le plus petit possible. Pour obtenir une valeur utilisable, nous devons comparer l'écart-type dans la famille à l'écart-type, pour la même distance, dans la population totale. Nous calculons:

$$\frac{\sigma_{tot}-\sigma_{fam.}}{\sigma_{tot}}\times 100 \qquad [6]$$

Nous présentons le résultat du calcul pour 16 familles finales trouvées par FASTCLUS.

Tableau 19. Différences entre les écart-type dans les 16 familles et dans la population totale pour les 44 distances caractéristiques de glmap.

famille	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12	d13	d14	d15	d16	d17	d18	d19	d20	d21	d22
1	27.	22.	29.	22.	20.	27.	17.	17.	13.	16.	31.	29.	7.	11.	42.	43.	13.	13.	39.	40.	25.	23.
2	37.	29.	31.	33.	28.	27.	32.	36.	32.	32.	21.	18.	5.	3.	42.	42.	6.	7.	40.	39.	41.	36.
3	36.	25.	20.	38.	24.	25.	25.	21.	28.	27.	22.	24.	31.	29.	22.	26.	25.	25.	19.	27.	31.	27.
4	38.	30.	19.	35.	29.	35.	42.	45.	31.	31.	23.	38.	27.	29.	25.	28.	20.	24.	30.	27.	32.	28.
5	13.	37.	46.	13.	35.	44.	15.	10.	30.	34.	47.	44.	29.	32.	44.	39.	31.	35.	45.	39.	6.	14.
6	29.	26.	31.	32.	24.	28.	26.	23.	26.	28.	27.	26.	29.	32.	31.	27.	25.	29.	31.	26.	15.	16.
7	34.	17.	31.	33.	17.	36.	35.	37.	17.	16.	35.	42.	38.	41.	21.	19.	33.	36.	23.	23.	29.	29.
8	7.	26.	25.	7.	22.	31.	15.	19.	17.	20.	23.	27.	4.	2.	33.	38.	5.	4.	32.	36.	27.	22.
9	18.	36.	43.	16.	32.	44.	26.	28.	33.	34.	43.	45.	25.	23.	34.	27.	20.	17.	35.	28.	17.	21.
10	2.	24.	25.	2.	25.	30.	5.	5.	21.	22.	23.	27.	16.	19.	25.	24.	16.	16.	27.	24.	26.	19.
11	31.	12.	31.	35.	12.	28.	32.	27.	11.	12.	30.	29.	8.	4.	27.	37.	9.	6.	22.	33.	32.	27.
12	31.	22.	26.	34.	22.	27.	26.	26.	17.	15.	26.	27.	23.	19.	26.	31.	23.	20.	21.	27.	18.	17.
13	10.	28.	19.	10.	27.	29.	7.	4.	27.	28.	19.	26.	22.	24.	34.	31.	19.	21.	33.	33.	32.	28.
14	8.	29.	24.	9.	30.	32.	9.	7.	27.	27.	23.	28.	27.	26.	32.	28.	29.	27.	30.	30.	32.	27.
15	2.	13.	41.	4.	15.	37.	6.	8.	17.	14.	45.	43.	29.	31.	43.	37.	27.	28.	46.	39.	33.	26.
16	34.	12.	31.	33.	12.	30.	36.	36.	12.	13.	29.	27.	25.	28.	34.	40.	25.	27.	32.	36.	6.	14.

famille	d23	d24	d25	d26	d27	d28	d29	d30	d31	d32	d33	d34	d35	d36	d37	d38	d39	d40	d41	d42	d43	d44
1	23.	22.	3.	1.	16.	24.	1.	3.	6.	-6.	5.	-3.	7.	6.	-3.	23.	26.	24.	8.	15.	25.	26.
2	41.	39.	3.	9.	22.	26.	17.	22.	16.	6.	17.	3.	8.	13.	1.	45.	35.	24.	16.	21.	32.	23.
3	29.	26.	13.	33.	41.	41.	-4.	25.	32.	-9.	5.	12.	28.	43.	1.	21.	35.	46.	9.	26.	46.	51.
4	32.	28.	-1.	21.	35.	40.	19.	38.	38.	3.	10.	26.	-3.	-5.	1.	19.	43.	34.	1.	23.	46.	32.
5	3.	10.	4.	3.	12.	15.	15.	13.	6.	9.	-5.	-8.	19.	15.	3.	3.	11.	17.	4.	13.	12.	13.
6	18.	17.	5.	29.	36.	33.	3.	25.	30.	-8.	-4.	-9.	25.	35.	3.	12.	30.	38.	3.	20.	42.	47.
7	27.	29.	11.	11.	31.	40.	16.	42.	36.	7.	28.	6.	31.	18.	-1.	40.	44.	26.	10.	28.	50.	41.
8	23.	16.	5.	12.	16.	28.	0.	19.	22.	14.	32.	19.	33.	20.	1.	30.	33.	21.	8.	26.	33.	24.
9	15.	16.	0.	12.	30.	26.	8.	25.	22.	7.	1.	-3.	-4.	-5.	-3.	4.	30.	27.	5.	19.	36.	36.
10	26.	17.	-3.	2.	21.	28.	0.	1.	6.	4.	-10.	-4.	8.	7.	5.	32.	26.	12.	9.	23.	32.	21.
11	31.	27.	1.	34.	45.	48.	3.	16.	28.	-6.	30.	25.	29.	35.	3.	26.	33.	40.	29.	48.	46.	39.
12	21.	20.	11.	29.	39.	34.	16.	21.	22.	-6.	0.	4.	8.	22.	0.	23.	23.	29.	9.	12.	28.	34.
13	31.	28.	-1.	22.	34.	37.	14.	29.	31.	4.	10.	17.	13.	4.	0.	26.	43.	35.	6.	22.	44.	37.
14	36.	31.	17.	17.	30.	31.	10.	12.	16.	13.	9.	13.	29.	21.	3.	15.	20.	23.	13.	21.	23.	22.
15	26.	17.	3.	18.	31.	38.	9.	30.	35.	-7.	12.	14.	-6.	-3.	1.	4.	23.	31.	9.	30.	46.	39.
16	4.	11.	3.	17.	21.	32.	28.	23.	15.	1.	-2.	-6.	40.	33.	-1.	18.	30.	29.	9.	18.	27.	23.

Nous présentons également le résultat du même calcul dans le tableau 20 mais pour un nombre de familles finales de 6, ce qui était fortement déconseillé par les indices statistiques.

Nous constatons effectivement que les distances à l'intérieur des familles possèdent des distributions plus larges ce qui se traduit par des écarts-types plus grands donc de faibles valeurs pour l'équation [6]. Nous regardons ensuite les valeurs des distances les plus "concentrées" en terme de distribution pour chaque famille. Nous avons souligné les écart-types les plus faibles à l'intérieur des familles (donc possédant un gros pourcentage de différence d'avec les écart-types de distances pour la population totale). Les distances correspondantes vont être les plus importantes dans la description des individus pour chaque famille. Nous constatons néanmoins que même les plus grands pourcentages de différence atteignent à peine 50% ce qui montre que la classification n'est pas excellente. En effet, à la limite, nous pourrions avoir dans chaque famille des distances caractéristiques pour lesquelles

Tableau 20. Différences entre les écart-type dans les 6 familles et dans la population totale pour les 44 distances caractéristiques de glmap.

distances	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
cluster																						
1	9.	23.	16.	8.	24.	15.	13.	15.	21.	19.	15.	13.	5.	6.	25.	24.	6.	4.	26.	22.	21.	17.
2	18.	18.	17.	19.	17.	21.	11.	8.	19.	19.	16.	19.	28.	29.	5.	9.	24.	26.	4.	8.	10.	9.
3	-5.	13.	16.	-5.	13.	16.	-5.	-7.	9.	9.	15.	15.	31.	35.	28.	33.	32.	36.	26.	30.	3.	10.
4	-4.	25.	11.	-5.	24.	24.	1.	3.	24.	24.	16.	28.	27.	28.	6.	5.	23.	25.	7.	7.	18.	16.
5	34.	14.	31.	35.	13.	28.	32.	31.	12.	13.	26.	25.	5.	1.	26.	33.	6.	3.	22.	29.	26.	23.
6	-4.	18.	19.	-3.	18.	21.	-3.	-4.	13.	15.	17.	19.	3.	6.	23.	25.	8.	8.	20.	22.	24.	22.

distances	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44
cluster																						
1	20.	16.	-3.	7.	17.	16.	1.	-1.	1.	4.	-2.	-6.	-6.	-1.	4.	16.	17.	12.	6.	15.	13.	15.
2	10.	10.	8.	22.	31.	32.	0.	21.	26.	-7.	2.	1.	19.	29.	0.	15.	30.	36.	1.	20.	41.	42.
3	5.	10.	3.	4.	9.	18.	22.	12.	3.	1.	1.	-4.	35.	20.	1.	0.	9.	6.	-1.	6.	14.	13.
4	17.	13.	0.	3.	22.	30.	3.	30.	29.	7.	2.	7.	-1.	-3.	-1.	15.	34.	26.	4.	20.	43.	33.
5	25.	24.	1.	25.	28.	27.	1.	-5.	1.	-3.	1.	3.	17.	28.	0.	5.	21.	32.	9.	13.	25.	29.
6	24.	22.	0.	-4.	18.	23.	3.	-7.	-4.	4.	0.	-2.	8.	11.	-1.	12.	18.	16.	4.	8.	21.	28.

l'écart-type serait de 0 comme lors de la classification de la population de l'alanine ce qui donnerait des valeurs de 100% dans l'équation [6] soit une séparation optimale des molécules.

Nous pouvons aussi calculer une quantité équivalente sur les moyennes:

$$\frac{\mu_{tot}-\mu_{fam.}}{\mu_{tot}}\times 100 \qquad [7]$$

dont nous présentons le résultat dans le tableau 21.

Nous avons indiqué quelles sont les distances pour lesquelles la moyenne dans la famille est très différente de la moyenne dans la population totale. Si le signe de la différence est négatif, cela signifie que la moyenne de distance dans la famille concernée est supérieure à la moyenne pour cette distance dans la population totale ce qui indique que cette famille contient des molécules plutôt étendues. Si le signe est positif, la distance dans la famille est

inférieure en moyenne à la même distance dans la population totale ce qui indique que cette famille contient des molécules plutôt repliées.

En conclusion, nous pouvons faire la même remarque qu'en ce qui concerne les Tableau 18., Tableau 19., Tableau 20., soit que nous n'obtenons des pourcentages de différence d'au plus 40%, ce qui signifie encore une classification peu efficace.

Tableau 21. Différences entre les moyennes dans les 16 familles et dans la population totale pour les 44 distances caractéristiques de glmap.

famille	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12	d13	d14	d15	d16	d17	d18	d19	d20	d21	d22
1	28.	9.	-6.	29.	8.	-1.	30.	29.	10.	10.	-7.	-3.	16.	17.	-23.	-19.	15.	17.	-22.	-17.	-15.	-12.
2	-15.	-18.	21.	-14.	-19.	22.	-15.	-14.	-18.	-17.	23.	24.	14.	16.	-21.	-16.	13.	15.	-19.	-14.	-24.	-20.
3	2.	35.	10.	0.	35.	10.	3.	5.	14.	14.	11.	12.	-8.	-7.	28.	24.	-7.	-6.	28.	24.	-7.	-7.
4	-18.	-20.	-30.	-19.	-21.	-33.	-18.	-17.	-20.	-21.	-30.	-34.	-2.	-2.	27.	27.	-3.	-4.	24.	23.	-7.	-6.
5	25.	-14.	-19.	27.	-12.	-16.	28.	26.	-14.	-15.	-21.	-19.	-16.	-17.	-22.	-20.	-15.	-16.	-20.	-19.	21.	20.
6	-3.	34.	15.	-4.	33.	14.	-5.	-4.	32.	33.	14.	13.	-11.	-11.	-7.	-5.	-11.	-10.	-6.	-4.	25.	21.
7	-14.	11.	-29.	-14.	11.	-26.	-15.	-14.	13.	13.	-28.	-25.	-13.	-13.	17.	18.	-12.	-12.	15.	16.	-17.	-15.
8	-3.	-14.	-15.	-3.	-12.	-13.	-4.	-4.	-12.	-14.	-15.	-13.	11.	8.	-25.	-24.	11.	8.	-26.	-24.	-7.	-4.
9	-10.	-25.	-29.	-9.	-24.	-30.	-9.	-9.	-23.	-24.	-30.	-30.	-4.	-5.	2.	5.	-5.	-6.	1.	4.	31.	26.
10	-2.	-12.	6.	-2.	-12.	8.	-2.	-3.	-11.	-11.	6.	8.	22.	23.	15.	11.	21.	22.	14.	10.	27.	22.
11	-7.	9.	28.	-8.	9.	29.	-10.	-8.	10.	10.	30.	31.	7.	6.	-20.	-21.	7.	6.	-18.	-20.	-7.	-6.
12	-9.	-4.	28.	-10.	-5.	27.	-10.	-9.	-6.	-4.	27.	26.	28.	29.	3.	1.	26.	28.	4.	2.	-1.	-3.
13	22.	-12.	14.	23.	-13.	6.	23.	22.	-14.	-13.	13.	5.	-5.	-4.	32.	28.	-5.	-4.	32.	28.	-2.	-3.
14	22.	-8.	15.	23.	-8.	9.	24.	22.	-8.	-7.	15.	8.	-11.	-12.	22.	16.	-11.	-12.	23.	16.	-5.	-5.
15	6.	0.	-40.	8.	2.	-35.	7.	5.	0.	-2.	-41.	-37.	-9.	-11.	-22.	-18.	-8.	-10.	-22.	-17.	-15.	-10.
16	-11.	6.	21.	-12.	6.	23.	-15.	-14.	6.	6.	23.	24.	-10.	-11.	-19.	-19.	-9.	-10.	-17.	-16.	15.	15.

famille	d23	d24	d25	d26	d27	d28	d29	d30	d31	d32	d33	d34	d35	d36	d37	d38	d39	d40	d41	d42	d43	d44
1	-16.	-14.	-4.	4.	22.	28.	5.	18.	20.	1.	7.	8.	6.	9.	3.	2.	15.	23.	3.	12.	27.	35.
2	-24.	-22.	4.	1.	18.	22.	7.	30.	28.	-2.	14.	12.	3.	2.	4.	0.	5.	2.	2.	13.	21.	16.
3	-6.	-7.	-10.	-17.	-24.	-29.	-4.	-11.	-17.	6.	2.	-4.	-9.	-16.	-1.	-6.	-17.	-25.	-4.	-12.	-24.	-32.
4	-7.	-5.	0.	-12.	-21.	-27.	-15.	-21.	-25.	-2.	-8.	-12.	6.	3.	-3.	-10.	-18.	-18.	-5.	-17.	-27.	-26.
5	22.	21.	6.	20.	20.	22.	8.	4.	11.	0.	-1.	6.	3.	10.	3.	4.	14.	25.	4.	8.	16.	25.
6	25.	21.	-5.	-11.	-16.	-13.	2.	-4.	0.	2.	0.	5.	-7.	-13.	2.	-5.	-14.	-21.	6.	2.	-9.	-16.
7	-18.	-17.	7.	11.	1.	-5.	11.	1.	-2.	0.	13.	13.	-4.	-2.	4.	-6.	-8.	-6.	1.	-3.	-12.	-10.
8	-6.	-3.	9.	21.	20.	19.	0.	-9.	-8.	-3.	-15.	-13.	0.	10.	0.	19.	32.	44.	-6.	-5.	11.	29.
9	30.	27.	6.	-3.	-9.	-4.	-10.	-12.	-3.	-3.	3.	8.	7.	8.	2.	-6.	-11.	-8.	8.	5.	-4.	-4.
10	26.	21.	1.	4.	19.	31.	1.	15.	23.	-2.	2.	1.	2.	5.	-6.	1.	10.	14.	10.	25.	38.	38.
11	-7.	-6.	-4.	-9.	-13.	-15.	2.	-5.	-10.	1.	-14.	-13.	-8.	-13.	-6.	18.	12.	-2.	-11.	-12.	-12.	-21.
12	-1.	-4.	-6.	-12.	-4.	-1.	-16.	-6.	-4.	4.	3.	1.	-9.	-14.	1.	3.	-5.	-15.	1.	7.	2.	-8.
13	-2.	-2.	-5.	-10.	-12.	-19.	-10.	-11.	-16.	-3.	2.	-4.	15.	18.	-2.	-9.	-10.	-6.	-4.	-12.	-14.	-8.
14	-5.	-5.	10.	27.	24.	14.	8.	6.	-2.	0.	4.	-1.	1.	10.	-3.	2.	11.	24.	-5.	-6.	2.	15.
15	-13.	-9.	-5.	-8.	-20.	-22.	0.	-15.	-14.	-1.	-13.	-11.	5.	3.	1.	-1.	-6.	-7.	-5.	-16.	-20.	-20.
16	15.	15.	2.	5.	14.	21.	13.	24.	28.	-1.	-3.	2.	-6.	-7.	0.	0.	4.	-2.	7.	18.	25.	17.

Si nous examinons alors les distances caractéristiques de chaque familles (soit celles possédant les valeurs les plus élevées pour résultats des équations [6] et [7]) nous pouvons trouver les caractéristiques conformationnelles de chaque famille reliées à ces distances. Les 20 premières distances décrivent la chaîne principale, les 24 suivantes décrivent les chaînes latérales.

Si nous faisons la somme des différences négatives d'une part, et des différences positives d'autre part, pour chaque type de distance, nous aurons la tendance générale de repliement (somme positive élevée) et d'extension (somme négative élevée) pour chaque famille. Nous présentons ces sommes calculée pour chacune des 16 familles dans le tableau suivant 22.

Tableau 22. Somme des résultats de l'équation (μ_{tot}-$\mu_{fam.}$/ μ_{tot}) positifs d'une part et négatifs d'autre part pour chaque type de distance dans chaque famille de glmap.

Famille	Chaîne principale (D1 à D20)		Chaînes latérales (D21 à D44)	
	$\mu_{fam.} < \mu_{tot.}$	$\mu_{fam.} > \mu_{tot.}$	$\mu_{fam.} < \mu_{tot.}$	$\mu_{fam.} > \mu_{tot.}$
1	218	-97	249	-62
2	149	-200	202	-92
3	295	-28	7	-291
4	101	-292	10	-294
5	106	-274	294	-1
6	186	-83	112	-134
7	115	-217	61	-126
8	38	-221	213	-80
9	13	-271	161	-77
10	164	-53	336	-8
11	183	-112	32	-194
12	227	-55	22	16
13	249	-69	34	-160
14	215	-77	159	-37
15	29	-274	9	-230
16	115	-164	239	-20

Nous présentons enfin la valeur de la moyenne pour chacune des 44 distances dans chacune des 16 familles notées Md de 1 à 44. La molécule ayant une conformation où toutes les distances interatomiques sont égales à la moyenne pour une famille donnée, représentera le centre de masse de la famille.

Tableau 23. Moyennes par familles pour les 44 distances caractéristiques de glmap.

famille	Md1	Md2	Md3	Md4	Md5	Md6	Md7	Md8	Md9	Md10	Md11	Md12	Md13	Md14	Md15	Md16	Md17	Md18	Md19	Md20	Md21	Md22
1	6.10	8.40	11.8	6.50	8.57	10.5	6.65	6.32	8.74	8.59	12.2	11.0	7.08	6.85	12.9	11.5	6.54	6.30	12.2	10.7	9.65	8.13
2	9.74	10.9	8.82	10.4	11.1	8.10	10.9	10.2	11.4	11.2	8.74	8.13	7.18	6.99	12.7	11.2	6.68	6.47	11.9	10.4	10.4	8.71
3	8.34	5.97	10.0	9.14	6.06	9.34	9.24	8.49	6.38	6.29	10.1	9.44	9.04	8.87	7.57	7.36	8.23	8.06	7.23	6.92	8.96	7.73
4	10.0	11.1	14.5	10.8	11.2	13.9	11.2	10.4	11.6	11.6	14.8	14.3	8.53	8.50	7.64	7.06	7.95	7.92	7.64	7.00	9.00	7.65
5	6.37	10.5	13.3	6.63	10.4	12.1	6.78	6.62	11.0	11.0	13.8	12.7	9.73	9.75	12.8	11.6	8.84	8.85	12.0	10.8	6.62	5.82
6	8.75	6.09	9.51	9.49	6.23	8.98	9.98	9.26	6.61	6.47	9.82	9.32	9.32	9.20	11.3	10.2	8.54	8.40	10.6	9.51	6.29	5.70
7	9.69	8.18	14.4	10.4	8.25	13.1	10.9	10.2	8.45	8.36	14.6	13.4	9.47	9.37	8.70	7.89	8.65	8.57	8.55	7.62	9.86	8.34
8	8.78	10.5	12.8	9.38	10.4	11.8	9.85	9.28	10.8	10.9	13.1	12.1	7.48	7.63	13.2	12.0	6.86	7.01	12.6	11.3	8.96	7.55
9	9.32	11.5	14.4	9.94	11.5	13.5	10.3	9.72	11.9	11.9	14.8	13.9	8.75	8.68	10.3	9.14	8.12	8.04	9.91	8.75	5.84	5.32
10	8.64	10.3	10.5	9.24	10.4	9.62	9.71	9.14	10.7	10.6	10.7	9.86	6.55	6.42	8.94	8.65	6.12	5.97	8.58	8.17	6.11	5.63
11	9.11	8.41	8.01	9.85	8.43	7.36	10.4	9.66	8.65	8.66	7.95	7.33	7.83	7.84	12.6	11.7	7.15	7.16	11.8	10.9	9.04	7.69
12	9.26	9.57	8.08	10.0	9.76	7.65	10.4	9.67	10.2	9.96	8.34	7.94	6.07	5.90	10.2	9.55	5.72	5.52	9.59	8.94	8.49	7.43
13	6.58	10.3	9.65	7.00	10.5	9.80	7.27	6.93	11.0	10.8	9.87	10.1	8.78	8.66	7.12	6.94	8.07	7.95	6.85	6.60	8.61	7.43
14	6.65	10.0	9.47	7.00	10.0	9.50	7.23	6.95	10.4	10.3	9.73	9.78	9.34	9.31	8.18	8.16	8.54	8.51	7.72	7.61	8.83	7.59
15	7.97	9.19	15.6	8.36	9.08	14.1	8.79	8.45	9.69	9.82	16.1	14.6	9.17	9.25	12.8	11.4	8.33	8.41	12.2	10.7	9.63	7.98
16	9.46	8.67	8.83	10.2	8.75	8.07	10.9	10.2	9.07	8.98	8.75	8.08	9.23	9.21	12.5	11.5	8.38	8.36	11.7	10.6	7.13	6.16

famille	Md23	Md24	Md25	Md26	Md27	Md28	Md29	Md30	Md31	Md32	Md33	Md34	Md35	Md36	Md37	Md38	Md39	Md40	Md41	Md42	Md43	Md44
1	9.67	8.14	4.10	5.73	6.00	6.59	3.63	4.83	6.18	5.47	3.61	5.55	5.35	6.75	3.86	5.53	6.32	6.77	5.48	6.60	6.52	6.58
2	10.3	8.67	3.79	5.93	6.30	7.20	3.56	4.12	5.55	5.68	3.37	5.35	5.49	7.27	3.82	5.68	7.10	8.63	5.55	6.56	7.03	8.44
3	8.80	7.59	4.33	6.98	9.50	11.9	3.98	6.54	9.01	5.23	3.84	6.29	6.18	8.65	4.04	6.00	8.75	11.0	5.90	8.42	11.1	13.3
4	8.89	7.51	3.93	6.70	9.31	11.7	4.38	7.13	9.59	5.66	4.23	6.80	5.32	7.19	4.12	6.21	8.79	10.4	5.94	8.81	11.3	12.7
5	6.48	5.62	3.71	4.78	6.17	7.16	3.51	5.61	6.87	5.54	3.92	5.70	5.48	6.66	3.88	5.43	6.39	6.56	5.41	6.88	7.54	7.58
6	6.20	5.59	4.13	6.59	8.91	10.4	3.74	6.09	7.72	5.44	3.91	5.75	6.08	8.41	3.92	5.96	8.46	10.6	5.31	7.33	9.73	11.7
7	9.84	8.33	3.67	5.28	7.63	9.62	3.41	5.80	7.86	5.56	3.38	5.27	5.87	7.60	3.84	6.02	8.05	9.36	5.63	7.76	9.97	11.1
8	8.78	7.35	3.58	4.70	6.13	7.45	3.83	6.42	8.34	5.69	4.52	6.83	5.65	6.72	4.00	4.62	5.10	4.95	6.00	7.88	7.94	7.11
9	5.81	5.19	3.71	6.14	8.33	9.58	4.22	6.58	7.90	5.70	3.80	5.55	5.26	6.84	3.91	6.03	8.24	9.49	5.21	7.11	9.28	10.5
10	6.13	5.59	3.90	5.70	6.20	6.32	3.78	4.99	5.95	5.64	3.84	6.01	5.57	7.09	4.23	5.64	6.69	7.54	5.07	5.62	5.54	6.26
11	8.92	7.57	4.10	6.51	8.69	10.6	3.75	6.17	8.45	5.52	4.43	6.84	6.15	8.41	4.22	4.67	6.54	8.94	6.27	8.38	9.99	12.2
12	8.43	7.38	4.15	6.68	8.00	9.31	4.44	6.23	8.01	5.35	3.77	6.02	6.20	8.48	3.94	5.48	7.83	10.1	5.61	6.99	8.76	10.9
13	8.47	7.29	4.12	6.54	8.61	10.9	4.20	6.50	8.92	5.71	3.82	6.28	4.83	6.13	4.05	6.18	8.19	9.13	5.88	8.38	10.2	10.9
14	8.72	7.48	3.54	4.36	5.86	7.89	3.50	5.50	7.84	5.57	3.74	6.10	5.59	6.70	4.10	5.56	6.61	6.70	5.97	7.93	8.72	8.52
15	9.43	7.77	4.14	6.42	9.19	11.2	3.83	6.73	8.76	5.58	4.42	6.72	5.41	7.23	3.94	5.74	7.87	9.41	5.92	8.69	10.7	12.1
16	7.04	6.02	3.85	5.68	6.57	7.23	3.34	4.47	5.55	5.60	4.02	5.96	6.02	7.97	4.00	5.67	7.18	8.99	5.25	6.18	6.74	8.38

Nous avons alors plusieurs éléments nous permettant de dégager la conformation type dans chaque famille. Nous résumons ces caractéristiques dans le tableau suivant.

Tableau 24. Résumé des caractéristiques des 16 familles.

fam.	Chaîne principale	Chaînes latérales	Distances les plus représentatives trouvées par l'équation 1	Distances les plus représentatives trouvées par l'équation 2
1	repliée	repliées	**D15** : 12.90 **D16** : 11.50 **D20** : 10.70	**D1** : 6.10 **D4** : 6.50 **D8** : 6.32
2	étendue	repliées	**D15** : 12.70 **D16** : 11.20 **D19** : 11.90 **D21** : 10.40 **D23** : 10.30 **D38** : 5.68	**D30** : 4.12 **D31** : 5.55
3	repliée	étendues	**D27** : 9.50 **D28** : 11.90 **D40** : 11.00 **D43** : 11.10 **D44** : 13.30	**D2** : 5.97 **D9** : 6.38 **D10** : 6.29 **D44** : 13.30
4	étendue	étendues	**D7** : 11.20 **D8** : 10.40 **D28** : 11.70 **D39** : 8.79 **D43** : 11.30	**D3** : 14.50 **D6** : 13.90 **D12** : 14.30 **D28** : 11.70 **D43** : 11.30
5	étendue	repliées	**D3** : 13.30 **D6** : 12.10 **D11** : 13.80 **D12** : 12.70 **D15** : 12.80 **D19** : 12.00	**D4** : 6.63 **D7** : 6.78
6	repliée		**D43** : 9.73 **D44** : 11.70	**D2** : 6.09 **D5** : 6.23 **D9** : 6.61 **D10** : 6.47
7	étendue	étendues	**D12** : 13.40 **D14** : 9.37 **D28** : 9.62 **D30** : 5.80 **D38** : 6.02 **D39** : 8.05 **D43** : 9.97 **D44** : 11.10	**D3** : 14.40 **D11** : 13.80
8	étendue	repliées	**D16** : 12.00 **D20** : 11.30	**D39** : 5.10 **D40** : 4.95 **D44** : 7.11

Tableau 25. Suite résumé des caractéristiques des 16 familles.

fam.	Chaîne principale	Chaînes latérales	Distances les plus représentatives trouvées par l'équation 1	Distances les plus représentatives trouvées par l'équation 2
9	étendue	repliées	**D3** : 14.40 **D6** : 13.50 **D11** : 14.80 **D12** : 13.90	**D6** : 13.50 **D11** : 14.80 **D12** :13.90 **D21** : 5.84 **D23** : 5.81
10	repliée	repliées	**D6** : 9.62 **D38** : 5.64 **D43** : 5.54	**D43** : 5.54 **D44** : 6.26
11		étendues	**D27** : 8.69 **D28** : 10.60 **D40** : 8.94 **D42** : 8.38 **D43** : 9.99	**D3** : 8.01 **D6** : 7.36 **D11** : 7.95
12	repliée	étendues	**D27** : 8.00 **D28** : 9.31	**D3** : 8.08 **D13** : 6.07 **D14** : 5.90 **D18** : 5.52
13	repliée	étendues	**D39** : 8.19 **D43** : 10.20	**D15** : 7.12 **D19** : 6.85
14	repliée	repliées	**D23** : 8.72	**D26** : 4.36
15	étendue	étendues	**D11** : 8.75 **D12** : 8.08 **D15** : 12.50 **D19** : 11.70 **D43** : 10.70	**D3** : 15.60 **D11** : 16.10 **D12** : 14.60
16		repliées	**D16** : 11.50 **D35** : 6.02	**D31** : 5.55 **D43** : 6.74

Nous avons une tendance générale des molécules dans chaque famille à présenter une chaîne principale plus ou moins étendue et des chaînes latérales avec plus ou moins d'interactions entre elles. Nous présentons en annexe C une molécule type pour chaque famille qui n'est pas

celle dont l'énergie est la plus basse mais plutôt celle dont les distances caractéristiques sont proches de la moyenne des distances pour cette famille c'est à dire le centre de masse de la famille.

Si nous examinons le tableau où nous avons résumé les caractéristiques de chaque famille, nous pouvons compiler quelles sont les distances qui caractérisent le plus souvent ces familles. Nous pouvons ainsi comparer ces résultats avec ceux de l'analyse en composantes principale où nous avons classé les distances par ordre d'importance en fonction de leurs participations respectives à la construction des composantes principales (voir page 96). Les distances classées par ordre de fréquence d'apparition dans le tableau résumé sont: D43, D3, D11, D12, D6, D28, D44, D15, D16, D19, D39, D23, D27, D38, D40. Ces observations recoupent les résultats de l'analyse en composantes principale puisque D28, D43 et D44 constituaient les distances majoritaires dans la construction de la première composante principales, D3, D6, D11, et D12 celles de la deuxième composante principale et D15, D16, D19 celles de la troisième composante principale.

3.2.4.4 Résultats de la classification pour gdmap

Nous avons procédé pour regarder les résultats de la classification par FASTCLUS de la même manière pour gdmap que nous l'avons fait pour glmap. Nous avions décidé suite aux indices révélés par les graphes du Pseudo F et du CCC de faire la classification de l'échantillon en 13 familles. Nous avons alors calculé les moyennes et écart-types des distances initiales à l'intérieur de chacune des 13 familles puis nous avons calculé l'équation [6] de manière à déterminer quelles distances possèdent la plus étroite distribution à l'intérieur de chaque famille. Cela nous indique quelles sont les distances caractéristiques qui déterminent les particularités conformationnelles dans chaque famille. Nous présentons le résultat de ce calcul pour les 13 familles dans le tableau suivant.

Tableau 26. Différences entre les écart-type dans les 13 familles et dans la population totale pour les 44 distances caractéristiques de gdmap.

famille	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12	d13	d14	d15	d16	d17	d18	d19	d20	d21	d22
1	2.	23.	36.	31.	20.	34.	25.	0.	4.	2.	-12.	-17.	-3.	23.	46.	33.	-5.	8.	38.	36.	17.	35.
2	15.	8.	29.	36.	5.	37.	33.	14.	18.	6.	8.	4.	0.	34.	45.	28.	7.	20.	45.	36.	32.	14.
3	-1.	32.	32.	28.	11.	8.	14.	3.	6.	5.	25.	16.	1.	3.	25.	35.	7.	13.	27.	34.	32.	17.
4	2.	4.	24.	28.	6.	1.	3.	0.	-2.	-6.	12.	14.	-2.	11.	16.	17.	5.	11.	19.	21.	23.	37.
5	10.	10.	23.	23.	17.	13.	1.	7.	-3.	-10.	18.	13.	1.	2.	10.	12.	4.	11.	16.	17.	-3.	41.
6	3.	8.	15.	33.	-2.	13.	31.	10.	35.	30.	15.	5.	15.	47.	34.	9.	33.	51.	47.	19.	25.	26.
7	10.	3.	24.	32.	17.	15.	12.	4.	6.	2.	8.	7.	0.	33.	24.	10.	13.	17.	31.	30.	29.	31.
8	1.	6.	18.	29.	0.	-3.	11.	0.	-10.	-4.	5.	0.	5.	22.	23.	13.	8.	23.	33.	24.	16.	26.
9	3.	3.	15.	21.	-2.	16.	18.	-6.	9.	12.	-3.	4.	2.	19.	13.	10.	3.	15.	26.	21.	23.	32.
10	6.	25.	37.	38.	0.	27.	30.	-10.	11.	5.	-12.	-2.	-2.	-8.	25.	34.	8.	28.	49.	44.	0.	12.
11	-1.	29.	40.	44.	-2.	25.	34.	-1.	6.	16.	2.	5.	1.	18.	41.	43.	12.	31.	50.	46.	30.	19.
12	6.	30.	35.	30.	11.	5.	11.	-6.	13.	9.	-14.	7.	3.	32.	34.	37.	15.	20.	33.	34.	18.	15.
13	1.	15.	25.	22.	8.	13.	6.	6.	-2.	-11.	11.	7.	7.	13.	32.	34.	3.	16.	26.	37.	24.	19.

famille	d23	d24	d25	d26	d27	d28	d29	d30	d31	d32	d33	d34	d35	d36	d37	d38	d39	d40	d41	d42	d43	d44
1	24.	14.	34.	34.	24.	29.	34.	35.	28.	37.	27.	24.	18.	19.	21.	20.	16.	19.	3.	7.	6.	8.
2	32.	34.	14.	38.	35.	35.	13.	13.	36.	43.	34.	38.	25.	24.	27.	31.	24.	27.	29.	28.	30.	28.
3	23.	34.	16.	17.	29.	25.	15.	17.	23.	21.	19.	17.	39.	43.	18.	17.	36.	40.	34.	26.	31.	22.
4	36.	20.	35.	35.	23.	24.	31.	29.	36.	36.	6.	5.	41.	42.	9.	7.	40.	40.	23.	22.	23.	24.
5	38.	0.	40.	38.	2.	0.	36.	38.	39.	39.	35.	35.	37.	35.	39.	40.	37.	35.	4.	13.	7.	16.
6	10.	27.	25.	14.	26.	25.	22.	22.	8.	11.	3.	1.	25.	33.	5.	3.	19.	27.	35.	29.	34.	28.
7	27.	28.	32.	25.	31.	32.	29.	28.	18.	16.	3.	3.	31.	35.	6.	3.	30.	33.	35.	27.	31.	24.
8	25.	17.	28.	27.	17.	17.	22.	21.	24.	24.	13.	12.	23.	23.	11.	11.	24.	24.	25.	22.	26.	18.
9	23.	18.	33.	33.	19.	16.	29.	30.	23.	29.	24.	26.	34.	30.	21.	26.	33.	32.	28.	26.	27.	25.
10	39.	1.	13.	36.	1.	0.	13.	13.	38.	35.	24.	25.	40.	32.	21.	22.	45.	36.	26.	19.	21.	15.
11	26.	31.	20.	26.	26.	24.	18.	18.	22.	20.	30.	32.	33.	32.	25.	26.	30.	34.	34.	29.	34.	28.
12	33.	19.	15.	28.	17.	16.	13.	13.	28.	24.	16.	16.	26.	28.	16.	18.	22.	23.	22.	21.	24.	25.
13	24.	25.	17.	22.	19.	18.	15.	17.	25.	22.	28.	33.	43.	38.	25.	29.	42.	37.	26.	23.	25.	22.

Nous avons souligné les valeurs les plus élevées dans ce tableau. Elles correspondent aux distances pour lesquelles l'écart-type est beaucoup plus étroit dans la famille concernée que dans l'échantillon entier. Ces distances seront caractéristiques de toutes les conformations de cette famille. Nous notons ici une certaine différence d'avec ce que nous observons pour glmap. En effet, le maximum de pourcentage de différence obtenue est en général plus élevé que ce que nous obtenions pour glmap et nous atteignons ici plus souvent des pourcentages de différence jusqu'à 50%. Or, nous avons ici un nombre de familles inférieur à ce que nous avions pour glmap (13 au lieu de 16). Plus le nombre de familles est élevé, plus la séparation entre les familles est bonne (si bien entendu les nombres de familles ont étés suggérés par les indices statistiques) ce qui devrait se traduire pour glmap par des écart-types plus faibles sur

les distances caractéristiques donc des pourcentages de différence plus grands. Comme nous le constatons, il n'en est rien et cela ne peut s'expliquer que par une différence fondamentale dans les populations des deux analogues: gdmap est un peptide plus contraint conformationnellement et, par conséquent, possède moins de conformations possibles. Un nombre de familles plus restreint peut ainsi mieux décrire toutes les possibilités conformationnelles de ce peptide. Nous avons également calculé la quantité équivalente sur les moyennes correspondant à l'équation [7] dont nous présentons le résultat dans le tableau suivant.

Tableau 27. Différences entre les moyennes dans les 13 familles et dans la population totale pour les 44 distances caractéristiques de gdmap.

famille	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12	d13	d14	d15	d16	d17	d18	d19	d20	d21	d22
1	-1.	-14.	-20.	-21.	-15.	-18.	-16.	-1.	-4.	-4.	8.	5.	0.	-10.	-19.	-19.	1.	-8.	-19.	-20.	-11.	-26.
2	11.	15.	3.	-5.	4.	-5.	-9.	-1.	10.	7.	2.	7.	2.	-6.	-8.	-2.	0.	-5.	-12.	-7.	-15.	9.
3	-5.	-9.	-7.	-3.	10.	9.	9.	-3.	-7.	-4.	-5.	-9.	4.	6.	4.	-8.	0.	2.	2.	-8.	-6.	26.
4	-2.	10.	23.	27.	4.	12.	14.	3.	2.	3.	-1.	2.	3.	5.	15.	21.	2.	9.	23.	32.	28.	1.
5	7.	22.	21.	24.	7.	4.	11.	2.	-2.	4.	-1.	5.	4.	7.	16.	26.	3.	7.	16.	26.	13.	-17.
6	4.	7.	3.	-4.	-4.	-11.	-17.	-2.	-13.	-15.	-2.	2.	-11.	20.	27.	29.	-12.	-13.	-3.	5.	-11.	-11.
7	7.	15.	30.	32.	8.	26.	26.	1.	8.	9.	2.	7.	4.	2.	14.	21.	2.	12.	27.	32.	-13.	-16.
8	1.	5.	19.	32.	1.	12.	21.	0.	-1.	0.	-1.	0.	-3.	2.	11.	14.	10.	24.	38.	39.	-9.	-12.
9	-4.	0.	-2.	-8.	-3.	-4.	-9.	4.	5.	1.	-1.	-1.	-1.	-5.	-7.	-4.	-3.	-6.	-10.	-10.	28.	-11.
10	-9.	-13.	-24.	-25.	-5.	-18.	-16.	2.	-11.	-8.	0.	-5.	1.	-1.	-11.	-16.	-4.	-14.	-24.	-28.	8.	0.
11	-5.	-14.	-22.	-28.	-1.	-9.	-15.	-2.	-1.	-5.	0.	-4.	-2.	-6.	-14.	-21.	-6.	-15.	-25.	-31.	-1.	32.
12	-6.	-17.	-4.	-2.	-12.	3.	3.	0.	12.	7.	-1.	-7.	4.	0.	-7.	-15.	-1.	5.	2.	-7.	-2.	-2.
13	-1.	-4.	-7.	-4.	7.	5.	8.	-2.	0.	5.	-2.	-4.	-2.	-5.	-8.	-13.	8.	9.	1.	-6.	-5.	25.

famille	d23	d24	d25	d26	d27	d28	d29	d30	d31	d32	d33	d34	d35	d36	d37	d38	d39	d40	d41	d42	d43	d44
1	-26.	-10.	-25.	-29.	-11.	-11.	-23.	-25.	-26.	-30.	-4.	-3.	17.	18.	-5.	-4.	15.	16.	13.	12.	13.	12.
2	-30.	-16.	8.	-28.	-15.	-15.	9.	10.	-31.	-28.	-9.	-9.	20.	21.	-9.	-9.	16.	18.	-18.	-16.	-19.	-17.
3	16.	-7.	26.	22.	-8.	-7.	25.	25.	18.	23.	-5.	-5.	-29.	-25.	-4.	-4.	-27.	-24.	-12.	-7.	-12.	-8.
4	-14.	29.	1.	-11.	32.	31.	3.	3.	-15.	-11.	13.	14.	-25.	-21.	13.	15.	-24.	-20.	-16.	-13.	-16.	-14.
5	-18.	15.	-16.	-16.	15.	13.	-17.	-17.	-19.	-17.	-16.	-18.	-21.	-20.	-15.	-18.	-19.	-18.	19.	18.	21.	21.
6	16.	-11.	-11.	12.	-11.	-10.	-9.	-9.	16.	13.	17.	15.	-11.	-14.	15.	14.	-10.	-13.	-6.	-6.	-4.	-4.
7	16.	-12.	-17.	17.	-14.	-14.	-16.	-15.	18.	18.	4.	4.	-20.	-17.	4.	5.	-19.	-16.	-17.	-13.	-17.	-14.
8	2.	-10.	-13.	4.	-9.	-9.	-11.	-11.	2.	5.	20.	21.	10.	6.	19.	20.	9.	5.	27.	21.	26.	21.
9	19.	29.	-10.	11.	30.	29.	-9.	-9.	20.	13.	-8.	-9.	30.	24.	-8.	-8.	30.	25.	-1.	-2.	-1.	-2.
10	-37.	10.	2.	-32.	11.	9.	0.	-1.	-38.	-33.	-8.	-9.	-19.	-15.	-7.	-8.	-19.	-15.	-11.	-7.	-11.	-7.
11	13.	-2.	32.	10.	-3.	-2.	29.	29.	11.	9.	-8.	-7.	32.	26.	-7.	-7.	32.	26.	-2.	-4.	-2.	-4.
12	29.	-3.	-3.	29.	-6.	-5.	-5.	-5.	30.	29.	26.	28.	-2.	-1.	24.	26.	0.	0.	-10.	-10.	-11.	-11.
13	19.	-6.	25.	16.	-6.	-5.	23.	23.	18.	16.	-11.	-11.	-8.	-7.	-10.	-10.	-6.	-6.	33.	28.	33.	28.

Nous avons indiqué les distances pour lesquelles la moyenne dans la famille est très différente de la moyenne dans la population totale. Si le pourcentage de différence est positif, la famiile

contient des molécules en moyenne plus repliées que l'ensemble de la population. Si le pourcentage de différence est négatif, la famille contient des molécules en moyenne plus étendues que l'ensemble de la population.

Nous voyons ici apparaître des différences d'avec ce qui est observé pour glmap: les pourcentages de différence sont généralement plus faibles et pour les distances relatives à la chaîne principale, ces pourcentages sont très peu élevés. Les distances reliées à la chaîne principale sont donc très peu utiles dans la classification pour trouver les caractéristiques conformationnelles de chaque famille à l'exception des distances D4, D16, D19 et D20 pour lesquelles les pourcentages de différences sont comparables à ceux observés pour les distances relatives aux chaînes latérales. Les caractéristiques conformationnelles dans chaque famille seront alors essentiellement relatives aux positions des chaînes latérales. Ceci contredit les résultat de l'ACP où nous avions constaté qu'une part sensiblement égale de la variabilité dans l'échantillon était reliée d'une part à la chaîne principale et d'autre part aux chaînes latérales. Cela indique que les différences entre les familles sont essentiellement dues aux positions des chaînes latérales où les pourcentages de différences sont plus élevés.

Nous présentons le tableau où nous avons additionné les différences négatives d'une part (correspond aux distances dont les moyennes sont supérieures dans la famille concernée à la moyenne dans la population totale) et les différences positives d'autre part (correspond aux distances dont les moyennes sont inférieures dans la famille concernée à la moyenne dans la population totale) pour chaque type de distances soit les 20 premières décrivant la chaîne principale et les 24 suivantes décrivant les chaînes latérales. Ceci nous indique l'extension et le repliement relatif de la chaîne principale d'une part et les interactions rapprochées ou non entre les chaînes latérales.

Nous présentons dans le tableau 29 les valeurs des moyennes des 44 distances dans chacune des 13 familles trouvées après la classification par FASTCLUS

Tableau 28. Somme des résultats de l'équation (μ_{tot}-$\mu_{fam.}$/ μ_{tot}) positifs d'une part et négatifs d'autre part pour chaque type de distance dans chaque famille de gdmap.

Famille	Chaîne principale (D1 à D20)		Chaînes latérales (D21 à D44)	
	$\mu_{fam.} < \mu_{tot.}$	$\mu_{fam.} > \mu_{tot.}$	$\mu_{fam.} < \mu_{tot.}$	$\mu_{fam.} > \mu_{tot.}$
1	14	-210	116	-269
2	60	-61	111	-284
3	45	-69	180	-190
4	208	-3	182	-202
5	210	-3	135	-283
6	97	-107	118	-151
7	287	0	88	-251
8	229	-6	219	-83
9	10	-76	288	-78
10	3	-232	40	-276
11	0	-225	282	-48
12	35	-80	221	-76
13	43	-58	288	-90

Tableau 29. Moyennes par familles pour les 44 distances caractéristiques de gdmap.

famille	Md1	Md2	Md3	Md4	Md5	Md6	Md7	Md8	Md9	Md10	Md11	Md12	Md13	Md14	Md15	Md16	Md17	Md18	Md19	Md20	Md21	Md22
1	3.94	6.68	9.07	11.0	4.38	6.93	8.90	5.64	4.07	6.31	5.18	6.96	4.00	6.25	8.75	10.2	5.58	8.08	10.5	11.8	9.39	11.7
2	3.48	5.00	7.32	9.48	3.66	6.15	8.37	5.63	3.50	5.65	5.54	6.84	3.93	6.03	7.96	8.77	5.66	7.91	9.90	10.5	9.72	8.50
3	4.11	6.39	8.10	9.37	3.44	5.31	7.02	5.71	4.19	6.31	5.92	7.99	3.83	5.35	7.09	9.25	5.67	7.36	8.63	10.6	8.95	6.91
4	3.90	5.32	5.85	6.65	3.65	5.15	6.62	5.38	3.83	5.87	5.68	7.21	3.89	5.39	6.26	6.76	5.54	6.80	6.80	6.71	6.05	9.17
5	3.63	4.59	5.95	6.92	3.55	5.63	6.87	5.47	3.99	5.80	5.69	6.98	3.84	5.28	6.17	6.37	5.50	7.00	7.43	7.25	7.32	10.9
6	3.73	5.49	7.36	9.39	3.97	6.52	8.97	5.68	4.41	6.95	5.75	7.17	4.43	4.51	5.40	6.14	6.34	8.46	9.11	9.30	9.37	10.3
7	3.63	4.98	5.29	6.13	3.51	4.34	5.68	5.49	3.58	5.52	5.51	6.84	3.82	5.54	6.32	6.79	5.57	6.60	6.40	6.62	9.59	10.8
8	3.85	5.60	6.14	6.12	3.79	5.13	6.08	5.58	3.93	6.07	5.71	7.34	4.12	5.53	6.55	7.40	5.10	5.74	5.50	5.96	9.26	10.4
9	4.06	5.89	7.74	9.79	3.92	6.07	8.34	5.36	3.70	5.99	5.67	7.41	4.03	5.93	7.83	8.94	5.82	7.93	9.74	10.8	6.12	10.3
10	4.25	6.66	9.41	11.3	4.01	6.90	8.90	5.46	4.34	6.56	5.64	7.71	3.95	5.73	8.13	9.98	5.89	8.57	10.9	12.5	7.76	9.26
11	4.09	6.70	9.24	11.6	3.84	6.38	8.85	5.68	3.93	6.35	5.64	7.65	4.07	6.00	8.41	10.4	6.00	8.62	11.0	12.0	8.52	6.29
12	4.15	6.87	7.88	9.21	4.28	5.69	7.47	5.57	3.45	5.60	5.68	7.86	3.83	5.69	7.86	9.90	5.71	7.13	8.66	10.5	8.65	9.48
13	3.94	6.12	8.13	9.44	3.54	5.58	7.10	5.66	3.89	5.75	5.72	7.59	4.08	5.94	7.95	9.69	5.19	6.87	8.74	10.4	8.87	6.99

famille	Md23	Md24	Md25	Md26	Md27	Md28	Md29	Md30	Md31	Md32	Md33	Md34	Md35	Md36	Md37	Md38	Md39	Md40	Md41	Md42	Md43	Md44
1	14.1	10.0	11.7	13.5	10.5	9.85	12.0	12.1	14.4	13.9	8.70	8.58	8.68	7.88	8.08	7.97	8.46	7.63	7.32	6.39	7.26	6.26
2	14.6	10.5	8.60	13.4	10.9	10.2	8.83	8.73	14.9	13.7	9.17	9.12	8.44	7.59	8.42	8.37	8.39	7.43	9.08	8.38	9.86	8.34
3	9.41	9.69	6.94	8.14	10.2	9.48	7.29	7.27	9.41	8.24	8.83	8.76	13.5	12.1	8.00	7.94	12.7	11.3	9.39	7.77	9.34	7.70
4	12.8	6.44	9.30	11.6	6.48	6.13	9.48	9.34	13.1	11.9	7.30	7.18	13.1	11.7	6.68	6.54	12.4	10.9	9.73	8.20	9.68	8.15
5	13.2	7.70	10.9	12.1	8.01	7.69	11.4	11.3	13.6	12.5	9.78	9.88	12.7	11.6	8.90	8.99	11.9	10.7	6.79	5.94	6.58	5.66
6	9.40	10.1	10.4	9.18	10.5	9.74	10.6	10.5	9.56	9.30	6.99	7.08	11.7	11.0	6.54	6.61	11.0	10.3	8.89	7.64	8.65	7.39
7	9.36	10.2	11.0	8.67	10.8	10.1	11.3	11.1	9.34	8.75	8.03	7.97	12.6	11.3	7.37	7.30	11.9	10.5	9.60	8.18	9.75	8.13
8	11.0	9.97	10.6	9.99	10.3	9.66	10.8	10.7	11.2	10.2	6.68	6.57	9.44	9.10	6.27	6.13	9.05	8.60	6.17	5.69	6.15	5.60
9	9.11	6.42	10.3	9.29	6.60	6.34	10.6	10.5	9.12	9.31	9.11	9.06	7.35	7.36	8.33	8.28	6.94	6.85	8.49	7.40	8.38	7.26
10	15.3	8.12	9.20	13.8	8.44	8.12	9.72	9.80	15.7	14.2	9.09	9.09	12.5	11.1	8.23	8.25	11.9	10.4	9.32	7.76	9.21	7.63
11	9.78	9.26	6.38	9.41	9.78	9.06	6.89	6.82	10.1	9.75	9.03	8.90	7.15	7.12	8.28	8.15	6.80	6.71	8.57	7.52	8.45	7.41
12	7.90	9.36	9.66	7.44	10.0	9.36	10.2	10.1	7.97	7.60	6.23	6.02	10.7	9.76	5.88	5.64	10.0	9.08	9.27	7.99	9.20	7.92
13	9.09	9.57	7.04	8.74	10.0	9.37	7.46	7.40	9.33	9.03	9.31	9.25	11.3	10.3	8.49	8.43	10.6	9.60	5.61	5.21	5.56	5.12

Tableau 30. Résumé des caractéristiques des 13 familles.

fam.	Chaîne principale	Chaînes latérales	Distances les plus représentatives trouvées par l'équation 1	Distances les plus représentatives trouvées par l'équation 2
1	étendue	étendues	D3 : 9.07 D22 : 11.70 D15 : 8.75 D30 : 12.10 D19 : 10.50 D32 : 14.40 D20 : 11.80	D26 : 13.50 D32 : 13.90
2		étendues	D4 : 9.48 D27 : 10.90 D6 : 6.15 D28 : 10.20 D15 : 7.96 D31 : 14.90 D19 : 9.90 D32 : 13.70 D20 : 10.50 D34 : 9.12 D26 : 13.40	D23 : 14.60 D26 : 13.40 D31 : 14.90 D32 : 13.70
3		étendues	D16 : 9.25 d36 : 12.10 D39 : 12.70 D40 : 11.30	D35 : 13.50
4	repliée	étendues	D22 : 9.17 D32 : 11.90 D23 : 12.80 D35 : 13.10 D25 : 9.30 D39 : 12.40 D26 : 11.60 D40 : 10.90 D31 : 13.10	D20 : 7.25 D24 : 6.44
5	repliée	étendues	D22 : 10.90 D33 : 9.78 D23 : 13.20 D34 : 9.88 D25 : 10.90 D35 : 12.70 D26 : 12.10 D36 : 11.70 D29 : 11.40 D37 : 8.90 D30 : 11.30 D38 : 8.99 D31 : 13.60 D39 : 11.90 D32 : 12.50 D40 : 10.70	D20 : 7.25
6			D9 : 4.41 D14 : 4.51 D18 : 8.46 D19 : 9.11 D41 : 8.89	D16 : 6.14
7	repliée	étendues	D35 : 11.30 D41 : 9.80	D3 : 5.29 D4 : 6.12 D20 : 6.62
8	repliée	repliées	D4 : 6.12 D19 : 5.50 D20 : 5.96	D4 : 6.12 D19 : 5.50 D20 : 5.96

Tableau 31. Suite résumé des caractéristiques des 13 familles.

fam.	Chaîne principale	Chaînes latérales	Distances les plus représentatives trouvées par l'équation 1	Distances les plus représentatives trouvées par l'équation 2
9		repliées	**D25** : 10.30 **D26** : 9.29 **D35** : 7.35 **D39** : 6.94	**D21** : 6.12 **D24** : 6.42 **D28** : 6.34 **D35** : 7.35 **D39** : 6.94
10	étendue	étendues	**D3** : 9.41 **D26** : 13.80 **D4** : 11.30 **D31** : 15.71 **D19** : 10.90 **D35** : 12.50 **D20** : 12.50 **D39** : 11.90 **D23** : 15.30 **D40** : 10.40	**D20** : 12.50 **D23** : 15.30 **D26** : 13.80 **D31** : 15.70 **D32** : 14.20
11	étendue	repliées	**D3** : 9.24 **D19** : 11.00 **D4** : 11.60 **D20** : 12.80 **D15** : 8.41 **D21** : 8.52 D16 : 10.40	**D20** : 12.80 **D30** : 6.82 **D22** : 6.29 **D35** : 7.15 **D25** : 6.38 **D39** : 6.80 **D29** : 6.89
12		repliées	**D3** : 7.88 **D16** : 9.90	**D23** : 7.90 **D26** : 7.44 **D31** : 7.97 **D32** : 7.60 **D34** : 6.02
13		repliées	**D20** : 10.40 **D35** : 11.30 **D36** : 10.30 **D39** : 10.60 **D40** : 9.60	**D41** : 5.61 **D42** : 5.21 **D43** : 5.56 **D44** : 5.12

Nous avons résumé comme pour glmap, toutes les caractéristiques des 13 familles dans les tableaux 30 et 31.

Nous avons, pour chaque famille, recherché les molécules dont les distances caractéristiques déterminées par l'étude des tableaux précédents sont les plus proches des moyennes dans cette famille. Nous présentons l'une de ces molécules pour chaque famille en annexe C.

Nous avons compilé sur les 13 familles de gdmap le nombre de fois où apparaît chaque

distance en tant que distance caractéristique. Les distances classées par ordre de fréquence d'apparition décroissante sont D20, D26, D35, D39, D19, D31, D32, D4, D23, D3, D40, D22, D16, D36, D25, D15, D34 et D41. Nous retrouvons ici les distances les plus importantes notées après examen des premières composantes principales dans l'étude par ACP. En effet, D20, D19 et D4 composaient en majorité la première composante principale et D26, D31, D32 et D23 participaient à la construction de la deuxième composante principale. La troisième composante principale était composée des distances D41 à D44 et seule D41 apparaît ici 3 fois comme distance caractérisant un famille. Les trois autres n'apparaissent comme distances importantes que dans la description de la famille 13 avec des moyennes faibles traduisant une interaction rapprochée entre le cycle de la chaîne latérale de la Phe et la chaîne latérale de l'Asp. Les distances D35, D39, D40, D36 participaient à la formation de la quatrième composante principale. Les seules distances relatives à la chaîne principale qui sont ici notées comme importantes c'est à dire D4, 16, D19 et D20 traduisent respectivement des pseudo-cycles à 16, 14, 14 et 17 membres soit des interactions à longue distance, conduisant à un repliement de la molécule d'un bout à l'autre si les distances sont courtes ou une extension totale si ces dernières sont longues. De plus, D20 et D19 sont les distances classées respectivement 1 et 4 dans l'ordre de fréquence d'apparition comme distances caractéristiques. Ceci concilie les conclusions de l'ACP et les observations sur les résultats de l'équation [7] où nous avions noté le peu d'importance des distances relatives à la chaîne principale dans la discrimination entre les familles. En effet, si il y a peu de distances importantes, ces dernières caractérisent néanmoins de nombreuses familles.

3.2.5 Comparaison des deux analogues glmap et gdmap

Si nous résumons, en parallèle pour chaque peptide, les caractéristiques conformationnelles découvertes grâce à l'ACP puis au regroupement, nous pouvons déterminer les quelques différences essentielles entre les conformations de chacun de ces deux analogues.
En ce qui concerne les corrélations entre les distances, dans les deux cas nous n'observons

pas de corrélations significatives entre les distances décrivant la chaîne principale et celles décrivant les chaînes latérales. Lors de l'analyse en composantes principales, nous avons également noté que ces dernières étaient composées soit de distances reliées à la chaîne principale, soit de distances reliées aux chaînes latérales et ce pour les deux peptides. Néanmoins, grâce aux résultats de l'ACP, nous constatons que glmap est caractérisé par des interactions à plus courte distance sur la chaîne principale que gdmap. En effet, les premières composantes principales de glmap sont constituées majoritairement par des distances traduisant des interactions à courte distances centrées sur le résidu Met soit D3, D6, D11, D12, traduisant des pseudo C_{13}, C_{10}, C_8, C_{11}, ainsi que les distances traduisant les interactions entre l'Asp et la Phe alors que gdmap est caractérisé par D4, D19 et D20 qui traduisent respectivement des pseudo-cycles à 13, 14 et 17 membres. Quant aux interactions entre chaînes latérales, on note pour glmap une importance des interactions entre la chaîne latérale du Trp et celles de tous les autres résidus ainsi qu'entre les chaînes latérales de la Phe et de l'Asp. Pour gdmap, les interactions entre chaînes latérales sont importantes entre le Trp et la Phe, l'Asp et la Phe et entre la Met et l'Asp et la Phe.

Les observations résultat de l'étude de classification sont sensiblement les mêmes. Il faut noter dans l'examen des distances caractérisant chaque famille que si glmap est caractérisé à la fois par les positions de ses chaînes latérales et le repliement de sa chaîne principale, gdmap est essentiellement caractérisé par les interactions relatives aux chaînes latérales. Nous avons également observé la forte prédominance de 4 distances seulement dans la description de la chaîne principale qui toutes traduisent des interactions à longue distance c-à-d entre les résidus terminaux.

Ces observations s'expliquent aisément par la présence du D-Trp dans l'analogue gdmap. En effet, l'encombrement de la chaîne latérale du Trp empêche tout repliement à courte distance entre les résidus entourant le Trp et c'est pourquoi nous observons dans le cas de gdmap presqu'exclusivement un repliement impliquant les résidus terminaux du peptide. De plus, la

présence du D-Trp limite les possibilités conformationnelles du peptide ce qui se traduit par un nombre de familles inférieur pour la description de l'échantillon d'une part, et, d'autre part, peut expliquer pourquoi 4 des 20 distances initiales suffisent à la description de la conformation de la chaîne principale de ce peptide.

En conclusion, nous observons que glmap possède plus de possibilités conformationnelles que gdmap en ce qui concerne la chaîne principale. Pour ce qui est des chaînes latérales, les interactions ne se produisent pas entre les mêmes résidus pour les deux peptides puisque là encore, la chaîne latérale du résidu Trp n'est plus aussi accessible pour les chaînes latérales des autres résidus.

3.3 Remarques et conclusion

En comparaison avec les classification effectuées sur les molécules précédentes soit l'alanine et APYA, les résultats de la classification sont ici nettement moins bons. Nous pouvons dégager deux causes à cela: l'une inhérente à la méthode d'échantillonnage elle-même c-à-d le choix des angles variables et fixes, la taille de l'échantillon et sa complexité et l'autre relative à la classification comme telle c-à-d tout ce qui est relatif au choix des distances initiales, à l'interprétation des indices statistiques, à la méthode de classification choisie.
Nous pensons que l'un des problèmes lorsqu'on essaie d'étudier une molécule flexible et de grande taille est que nous devons générer un échantillon représentatif correct. Si il est relativement facile de le faire pour une molécule de taille restreinte comme celles étudiées précédemment, cela devient ici un travail délicat au point de vue théorique (déterminer la taille de l'échantillon) et pratique (temps de calcul mis en jeu). De plus, nous pensons avoir fait une erreur dans le choix de nos angles variables. En effet, nous avons noté dans l'étude de l'alanine que contraindre certains angles pouvait empêcher l'accès à certains minima en créant des barrières de rotations artificielles, or nous contraignons toujours nos angles ω à 180 degré ce qui reproduit la réalité observée dans les peptides mais peut néanmoins empêcher

la molécule de converger vers un minimum local lors du processus de minimisation (fait que nous avons observé directement pour l'étude de l'alanine par une dispersion des molécules autour des minima réels sur les cartes de Ramachandran lorsque certains angles avaient été contraint pendant la minimisation). Dans le cas d'un petit peptide comme l'alanine qui ne peut adopter de conformations très repliées, la restriction des angles de liaison peptidique ω ne gène pas la minimisation alors que pour un peptide de grande taille, elle peut empêcher la convergence même si dans la conformation finale, les angles ω sont effectivement proches de 180 degré (tel que généralement observé dans les peptides et protéines). Ceci a pour effet de compliquer artificiellement l'hypersurface conformationnelle de la molécule rendant par conséquent la classification plus difficile puisque les frontières entre classes sont moins bien définies.

L'autre point est relatif aux limitations des méthodes de classification. D'abord l'interprétation des indices statistiques dont nous avons déjà discuté. S'ils ne sont pas en eux-même plus compliqués à utiliser pour un grand que pour un petit échantillon, ils nuisent indirectement ici au succès de la classification. En effet, les indices statistiques suggérant généralement plus qu'un nombre optimal de familles, il est bon de procéder à plusieurs classifications avec chacun des nombres suggérés. Nous ne pouvons ici nous permettre de le faire puisque la classification complète et l'étude des résultats pour un tel échantillon est très longue. Le deuxième point est relatif au choix des distances caractérisant les conformations de la molécule. Lorsque nous étudions de petites molécules, nous pouvons nous permettre de choisir un ensemble de distances caractéristiques de grande taille. Pour une grosse molécule, le nombre de distances possibles devient très élevé et nous en privilégions probablement certaines à tort en les choisissant comme variables plutôt que d'autres ce qui biaise les résultats de la classification. Le troisième point est relatif au choix de la méthode de classification. Lors de nos précédentes études sur l'alanine et APYA, nous avons procédé à plusieurs classifications successives à partir de différents types de classification de manière à vérifier, par l'obtention d'un consensus entre les résultats, que notre classification était

correcte. Nous avons du ici procéder au choix d'une méthode en fonction de la faisabilité pratique de la classification et c'est pourquoi nous avons choisi FASTCLUS qui, étant une méthode de classification non-hiérarchique est apte à classer de très grands échantillons relativement rapidement. Nous ne pouvons comparer les résultats de classification avec ceux obtenus par une méthode hiérarchique comme il est de règle de le faire. Nous aurions également désiré tester une méthode de classification "floue" pour cet échantillon. En effet, nous avons ici une hypersurface conformationnelle extrêmement complexe ce qui entraîne l'existence d'une multitude de minima métastables qui sont très proches les uns des autres. L'échantillon des minima devient alors quasiment continu et les frontières entre les familles de conformations sont alors très délicates à tracer. Lors d'une classification floue, les individus sont affectés à une famille avec une certaine probabilité. Par suite, les individus situés aux frontières sont plus faciles à classer et ce type de classification serait plus adapté à ce type d'échantillon.

Néanmoins, nous pensons que toutes ces limitations ne signifient pas que ces méthodes sont inutilisables sur de très grands échantillons de conformations. En effet, la plupart des points soulevés ici ne représentent une limitation que par rapport à la capacité de calcul disponible. Il faut en revanche porter une attention particulière à la génération de l'échantillon de conformation qui peut indirectement provoquer des problèmes dans la classification subséquente.

3.4 Comparaison avec les conformations connues et activité biologique

Comme nous l'avons souligné au début de ce chapitre, CCK et tous les fragments dérivés existant à l'état naturel dans l'organisme exercent des rôles biologiques importants et diversifiés. Il existe deux cibles d'actions essentielles pour ces molécules ce qui se traduit par deux récepteurs potentiels notés CCK-B (pour "brain") largement réparti dans tout le CNS et CCK-A (pour "alimentary") présent essentiellement dans le système gastro-intestinal mais

également en faible quantité et très localisé, dans le CNS. Ces deux récepteurs de nature protéique ont récemment étés clonés et la séquence primaire est déterminée: CCK-B est constitué d'un enchaînement de 447 acides aminés chez l'humain (79) et CCK-A est constitué d'un enchaînement de 448 acides aminés chez le rat (80). Le fragment CCK-5 objet de notre étude est supposé agir plutôt sur le récepteur CCK-B. L'existence de ces deux récepteurs distincts à rendu assez difficile l'étude des relations structure-activité des molécules actives sur chaque récepteur d'autant que les différents fragments CCK sont tantôt sélectifs c-à-d actifs seulement sur l'un des récepteurs tantôt actifs sur les deux récepteurs. L'étude de l'activité des molécules biologiques est généralement un processus indirect. En effet, on peut rarement déterminer la conformation du récepteur in vivo si même on a la chance de connaître la séquence primaire. De plus, la taille du récepteur étant très élevée par rapport à celle de la molécule active, il est difficile de déterminer le site actif du récepteur c-à-d l'endroit où la molécule va s'attacher. On passe donc généralement par le biais de l'étude de familles d'agonistes et d'antagonistes de la molécule qui permettent de déterminer les caractéristiques structurales nécessaires et suffisantes à la reconnaissance de la molécule par le récepteur.

Les agonistes et antagonistes sont des molécules relativement rigides dont on connaît et contrôle la conformation spatiale. A partir d'une molécule possédant une certaine activité sur un récepteur, on effectue une série de modification chimique en mesurant à chaque fois l'activité de la nouvelle molécule. Ce type d'étude remplit deux objectifs: déterminer autant que possible les raisons structurales de l'activité de la molécule et développer un produit le plus actif possible qui permet de traiter certains désordres biologiques. L'étude de la molécule endogène généralement de nature peptidique donc flexible commence généralement par une rigidification qui réduit le nombre de degrés de libertés de la molécule via l'échange d'un résidu d'acide aminé par une proline par exemple. On peut ainsi acquérir des informations sur les pré-requis conformationnels nécessaires à la reconnaissance de la molécule endogène par le récepteur. Le cas des molécules CCK est complexe car pendant très longtemps, seuls des

agonistes et antagonistes de CCK-A étaient connus alors qu'on arrivait pas à développer de molécules ayant une activité spécifique sur CCK-B.

Notre étude a consisté à déterminer toutes les possibilités conformationnelles pour la molécule de CCK-5 et nous allons maintenant comparer les structures que nous proposons théoriquement avec celles proposées dans la littérature.

3.4.1 Comparaison avec des études publiées des fragments CCK

Les molécules agonistes de CCK sont plutôt de nature peptidique c-à-d des peptides plus ou moins modifiés et de grande taille c-à-d d'une taille comparable avec celle des fragments endogènes.

Le groupe de Taga a publié deux études en 1994 sur la conformation de la tétragastrine (soit le fragment CCK-4) dans le DMSO par la méthode Monte-Carlo en prenant en compte les effets de solvant. Ils ont effectué l'étude de ce fragment par RMN dans le même solvant. L'étude publiée dans la référence (81) donne les 7 conformations les plus basses en énergie notée a, b, c, d, e, f, g résultant de la stratégie de recherche de minima locaux d'énergie par méthode Monte-Carlo mis au point par les auteurs. Les auteurs notent que ces conformations sont consistantes avec les études expérimentales.

Ces 7 conformations ont été systématiquement comparées par superposition graphique avec chacune des 16 conformations types proposées dans notre étude du fragment CCK-5 ainsi qu'avec les 13 conformations types de son analogue d-Trp. Nous remarquons d'abord que les molécules proposées sont très proches conformationnellement ce qui montre que la stratégie d'exploration par méthode Monte-Carlo est peu efficace. Nous présentons sur la figure suivante une superposition des 7 conformations proposées par l'équipe de Taga:

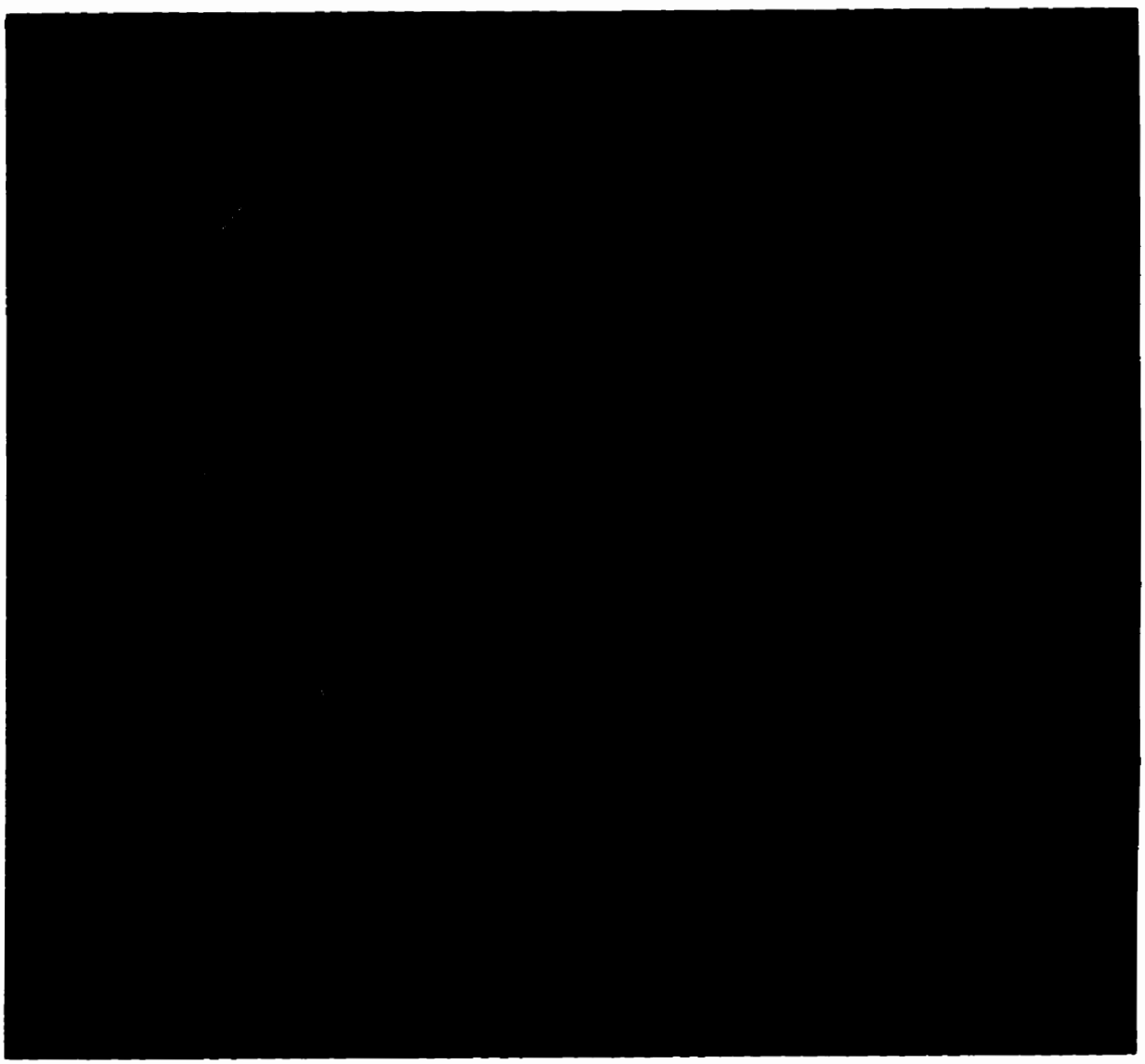


Figure 41. Superposition des 7 molécules proposées par l'équipe de Taga, résultat de la simulation Monte-Carlo.

On constate que les 7 molécules proposées sont toutes superposables avec la molécule type de la famille 3 telle que nous l'observons sur la figure suivante. Nous obtenons également des superpositions correctes avec les molécules types des familles 7, 5 et 13 dans un ordre décroissant c-à-d en allant de la meilleure vers la moins bonne superposition graphique. Nous

présentons dans la figure suivante la superposition d'une des molécule proposée par l'équipe de Taga avec la molécule type de la famille 3 de notre échantillon.

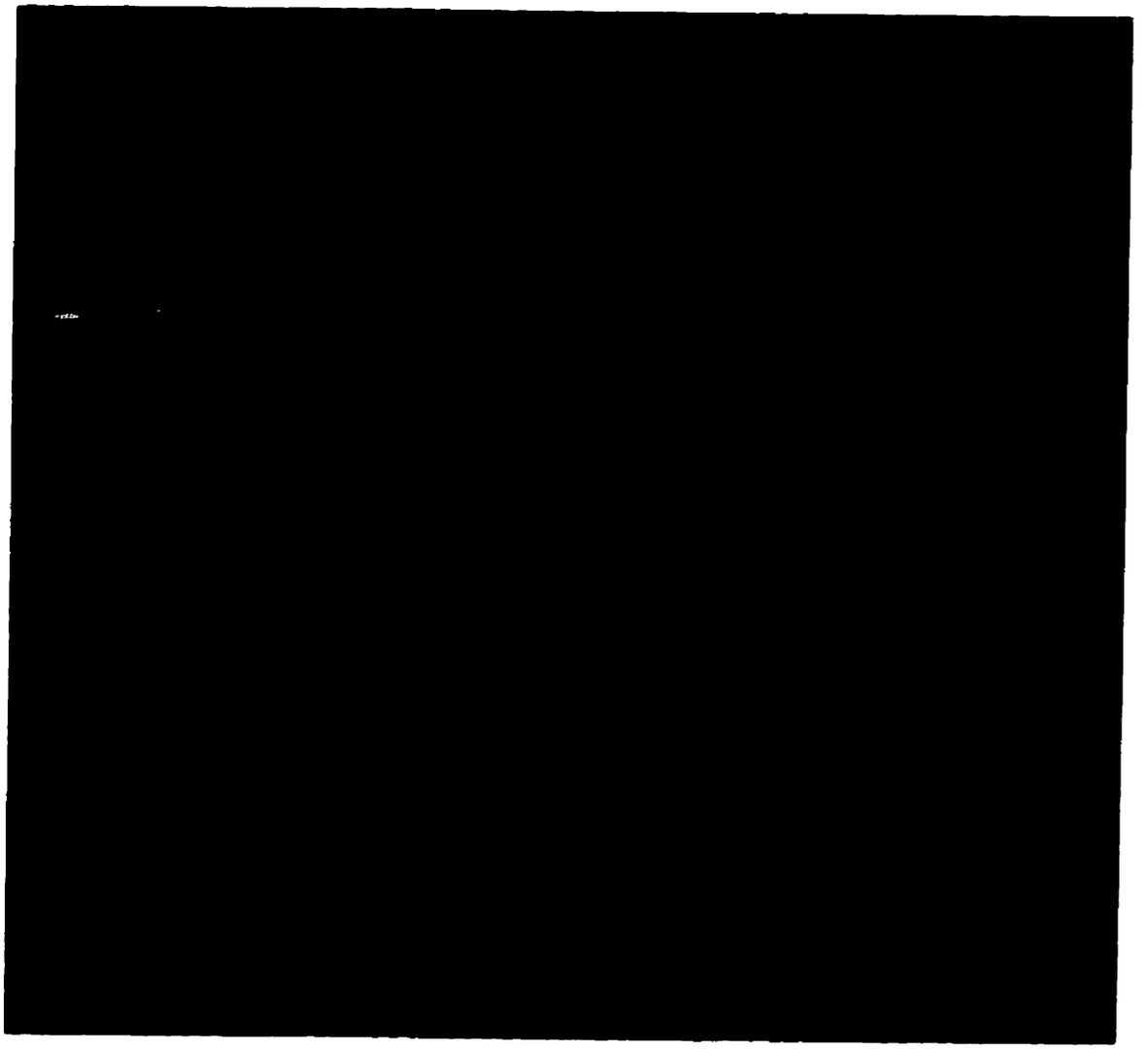


Figure 42. Superposition de la molécule type de la famille 3 de CCK-5 avec la molécule **a** proposée par l'équipe de Taga, résultat de la simulation Monte-Carlo.

Le même groupe de chercheurs publie une autre étude complémentaire sur la même molécule

(82) où l'ensemble des simulations Monte-Carlo effectuées est utilisé pour interpréter les résultats de l'étude RMN de la tétragastrine. Ils proposent ainsi une conformation de la molécule en solution qui n'est autre que la molécule **a**, proposée dans l'étude (81) qui est représentée à la figure suivante.

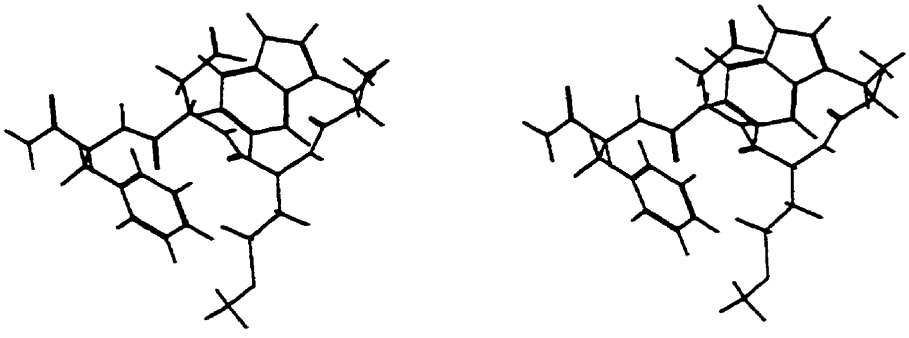


Figure 43. Vue stéréographique de la molécule de tétragastrine proposée par le groupe de Taga Confrontation des résultats obtenus par RMN avec la simulation Monte-Carlo.

Nous avons donc montré que la famille 3 dans notre échantillon correspond à la conformation de la tétragastrine en solution. Il faut remarquer la position des chaînes latérales du Trp et de la Phe dont les cycles sont orientés perpendiculairement l'un par rapport à l'autre. En ce qui concerne l'analogue d-Trp, l'ensemble des superpositions effectuées est beaucoup moins bon que pour CCK-5. Seule la famille #2 présente une superposition correcte avec les molécules proposées par le groupe de Taga tel que nous le présentons sur la figure 44.

Néanmoins, il faut remarquer que la chaîne latérale de la molécule ne peut être orientée comme sur la molécule a du fait de la stéréochimie différente du carbone asymétrique du résidu Trp.

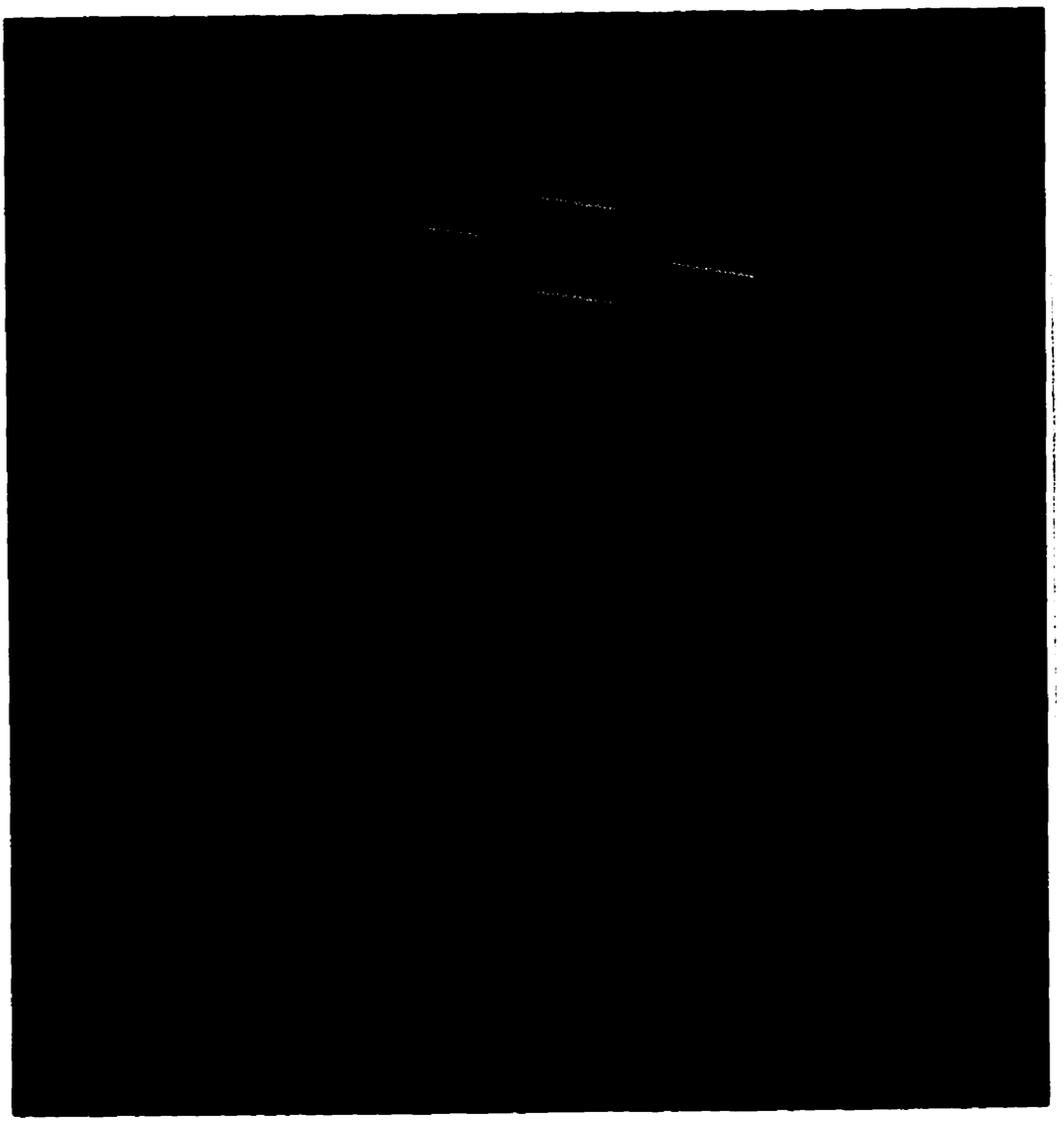


Figure 44. Superposition de la molécule type de la famille 2 de d-Trp CCK-5 avec la molécule **a** proposée par l'équipe de Taga, résultat de la simulation Monte-Carlo.

En résumé, nous avons montré que l'une de nos familles de CCK-5 présente une conformation très proche de celle du fragment CCK-4 en solution dans le DMSO ce qui n'est pas le cas pour l'analogue d-Trp qui ne peut adopter une conformation où les chaînes latérales des résidus Trp et Phe soient correctement positionnées. Nous remarquons ici que l'exploration

de l'hyperespace conformationnel à partir d'une simulation Monte-Carlo proposée par le groupe de Taga est inefficace dans la découverte de tous les minima métastables. En effet, tous les minima qu'ils proposent se distinguent exclusivement par la position de la chaîne latérale du Trp et leur échantillonnage consiste en fait à se promener sur l'hypersurface conformationnelle autour d'un puits d'énergie puisque chaque nouvelle molécule est une petite variation conformationnelle de la molécule précédente.

Ceci souligne le problème rencontré par les méthodes d'exploration d'un hyperespace conformationnel complexe par méthode Monte-Carlo ou de dynamique moléculaire. En effet, pour obtenir un véritable échantillonnage total par ces méthodes, il faudrait générer un nombre de conformations énorme requérant par conséquent un temps de calcul prohibitif. Ce problème est évité dans notre approche par la méthode d'échantillonnage choisie aléatoire et itérative des molécules.

L'autre étude que nous avons retenue pour la confrontation de nos résultats a été proposée par le groupe de Roques. Ils présentent une analyse conformationnelle de 4 fragments CCK dont CCK-5 (83). Ils utilisent la méthode de Monte-Carlo Métropolis pour l'exploration de l'espace conformationnel de ces molécules en étudiant les formes acides, neutres et basiques pour chaque fragment de manière à simuler les effets de pH. Tous ces fragments sont des ligands de CCK-B. Ils comparent les résultats de l'analyse conformationnelle avec les résultats expérimentaux obtenus par RMN et CD. Pour CCK-5, ils proposent ainsi 11 minima répartis en 6 conformations dans le cas de la molécule acide (qu'ils notent A, B, C, D, E, F), 3 conformations dans le cas de la molécule neutre (notés G, H et I) et 2 conformations dans le cas de la molécule basique (notées J et K).

Nous avons effectué une comparaison systématique par superposition graphique de chacune des 3 minima neutres qu'ils proposent (en conservant la dénomination des minima soit G, H et I) avec les 16 molécules types de chaque famille résultant de la classification de notre

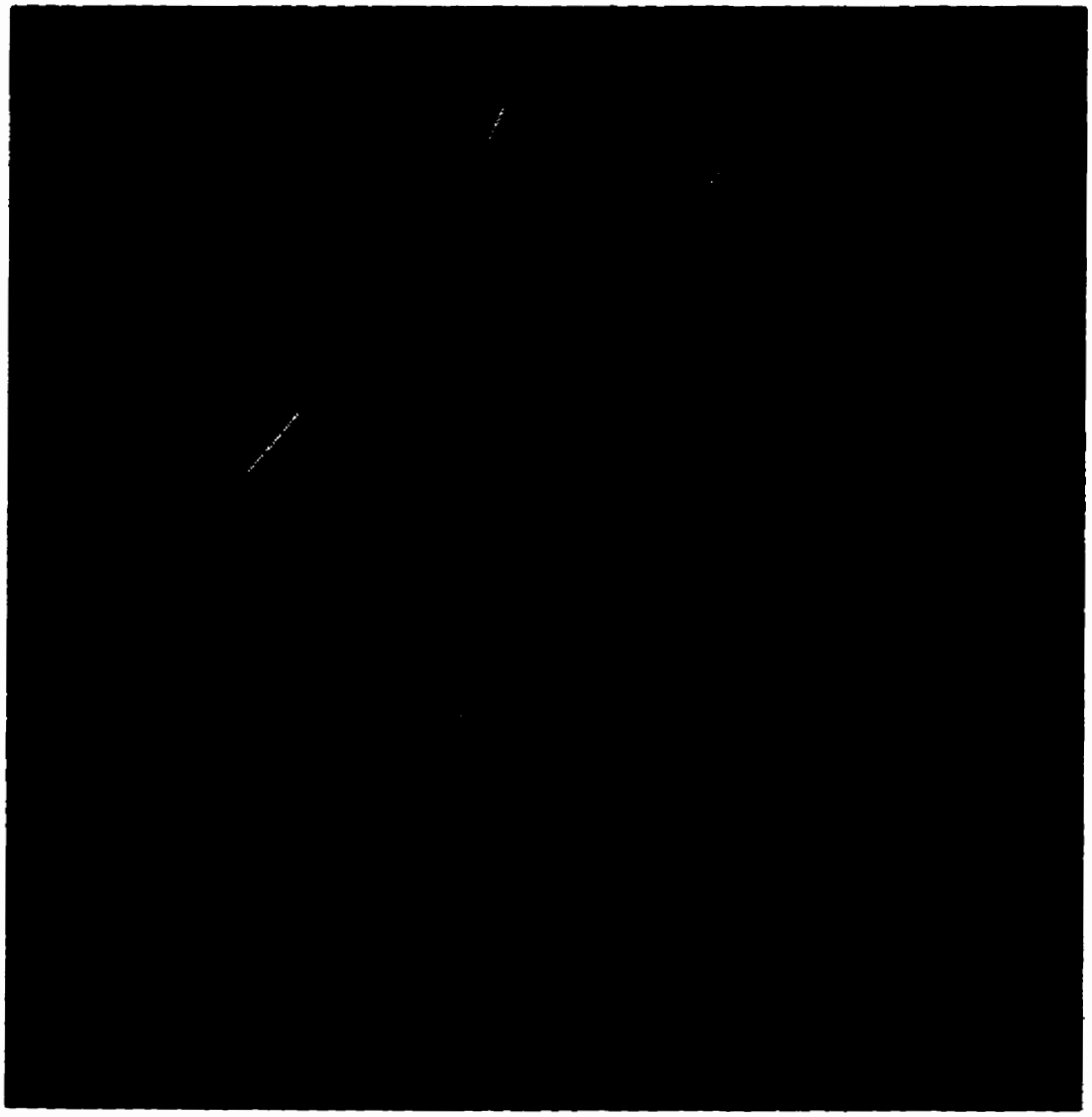


Figure 45. Superposition des 3 molécules G, H et I proposées par l'équipe de B.P. Roques, résultat de la simulation Monte-Carlo Métropolis.

échantillon puis des 13 molécules types de l'analogue d-Trp sachant que nous avons généré notre échantillon à partir de la forme neutre de la molécule.

Nous observons la même caractéristique que pour les 7 molécules proposées par le groupe

de Taga soit que les molécules proposées (les 3 minima neutres) sont très proches conformationnellement ce qui apparaît à la figure montrant la superposition des 3 conformations G, H et I.

Ces trois molécules se distinguent encore ici par la position du cycle de la chaîne latérale du résidu Trp.
Les superpositions graphiques avec CCK-5 donnent les résultats suivants:
- la molécule G est superposable avec les molécules types des familles #4, #7 et #13
- la molécule H est superposable avec les molécules types des familles #1, #4, #5, #6, #7 et #13
- la molécule I est superposable avec les molécules types des familles #3 et #13

Nous constatons que la famille la plus représentative pour les molécules G, H et I est la famille #13 et nous présentons les 3 superpositions graphiques de chacune de ces molécules avec la molécule type de notre famille #13 sur les figures suivantes.

Nous remarquons que dans l'étude précédente, la famille #13 avait également été superposée avec succès aux molécules proposées. En revanche, la position des chaînes latérales est assez différente ici. D'ailleurs, les auteurs donnent la distance entre les cycles des chaînes latérales pour Trp et Phe et suggèrent 10Å pour les molécules acides et 13Å pour les molécules neutres. Ces distances sont en contradiction avec la conformation observée par le groupe de Taga où les cycles sont très proches ainsi qu'avec leurs propres résultats pour la molécule H où les cycles des chaînes latérales des résidus Trp et Phe sont à moins de 6Å de distance.

En conclusion, la comparaison de nos molécules avec les différentes études publiées montre que nous avons reproduit les résultats de calcul aussi bien que les résultats expérimentaux. Nous avons montré de plus que les méthodes d'échantillonnage Monte-Carlo utilisées par les auteurs conduisent à un échantillonnage autour d'un minima et, par conséquent,

correspondantes à l'une ou plusieurs des familles conformationnelles déterminée par la classification de notre échantillon. Ainsi, nous pouvons proposer que chacune de nos familles reproduit une possibilité conformationnelle correspondant à des conditions d'environnement différentes : milieux hydrophile, lipophile,...

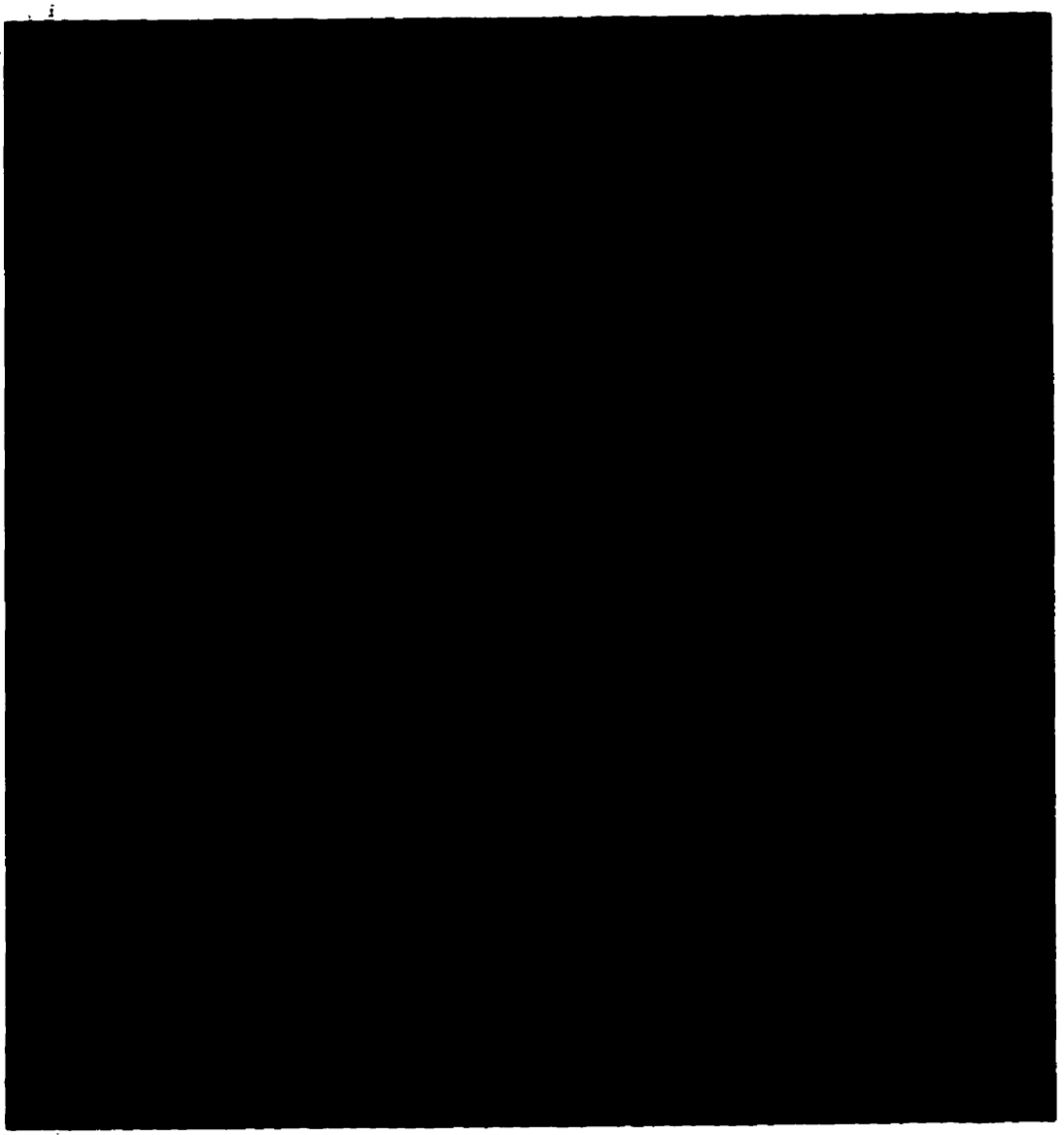


Figure 46. Superposition de la molécule type de la famille 13 de CCK-5 avec la molécule G proposé par l'équipe de Roques, résultat de la simulation Monte-Carlo Métropolis.

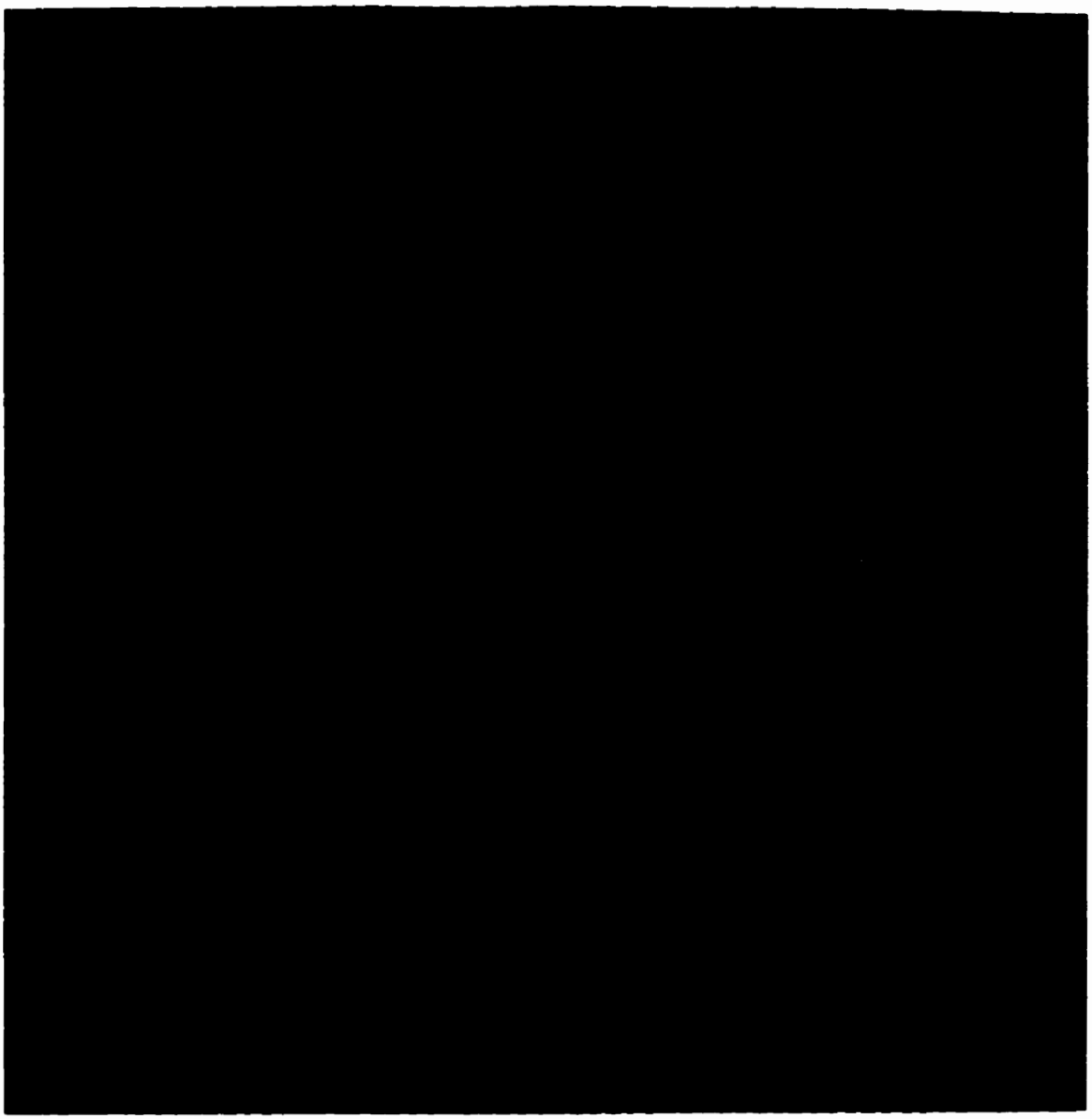


Figure 47. Superposition de la molécule type de la famille 13 de CCK-5 avec la molécule H proposée par l'équipe de Roques, résultat de la simulation Monte-Carlo Métropolis.

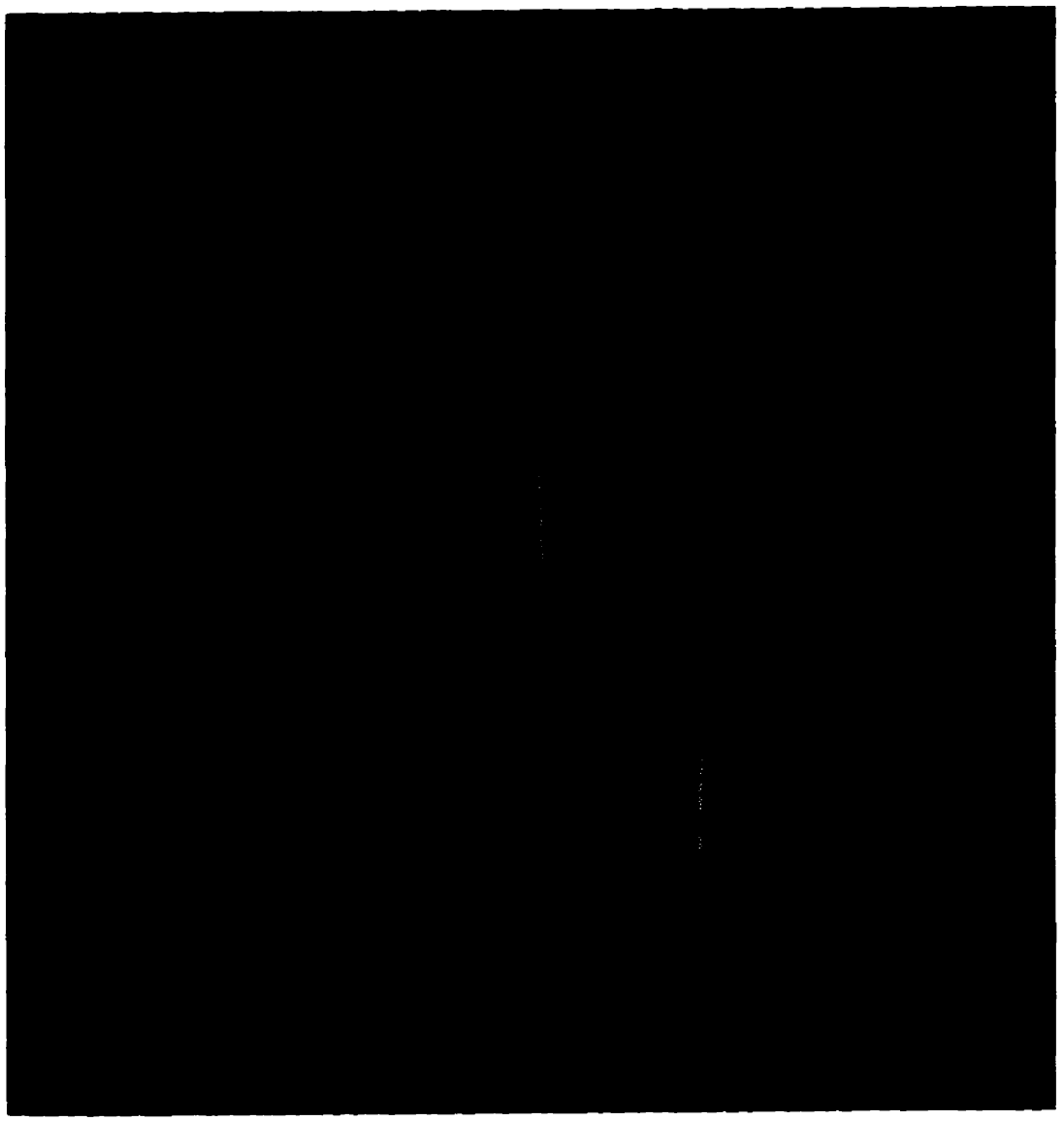


Figure 48. Superposition de la molécule type de la famille 13 de CCK-5 avec la molécule I proposé par l'équipe de Roques, résultat de la simulation Monte-Carlo Métropolis.

3.4.2 Comparaison avec les études des familles d'antagonistes de CCK

Les nombreuses études sur les familles d'antagonistes CCK a conduit a déterminer récemment

les pré-requis conformationnels nécessaires à la reconnaissance des molécules par les récepteurs CCK-A et CCK-B. Ces études réalisées par CoMFA (Comparative molecular field alignment technique) et QSAR (Quantitative structure-activity relationships) permettent de déterminer un ensemble de distances entre différents groupement de la molécule. Une molécule possédant les groupement adéquat correctement positionnés sera reconnaissable par le récepteur.

3.4.2.1 Récepteur CCK-B

Récemment, une nouvelle classe d'antagoniste a été proposée dans laquelle les auteurs ont développé une molécule d'activité comparable à celle de L-365,260 mais avec une sélectivité encore plus élevée (84). Malheureusement la stéréochimie absolue du produit n'a pas été déterminée ce qui ne nous permet pas de comparer leur molécule avec nos conformations. Une autre étude récente propose à nouveau une classe d'antagonistes dérivés des benzodiazépines où la molécule finale obtenue possède une activité supérieure à celle de L-365,260 (85). La modification apportée consiste simplement à remplacer le groupement CH_3 attaché sur l'azote du cycle benzodiazépine par un groupement CO-NH tBut ce qui ne change pas drastiquement a priori la conformation du squelette de la molécule. Les informations structurales ont été prises dans l'article (86), où le groupe de Vittoria a déduit des 5 classes connues d'antagonistes de CCK-B rapportées dans la littérature à ce moment là un modèle de pharmacophore. L'antagoniste le plus puissant est L-365,260 schématisé sur la figure 49.

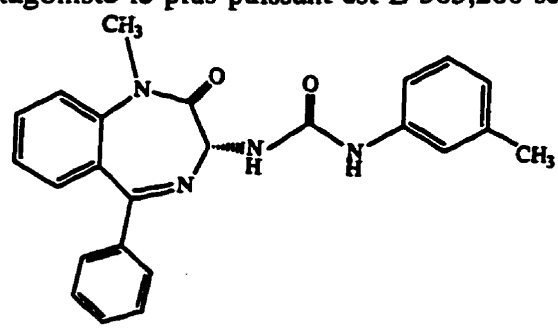


Figure 49. Schéma de L-365,260, antagoniste de CCK-B de la famille des benzodiazépines.

La géométrie du pharmacophore déterminé par l'équipe de Vittoria se présente ainsi:

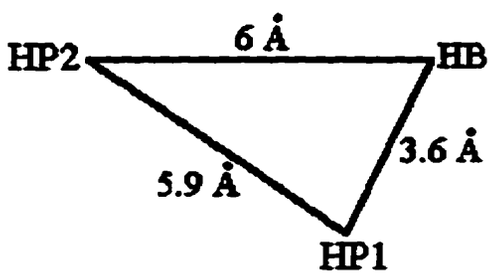


Figure 50. Géométrie du pharmacophore dérivée de l'étude des 5 familles d'antagonistes de CCK-B.

Les auteurs déterminent ainsi les sous-structures qui doivent être présentes pour qu'une molécule puisse être reconnue:

- l'oxygène d'un carbonyle ou d'une fonction amide ou urée capable de former un lien hydrogène avec un site donneur du récepteur (noté site HB)
- un système aromatique entouré dans la cavité du récepteur d'un environnement hydrophobe (site HP1)
- un fragment aromatique ou aliphatique dont la taille et la forme peuvent varier se liant à une autre région hydrophobe (HP2).

Ces 3 distances (HB-HP1, HB-HP2, et HP1-HP2) ont servi de contraintes pour trier la population de CCK-5. D'après les informations recueillies dans la littérature, le point HP1 du pharmacophore correspond au cycle de la chaîne latérale du Trp et le fragment HP2 au cycle

de la Phe. En revanche, en ce qui concerne la fonction carbonyle, nous avons en tout 6 carbonyles possible pouvant se placer au point HB du pharmacophore. La procédure utilisée a donc été la suivante:

- un premier tri des 15000 conformères qui vérifie la distance HP1-HP2 a été effectué ce qui a donné 366 molécules pour lesquelles 5.4Å > HP1-HP2 > 6.4Å
- ces 366 molécules ont été ensuite triées 6 fois en prenant pour HB chacun des 6 carbonyles possibles de la molécule. On rencontre 7 molécules de notre population pour lesquelles les contraintes de distance (3.1Å > HP1-HB > 4.1Å et 5.5Å > HP2-HB > 6.5Å) sont respectées. En outre, il y a un seul oxygène pour lequel nous ne trouvons aucune molécule susceptible de se placer dans la position requise. Leur énergie est comprise entre -20.39 kcal/mol et 5.47 kcal/mol.

Nous retrouvons donc à l'intérieur de notre population des molécules possédant les caractéristiques structurales nécessaires à leur reconnaissance par le récepteur CCK-B. De plus, l'énergie de ces molécules, possible ligands, n'est pas trop élevée par rapport au minimum global ce qui leur donne une probabilité d'existence non négligeable. La première molécule susceptible d'être reconnue par CCK-B est à 10.72 kcal/mol au dessus du minimum global.

De plus, nous avons recherché à quelle familles ces molécules appartiennent. Nous avons découvert une grande diversité dans les familles susceptibles de présenter les trois points pharmacophoriques correctement situés les uns par rapport aux deux autres. Ces 7 molécules appartiennent aux familles #9, #13, #3 (pour deux d'entre elles), #6, #5 et #10. De plus, il y a 5 des 6 oxygène qui peuvent se placer au point HB. Sachant qu'une famille regroupe des molécules dont les conformations sont proches les unes des autres, un apport d'énergie limité permet le passage facile de l'une à l'autre à l'intérieur d'une même famille et donc une molécule peut adapter sa conformation pour devenir active c-à-d se lier au récepteur.

Nous présentons sur la figure suivante la molécule la plus basse en énergie qui remplit les

exigences du récepteur. Elle fait partie de la famille #6.

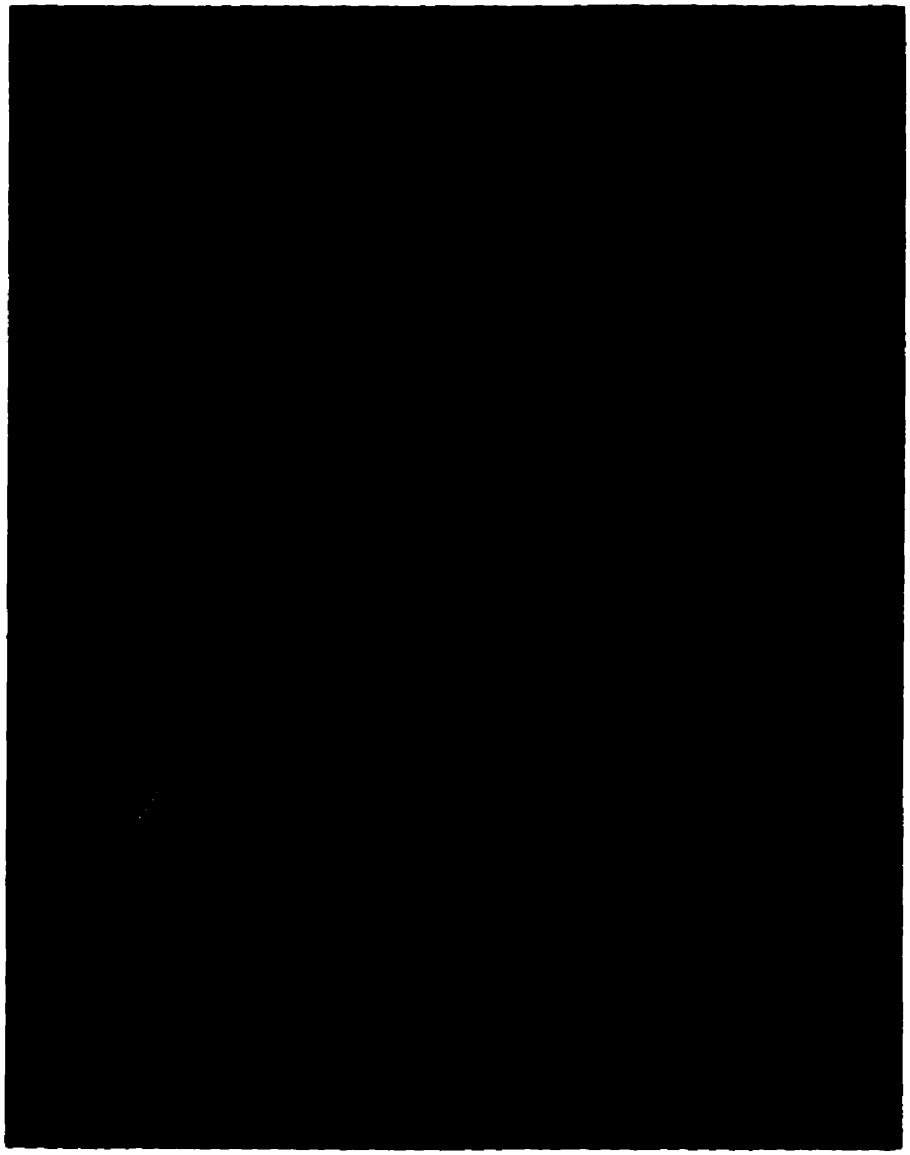


Figure 51. Molécule de la famille 6 de CCK-5 rencontrant les exigences conformationnelles du récepteur CCK-B telles que déterminés par le groupe de Vittoria

Le même travail a été effectué sur l'analogue d-Trp pour lequel nous avons passé la population au crible des contraintes de distance du pharmacophore. Nous obtenons les

résultats suivants:

- il existe 10 molécules qui vérifient les exigences conformationnelles du récepteur soit 5.4Å > HP1-HP2 > 6.4Å, 3.1Å > HP1-HB > 4.1Å et 5.5Å > HP2-HB > 6.5Å
- il y a deux oxygènes sur six pour lesquels il est impossible de trouver une conformation adéquate.
- Les familles auxquelles appartiennent ces molécules sont les #9, #13, #3 (pour deux d'entre elles), #6, #5 et #10 avec un intervalle d'énergie compris entre -18.99 et -2.67 kcal/mol

Nous constatons ici que la première molécule susceptible d'être reconnue par le récepteur possède une énergie 14.59kcal/mol plus élevée que le minimum global. En revanche, un plus grand nombre de molécules vérifient les exigences conformationnelles du récepteur et 4 sur 6 oxygènes sont susceptibles de se placer au point pharmacophorique HB.

En conclusion, nous ne pouvons expliquer ainsi définitivement pourquoi l'analogue d-Trp est inactif alors que le fragment CCK-5 est agoniste. Néanmoins, il faut souligner que la géométrie du pharmacophore a été obtenue à partir de familles d'antagonistes et il est probable que l'action agoniste du fragment requière des points pharmacophoriques supplémentaires. Nous pouvons d'ailleurs remarquer que la taille des molécules agonistes est plus élevée que celle des molécules antagonistes. De plus, ces dernières sont rigides or il n'a encore été développé à notre connaissance aucune molécule agoniste très rigide, les agonistes publiés étant des dérivés de peptides. Ceci laisse penser que la molécule agoniste peptidique subit des modifications conformationnelles à l'approche du récepteur, ce qui la rend pleinement active.

3.4.2.2 Récepteur CCK-A

Le fragment CCK-5 ne possède théoriquement pas d'activité sur le récepteur CCK-A. Nous désirons néanmoins vérifier comment ce fragment vérifie ou non les exigences

conformationnelles du récepteur. En effet, il est possible que l'absence d'activité in vivo du fragment sur le récepteur périphérique soit dû à d'autres causes que la conformation adoptée par ce fragment comme par exemple la dégradation du fragment avant même qu'il n'ait atteint le site du récepteur.

De très nombreuses familles d'antagonistes du récepteur CCK-A sont connues depuis longtemps et la littérature à ce sujet est abondante à l'inverse de CCK-B. Ceci est dû en partie aux processus physiologique dont ce récepteur fait partie en particulier les phénomènes de satiété et sécrétions pancréatiques. Les molécules actives sur ce récepteur sont par conséquent de possibles thérapeutiques pour les ulcères et troubles de l'alimentation. L'antagoniste le plus actif est encore à notre connaissance un dérivé benzodiazépine noté MK-329 dont le schéma apparaît à la figure suivante:

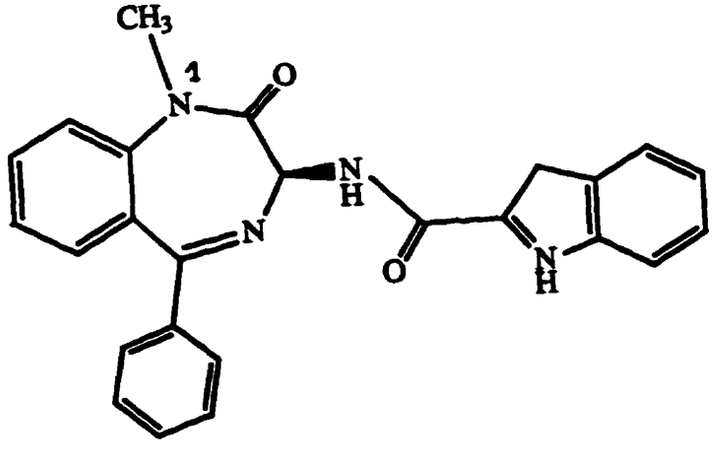


Figure 52. Schéma de MK-329, antagoniste de CCK-A de la famille des benzodiazépines.

L'équipe de Robba a utilisé plusieurs antagonistes connus de CCK-A pour en déduire les points pharmacophoriques du récepteur (87 , 88) .

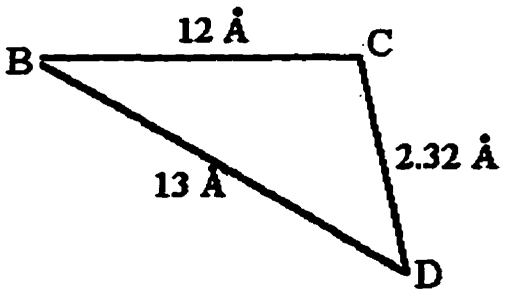


Figure 53. Géométrie du pharmacophore dérivée de l'étude de familles d'antagonistes de CCK-A.

La région notée B doit être occupée par un halogène, ou un cycle aromatique. Les région C et D doivent être occupées par des cycles aromatiques.

Le processus utilisé est ici un peu différent. En effet, les travaux publiés ne faisait aucunement état de l'endroit où avait été mesuré les distances du pharmacophore (à l'extérieur de la molécule ou directement sur le centre des cycles aromatiques?). Nous avons par conséquent utilisé le biais de mesurer les distances correspondantes sur la structure obtenue par diffraction de rayons-X du fragment CCK-4 (72, 75). Ce dernier, bien qu'inactif sur le récepteur CCK-A vérifie les exigences conformationnelles de CCK-A tel qu'ils les déterminent à partir des antagonistes de CCK-A. Ils présentent un schéma sur lequel on voit que le cycle aromatique qui se trouve théoriquement au point C est remplacé alors par la fonction carbonyle du résidu Asp. Les distances mesurées sur CCK-4 donnent la géométrie suivante pour le pharmacophore:

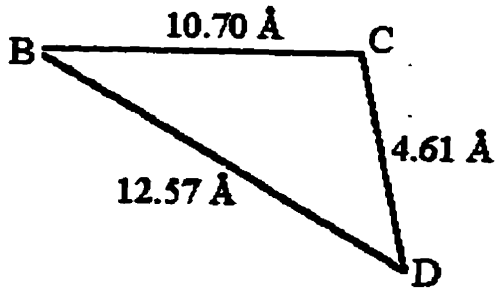


Figure 54. Géométrie du pharmacophore de CCK-A mesurée sur la structure RX de CCK-4.

Ces distances ont servies à trier l'échantillon de conformations de CCK-5 de manière à obtenir les conformations vérifiant les distances entre les points pharmacophoriques. On obtient que 662 des conformations de CCK-5 possèdent la distance correcte entre B et D soit les cycles des chaînes latérales des résidus Phe et Trp. Dans un deuxième temps, ainsi que nous avions procédé pour CCK-B, nous avons pris pour le point C successivement les 6 carbonyles possibles existants sur la molécule de CCK-5. Nous obtenons finalement que 17 molécules sont susceptibles d'être reconnues par le récepteur CCK-A. L'échelle d'énergie est comprise entre -16.58 et 12.31 kcal/mol et les conformères appartiennent aux familles #6, #7, #8, #9, #10 et #12, #13, #14, #15 et #16.

Nous constatons ici qu'un plus grand nombre de molécules de CCK-5 vérifient les exigences du récepteur CCK-A par rapport à CCK-B mais que la première molécule possède une énergie conformationnelle 14.53 kcal/mol plus élevée que le minimum global ce qui est plus élevé que pour CCK-B. De plus, il y a seulement 3 des oxygènes sur 6 possibles dans la molécule qui peuvent se placer au point pharmacophorique C.

En ce qui concerne la population de l'analogue d-Trp, après avoir effectué le même tri à partir des contraintes du récepteur CCK-A, nous obtenons 17 molécules qui vérifient les contraintes de distances déduites des travaux de l'équipe de Robba et dont l'énergie est comprise entre -19.66 et 9.50 kcal/mol. Là encore, seuls 3 oxygènes peuvent se placer au point pharmacophorique C.

En conclusion des ces comparaisons, nous constatons que nous ne pouvons définir exactement les raisons pour lesquelles la molécule CCK-5 est active sur le récepteur CCK-B et inactive sur CCK-A alors que l'analogue d-Trp de CCK-5 est inactif puisque dans chaque population et pour chaque récepteur, nous observons des molécules qui vérifient les exigences conformationnelles du récepteur en question. Nous pouvons néanmoins émettre un certain nombre de remarques:

- nous avons vu précédemment que CCK-5 est plus flexible que son analogue d-Trp. Ceci pourrait expliquer l'inactivité de l'analogue si on se place dans la théorie où la molécule subit des modifications conformationnelles à l'approche du récepteur.

- dans l'interaction avec CCK-A, le nombre d'oxygènes qui peuvent se placer au point pharmacophorique C est restreint (3 sur 6 possibilités dans la molécule). Or on ne sait pas a priori quel oxygène est supposé occuper cette position. Il est donc possible que l'oxygène supposé occuper ce point pharmacophorique soit justement l'un de ceux pour lesquels il est impossible de trouver une molécule vérifiant les 3 contraintes de distances du pharmacophore. C'est également ce qui peut arriver dans le cas de la liaison avec le récepteur CCK-B où, pour l'analogue d-Trp et pour 2 des 6 oxygènes, nous ne pouvons trouver de molécules vérifiant les contraintes conformationnelles du récepteur alors que pour les même oxygènes, il existe des molécules pour CCK-5 qui peuvent adopter la conformation reconnue par le récepteur.

Il ne faut pas oublier que les conformations sont triées ici à partir des contraintes des deux

types de récepteurs déterminées grâce aux molécules antagonistes de ces récepteurs. Comme nous l'avons déjà souligné, l'action agoniste d'une molécule requiert éventuellement des points pharmacophoriques additionnels qui peuvent faire la différence entre molécules actives et inactives.

L'étude publiée par l'équipe de Robba (87) souligne cette possibilité en montrant qu'effectivement, le fragment CCK-4 vérifie les exigences conformationnelles du récepteur CCK-A bien qu'elle soit inactive sur ce récepteur. Il avance l'hypothèse que la chaîne latérale de l'Asp occupe un site qui est libre sur les antagonistes de CCK-B sur lesquels ils ont construit leur modèle de pharmacophore. Ce site est l'azote numéroté 1 sur le cycle benzodiazépine de la figure 52.. Nous pouvons faire deux observations à ce sujet:

- les antagonistes actif sur CCK-A ne possèdent effectivement pas de fragment encombrant sur cette position.
- l'antagoniste actif sur CCK-B noté L-365,260 ne possède pas non plus de fragment encombrant sur cette position

l'activité des molécules vis à vis de l'un ou l'autre des récepteurs CCK-A ou CCK-B ne peut donc être due exclusivement à la présence d'un groupement encombrant à cet endroit.

Néanmoins, la présence d'un groupement encombrant à cet endroit n'empêche par l'activité des molécules sur CCK-B puisque le fragment CCK-4 est actif sur le récepteur CCK-B. Un groupement encombrant en cette position semble même être un facteur qui améliore la sélectivité des molécules sur CCK-B. En effet, une transformation apportée à L-365,260 (85) qui améliore la sélectivité à CCK-B est de remplacer le groupement CH_3 lié à l'azote 1 par un groupement encombrant CO-NH Tbut tel que le montre la figure suivante:

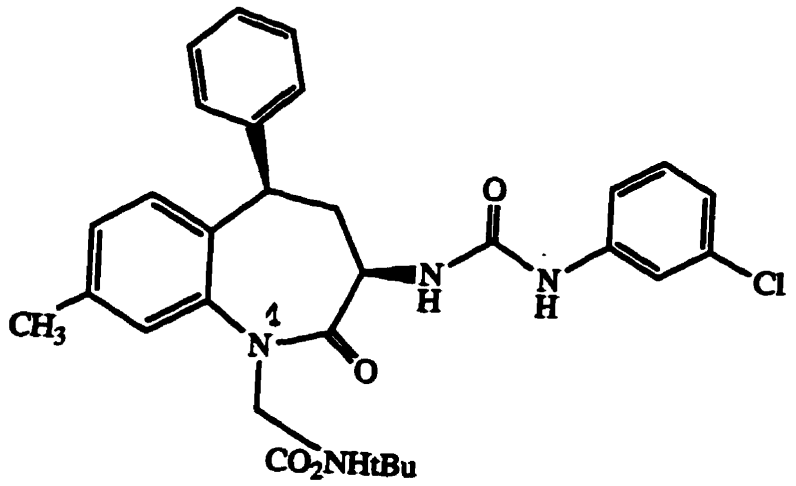


Figure 55. Schéma de CP-212,454, antagoniste de CCK-B de la famille des benzodiazépines développé sur le squelette de L-365,260.

Il faut enfin souligner que pour déterminer les distances pharmacophoriques du récepteur CCK-A nous sommes passés par le biais de mesurer ces distances sur une molécule qui vérifie les exigences conformationnelle du récepteur puisque nous ne savions pas entre quels points les distances pharmacophoriques fournies avaient été mesurées par les auteurs ce qui diminue la précision des distances obtenues.

CONCLUSION

Nous avons montré comment les méthodes d'analyses de données peuvent s'appliquer à l'étude d'échantillon de conformations générées par la méthode PEPSEA. Ces méthodes appliquées sur un échantillon adéquat permettent de rationaliser l'étude de molécules très flexibles possèdant une grande variété de conformations métastables. Grâce à ces méthodes d'analyse de données, il devient possible d'étudier l'ensemble de l'hypersurface conformationnelle d'une molécule. Nous avons relevé un certain nombre de difficultés relatives à l'emploi de ces méthodes. Un des problèmes rencontré est commun à toutes les méthodes de simulation: si nous voulons appliquer les méthodes d'analyses des données il nous faut échantillonner correctement toutes les conformations possibles. Cela implique deux exigences: le champ de force choisi doit être capable de calculer une énergie potentielle correcte, et l'échantillon produit doit être représentatif. Le premier point a été justifié par les travaux antérieurs du laboratoire et le dernier point est vérifié par le calcul de la moyenne de l'énergie pour des tailles croissante de l'échantillon jusqu'à obtention d'une valeur constante.

L'analyse en composantes principales appliquée à nos échantillon de molécules ne pose aucun problème sinon celui de la représentation des individus dans l'espace des premières composantes principales lorsque l'échantillon est de grande taille. Or notre but principal n'est pas l'exploration graphique de nos données, mais la construction de variables indépendantes. Le premier problème posé par la classification est celui de l'interprétation des indices statistiques. Nous avons résolu ce problème en utilisant conjointement plusieurs méthodes de données jusqu'à avoir un consensus pour les résultats des classifications. Lorsque l'échantillon est de grande taille néanmoins, le temps de calcul nécessaire à la classification devient un paramètre non négligeable et limite l'exploration à quelques méthodes de classifications. De plus la classification pose alors un problème intrinsèque car, quelque soit la méthode choisie, il devient difficile théoriquement de trouver les hyperplans délimitant les familles dans

l'hyperespace des variables composantes principales. En effet, une molécule complexe possédant de nombreux degrés de liberté est caractérisée par de nombreuses variables et l'hyperespace de sa surface d'énergie potentielle est par suite extrêmement complexe. De plus, l'évaluation des résultats au point de vue chimique devient difficile et il nous a fallut développer des méthodes pour juger la qualité des familles obtenues (pourcentage de code par résidu par famille, valeurs moyennes des distances caractéristiques...). Ceci constitue une des limitations de la méthode au point de vue pratique (temps de calcul) et théorique. Néanmoins, il faut remarquer que la complexité apportée par l'étude d'un échantillon du dernier peptide ne peut guère être dépassée. Le nombre de conformations accessibles à un peptide donné augmente très rapidement dans un premier temps avec la taille de ce peptide. De plus la complexité des résidus d'acides aminés (chaînes latérales importantes, nombreuses fonctionnalités) augmente les possibilités de conformations. On peut ainsi penser que l'efficacité de la méthode proposée pour ordonner les échantillons de conformations peptidiques est très limité puisque le nombre de 5 résidus d'acides aminés en constitue la limite pratique. Néanmoins,plus la taille du peptide augmente, plus la complexité de la surface conformationnelle, diminue. En effet, des éléments de structure tertiaire apparaissent qui vont progressivement figer les conformations et les possibilités conformationnelles vont diminuer au fur et à mesure que le nombre de résidus d'acides aminé augmente. Par conséquent, la taille de l'échantillon représentatif diminue. Les gros peptides et protéines possèdent une conformation bien définie et très peu de liberté conformationnelle limitée à de légères variations autour du site actif. La limite d'apparition de la structure tertiaire est de 5 acides aminés puisqu'elle permet la formation d'un pseudo-cycle à 13 membres traduisant une structure en hélice α, élément important de rigidification des peptides et protéines. La présence de certains résidus particuliers comme la proline, l'hydroxyproline ou la cystéine conduit également à une réduction de la liberté conformationnelle des composés. Les deux premiers induisent des structures tertiaires particulières qui sont les tournants β et le dernier forme des liaisons chimiques qui rigidifient les squelettes protéiques. Avec le dernier peptide étudié, nous avons donc montré que nous pouvions traiter une structure de la plus grande

flexibilité possible de par sa taille et sa composition en acides aminés divers.

Nous pensons ainsi avoir montré les avantages de ces techniques d'analyses: classification sans a priori, appréciation des corrélations entre les variables, rapidité de l'analyse et traitement de gros échantillons. Nous pouvons de plus utiliser les méthodes d'analyse de données dans le cadre de l'étude de molécules biologiquement actives pour lesquelles certaines caractéristiques reliées à leur activité sont connues: conformation de la partie active de la molécule ou distance caractéristique entre deux groupements nécessaire à l'activité. Si nous faisons l'analyse en composantes principales à partir de distances spécifiques correspondant à des caractéristiques connues, nous pourrons évaluer par exemple si ces distances décrivent pertinemment la population et évaluer le pourcentage de la population décrite par ces distances après la classification. Cela nous permettant d'estimer quel est la chance pour la molécule de se lier au récepteur dont nous connaissons les caractéristiques structurales. Nous pouvons ainsi évaluer par exemple la qualité d'agoniste ou d'antagoniste de plusieurs molécules. Nous pouvons de cette façon utiliser des résultats expérimentaux fournis par la RMN ou la fluorescence de rayons-X.

Nous pensons que les méthodes de classification constituent un outil d'exploration idéal pour l'étude des hypersurfaces conformationnelles de molécules complexes. Il existe de très nombreuses méthodes reliées de près ou de loin à l'analyse de données et nous n'avons exploré qu'une partie de ses méthodes. En effet, selon le but poursuivi, d'autre méthodes d'analyses des données sont susceptibles d'être utilisées comme c'est déjà le cas dans l'étude des relations structures-activité des composés biologiquement actifs où les différentes techniques d'analyse factorielle et de regroupement sont couramment utilisées. Il serait donc intéressant de développer ce genre d'approche à l'avenir dans l'étude de populations conformationnelles d'autant plus que les résultats obtenus permettent l'accès ultérieur à des paramètres quantitatifs comme le poids statistique et ce de manière à la fois simple et générale.

ANNEXE A

L'ANALYSE EN COMPOSANTES PRINCIPALES

Soit un tableau I comportant p variables quantitatives. Les n individus, caractérisés chacun par p variables, peuvent être représentés dans un espace vectoriel à p dimensions. Le but de l'ACP est de trouver un espace vectoriel de dimension restreinte (donc un sous-espace vectoriel de l'espace initial) dans lequel il soit possible d'examiner au mieux les individus. S'il existe un vecteur colonne u $(u_1, u_2,..., u_n)'$ et un vecteur ligne v $(v_1, v_2,..., v_p)$ tel qu'il soit possible de reconstituer le tableau I avec I = u.v', on aura reconstitué les np valeurs du tableau I par seulement (n+p) valeurs. En pratique, on cherchera l'approximation de rang q telle que $I = u_1v_1' + u_2 v_2' + ... + u_q v_q' + E_p$, (E_p étant une matrice résiduelle dont les termes sont petits). On pourra ainsi reconstituer le tableau I grâce à n(q+p) valeurs :

·	p_1	p_2	...	p_p
x_1	...	...	...	...
x_2	...	...	...	...
...	...	...	...	...
x_n	...	...	...	...

→

·	cp_1	cp_2	...	cp_p
x_1	...	...	...	...
x_2	...	...	...	...
...	...	...	...	...
x_n	...	...	...	...

[8]

On effectue un changement de base. L'espace des variables initiales est remplacé par l'espace des composantes principales. Les p variables quantitatives initiales qui sont plus ou moins corrélées entre elles sont remplacées par p nouvelles variables non corrélées.

A.1 Traitement des données initiales

La base de représentation la plus adéquate pour nos individus sera constituée par celle dont la projection des individus sur les axes aura la plus grande variance. Une transformation intéressante des variables initiales est de les réduire et de les centrer. En effet, h_i et h_j désignant les projections de 2 points i et j sur une droite H, on veut que le nuage de point soit le plus étalé possible soit rendre maximale la somme $\Sigma(h_i - h_j)^2$. Cette somme peut se développer en $2n\ \Sigma(h_i - \bar{h})^2$ avec $\bar{h}$ moyenne des projections. Si on centre les variables, i.e. que l'on prend l'origine au point moyen, $\bar{h}=0$. De plus, si les variables ne sont pas mesurées avec des échelles identiques, il faudra normer ces variables de manière à ce qu'elles aient toutes le même poids dans l'analyse en composantes principales:

$$x_{ij} = \frac{(r_{ij} - \bar{r}_j)}{\sigma_j} \qquad [9]$$

Si r_{ij} est un élément de rang i du tableau initial, on lui soustrait la moyenne sur la colonne j et on le divise par l'écart-type de cette colonne de manière a centrer et réduire chaque valeur de propriétés mesurée. Ainsi chaque propriété aura le même poids dans l'analyse subséquente. Le but de l'analyse consiste à déterminer une base du sous-espace vectoriel qui ajuste au mieux, dans l'optique de la méthode des moindres carrés, l'ensemble des n points. A cette fin, on montre que la solution de ce problème se ramène à la diagonalisation de la matrice de variance-covariance associée aux données expérimentales. Après avoir centré et réduit les variables initiales, cette matrice sera la matrice des corrélations entre les variables prises deux à deux, notée I'I, dont les éléments sont calculés selon la formule:

$$\rho_i = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n \sum x_i^2 - (\sum x_i)^2]\ [n \sum y_i^2 - (\sum y_i)^2]}} \qquad [10]$$

On mesure ainsi la corrélation entre chacune des p variables et toutes les autres. On obtient ainsi une matrice symétrique, la diagonale étant composée du coefficient de corrélation entre

chaque variable et elle-même et vaut donc 1 pour chaque élément.

A.2 Rappels sur les propriétés des matrices

Certaines propriétés des matrices permettent de voir comment le fait de diagonaliser la matrice de corrélations permet de représenter le tableau I au mieux.

. Produit scalaire

soit 2 vecteurs colonnes à n composantes x et y, le produit scalaire de ces deux vecteurs est par définition:

$$x'y = y'x = \sum_{i=1}^{n} x_i\, y_i \qquad [11]$$

si x'y = 0, alors x est orthogonal à y ce qui signifie géométriquement que cos (x, y) = 0 dans un espace à n dimensions.

. si X est une matrice quelconque, XX' et X'X sont des matrices symétriques semi-définies positives.

. Soit p vecteurs colonnes $(x_i)_{i=1, 2, ..., p}$ de dimension n, et p coefficients réels (c_1, c_2, ..., c_p), les vecteurs seront linéairement indépendants si et seulement si:

$$c_1x_1 + c_2x_2 + \ldots + c_px_p = 0 \quad \Leftrightarrow \quad c_1 = c_2 = \ldots = c_p = 0 \qquad [12]$$

. Le rang d'une matrice est égal au plus grand nombre de ses vecteurs colonnes linéairement indépendants. Une matrice est de plein rang si son rang est égal à sa plus petite dimension.

. Soit X une matrice carrée (p, p), elle est régulière si elle est de plein rang (i.e. rang X=p). De plus, dans ce cas, son déterminant est non nul.

. Pour toute matrice symétrique X, il existe une matrice orthogonale Y telle que Y'XY = Λ où Λ est une matrice diagonale.

. Soit une matrice carrée (p, p), λ un scalaire et x un vecteur (p, 1) non nul, on a Xx = λx

s o i t :

$$(X - \lambda I)\, x = 0 \qquad [13]$$

les colonnes de $(X - \lambda I)$ sont linéairement dépendantes donc $\det(X - \lambda I) = 0$ (équation caractéristique de X). Le scalaire λ est une valeur propre et x un vecteur propre de X associé à λ.

- L'équation caractéristique est un polynôme de degré p possédant p valeurs propres non forcément distinctes.
- La somme des valeurs propres est égale à la trace de la matrice; leur produit est égal au déterminant de la matrice.
- Si la matrice est symétrique, toutes les valeurs propres sont réelles; le nombre de ses valeurs propres non nulles est égal au rang de la matrice; les vecteurs propres, correspondants à des valeurs propres distinctes, pris deux à deux sont orthogonaux.

A.3 Diagonalisation de la matrice des corrélations

On cherche les vecteurs propres u_α de la matrice Γ. Cette dernière est une matrice symétrique (p,p) dont les p valeurs propres réelles (si elles sont toutes distinctes) permettront d'engendrer p vecteurs propres orthogonaux et linéairement indépendants par définition, dont l'ensemble constituera une base de $\Re^p$. Chacun des vecteurs propres sera donc un sous-espace vectoriel de dimension 1. Le premier vecteur propre sera la droite sur laquelle les individus projetés auront la plus grande variance. De cette manière, les p vecteurs propres seront classés par ordre de représentativité. Les q vecteurs propres suffisant à la reconstitution du tableau de données X seront les p premiers vecteurs propres puisque les p-q derniers vecteurs propres assumant une part négligeable dans la représentation des n valeurs de X peuvent être négligés.

A.4 Exemple simple d'ACP

Soit un tableau constitué de 31 individus et 3 variables:

n p	N1	N2	N3	n p	N1	N2	N3
1 Luc	**12.5**	**17.5**	**17.5**	17 Ludovic	**18.0**	**17.5**	**15.0**
2 Isabelle	**16.5**	**14.5**	**14.5**	18 Josée	**16.0**	**14.5**	**16.5**
3 Philippe	**14.0**	**16.0**	**15.5**	19 Franco	**16.0**	**13.5**	**13.5**
4 Jacques	**12.0**	**13.0**	**14.5**	20 Emmanuel	**17.0**	**17.5**	**15.0**
5 Véronique	**18.0**	**14.0**	**11.5**	21 Sophie	**18.0**	**18.5**	**17.0**
6 Pierre	**19.0**	**17.0**	**16.5**	22 Michel	**13.0**	**18.0**	**16.0**
7 Corinne	**16.0**	**17.5**	**16.0**	23 Eliane	**16.0**	**15.0**	**15.0**
8 Alain	**13.0**	**17.5**	**17.0**	24 Karine	**12.5**	**15.5**	**10.0**
9 Florence	**16.0**	**15.5**	**12.5**	25 Nelly	**11.5**	**13.5**	**13.0**
10 Pascale	**12.0**	**15.5**	**14.5**	26 Franck	**15.0**	**16.0**	**16.5**
11 Frédéric	**15.5**	**15.0**	**17.0**	27 Raphaëlle	**16.0**	**17.0**	**17.5**
12 Françoise	**18.0**	**17.5**	**14.5**	28 Christiane	**13.5**	**17.0**	**14.0**
13 Lucie	**17.0**	**16.5**	**14.5**	29 Jean-Paul	**13.5**	**12.5**	**15.5**
14 Nathalie	**14.0**	**15.0**	**14.0**	30 Marielle	**12.5**	**16.5**	**13.5**
15 Martin	**13.0**	**16.0**	**15.0**	31 Anne-Marie	**15.0**	**15.5**	**16.0**
16 Rachel	**18.0**	**15.5**	**14.0**				
Moyenne					**15.1**	**15.8**	**14.9**
Ecart-type					**2.2**	**1.6**	**1.7**

La première étape consiste en la transformation des p=3 variables initiales N1, N2 et N3 en variables centrées réduites, puis en calcul de la matrice des corrélations qui se présente alors comme suit:

	N1	N2	N3
N1	**1.0000**	**0.2424**	**0.1107**
N2	**0.2424**	**1.0000**	**0.4075**
N3	**0.1107**	**0.4075**	**1.0000**

La diagonalisation conduit à 3 valeurs propres distinctes (dont la somme est égale à la trace de la matrice des corrélations). En effet, chaque variable initiale étant centrée et réduite possède, par définition, une variance de 1 donc la somme des variances est égale au nombre de variables initiales. Ainsi, le rapport de chaque valeur propre à la somme de toutes les valeurs propres donne le pourcentage d'information retenu par chaque axe cp engendré par la valeur propre en question:

Valeurs Propres		Pourcentage de Variance	Pourcentages cumulés
1	**1.52774**	**0.509248**	**0.50925**
2	**0.90401**	**0.301388**	**0.81059**
3	**0.56824**	**0.189414**	**1.00000**

Les vecteurs propres associés à ces trois valeurs propres constituent une base de l'espace vectoriel de départ. On peut écrire ainsi une matrice appelée de vecteurs propres contenant les coefficients de combinaison linéaire permettant d'écrire les nouveaux vecteurs en fonction des anciens. On peut ainsi voir quel part prend chacune des variables initiales à la construction des nouvelles variables:

	cp1	cp2	cp3
N1	**0.433521**	**0.865337**	**0.251499**
N2	**0.667010**	**-0.120473**	**-0.735244**
N3	**0.605935**	**-0.486496**	**0.629416**

D'après les pourcentages de variance retenus par les valeurs propres, on peut voir que si on décide de ne garder qu'un axe de représentation, on choisira cp1 puisque celui-ci retient environ 50% de l'information totale contenue dans le tableau de données. Les 31*3=93 valeurs initiales du tableau X étant représentées par 31+3=34 valeurs seulement soit les composantes des 31 individus de départ calculées sur le nouvel axe cp1 plus les composantes de cet axe en fonction des axes de départ N1, N2, N3. Une meilleure représentation sera constituée par le sous-espace vectoriel de dimension 2 engendré par cp1 et cp2. Les 93 valeurs initiales étant alors reconstituées à 81% par 2(31+3)=68 valeurs.

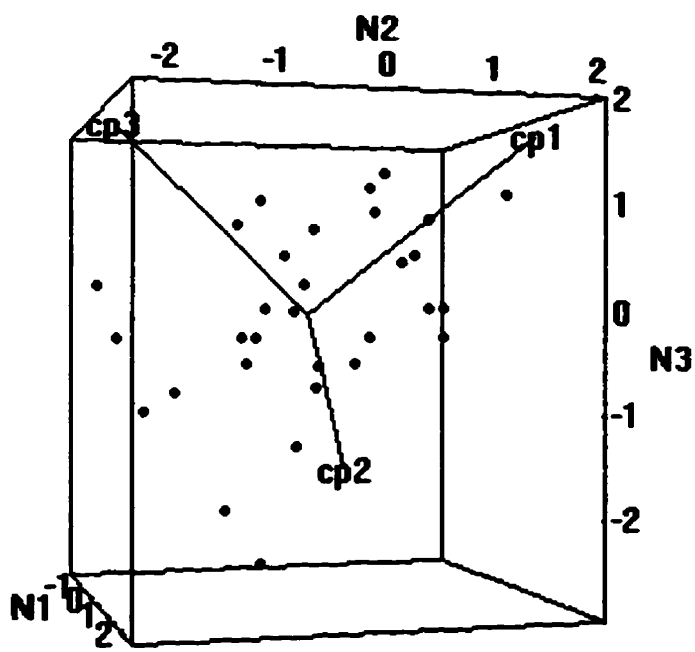


Figure 56. Projection des individus sur les variables initiales et sur les composantes principales

Il est évident que pour un tel tableau, l'intérêt de l'ACP est faible puisqu'il est tout à fait possible de représenter le tableau X dans l'espace de ses trois variables initiales et de visualiser les proximités entre les individus. Si nous traçons un graphique en trois dimensions

(figure 56) où nous plaçons les 31 points dans le repère des variables de départ N1, N2, N3, et que nous représentons ensuite les 3 axes composantes principales, nous constatons que ces derniers sont beaucoup plus adéquats pour décrire le nuage de points.

L'axe cp1 est l'axe de plus grande inertie pour le nuage de points. Ce qui signifie que si nous devions représenter notre échantillon par les coordonnées des individus sur un seul axe, nous choisirions cp1. Le deuxième axe de représentation choisi serait cp2.
L'axe cp1 est l'axe de plus grande inertie pour le nuage de points. Ce qui signifie que si nous devions représenter notre échantillon par les coordonnées des individus sur un seul axe, nous choisirions cp1. Le deuxième axe de représentation choisi serait cp2.

ANNEXE B

LES METHODES DE REGROUPEMENT

B. 1 Données initiales

Le tableau de donnée initiales est constitué comme dans le cas de l'ACP par un ensemble d'individus caractérisé par plusieurs variables. Le tableau ne doit pas posséder de structure a priori c'est à dire aucune dépendance fonctionnelles entre les variables. Nous avons vu que la première étape de l'ACP est de calculer les coefficients de corrélation entre les variables. Nos variables distances ne sont donc pas utilisables comme variables pour le regroupement puisqu'il existe plus ou moins des relations entre celles-ci. L'ensemble des individus pour le regroupement sera donc caractérisé par les variables composantes principales déterminées par l'ACP, ces dernières étant non-corrélées entre elles par construction.

B. 2 Définitions

n nombre d'individus

v nombre de variables

G nombre de classes (à n'importe quel niveau de la hiérarchie)

$x_i$ième observation (vecteur ligne)

C_KKième classe (sous-ensemble de {1, 2, ..., n})

N_K nombre d'individus dans C_K

$\bar{x}$ vecteur moyen de l'échantillon

$\bar{x}_K$ vecteur moyen de la classe C_K

$\|x\|$ longueur euclidienne du vecteur x (racine carrée de la somme des carrés des éléments

de x)

$$T\sum_{i=1}^{N} || x_i - \bar{x} ||^2$$

$$W_K\sum_{i\in C_k} || x_i - \overline{x_K} ||^2$$

P_G ΣW_J soit la somme sur les G classes au Gième niveau de la hiérarchie

B_{KL} $W_M - W_K - W_L$ si $C_M = C_K \cup C_L$

d(x, y) mesure de distance ou de dissimilarité entre les classes C_K et C_L

B.2.1 Calcul de d(x, y)

Soit deux individus k et l repérés dans l'espace des v variables qui les caractérisent (les composantes principales dans notre cas), les vecteurs représentant ces individus sont:

$x_k = (x_{k1}, x_{k2}, \ldots, x_{kv})$ et $x_L = (x_{l1}, x_{l2}, \ldots, x_{lv})$

Si on note $D(x_k, x_L)$ la distance entre les individus k et l, celle-ci peut être calculée comme:

$$D(x_k, x_l) = \sqrt{\sum_{j=1}^{v} (x_{kj} - x_{lj})^2} \qquad [14]$$

qui est la distance euclidienne entre les individus k et l dans l'espace à v dimensions.

Un autre distance est parfois utilisée comme mesure de similarité: la distance de Minkowski:

$$D_r(x_k, x_l) = \left(\sum_{j=1}^{v} |(x_{kj} - x_{lj})|^r\right)^{\frac{1}{r}} \qquad [15]$$

avec $r \geq 1$. Lorsque r=2 on obtient, comme cas particulier, la distance euclidienne.

On construit ainsi une matrice de distances. La matrice de départ sera donc une matrice (n, n) dans laquelle seront réunies les distances entre chaque individu par rapport à tous les

autres. Par la suite, lorsque les individus seront regroupés, la matrice sera de dimension G et les distances seront calculées entre les classes.

B.2.2 Différentes méthodes pour le calcul de regroupement hiérarchique

La distance entre deux classes peut être définie directement ou de manière itérative c'est à dire par une équation qui met à jour la matrice de distances à chaque fois que deux classes sont fusionnées. On suppose que les classes C_K et C_L sont fusionnées pour donner la classe C_M et les formules donnent la distance entre la classe C_M et n'importe quelle autre classe C_J. Les équations donnant la distance pour les différentes méthodes sont tirées des manuels de SAS puisque nous avons utilisé ce logiciel pour nos calculs et peuvent différer de ce qui est décrit dans la littérature pour la même méthode de classification .

B.2.2.1 Méthode hiérarchique AVERAGE

La distance entre deux classes est définie par:

$$D_{KL} = \sum_{i \in C_K} \sum_{j \in C_L} d(x_i, x_j) / (N_K N_L) \qquad [16]$$

Si d (x, y) = $\| x - y \|^2$ alors:

$$D_{KL} = || \overline{x_K} - \overline{x_L} ||^2 + W_K / N_K + W_K / N_L \qquad [17]$$

La distance entre la nouvelle classe C_M et une autre classe C_J est:

$$D_{JM} = (N_K D_{JK} + N_L D_{JL}) / N_M \qquad [18]$$

Dans ce calcul, la distance entre deux classes est la distance moyenne entre les paires d'individus pris deux à deux (un dans chaque classe). Cette méthode est biaisée dans le sens

qu'elle tend à fusionner les classes de faible variance (intra-classe) pour produire des classes de variances à peu près identique.

B.2.2.2 Méthode hiérarchique de WARD

La distance entre deux classes est définie par:

$$D_{KL} = B_{KL} = || \overline{x_K} - \overline{x_L} ||^2 / (1/N_K + 1/N_L) \qquad [19]$$

Si d (x, y) = $\| x- y\|^2/2$ alors:

$$D_{JM} = ((N_J + N_K) D_{JK} + (N_J + N_L) D_{JL} - N_J D_{KL}) / (N_J + N_M) \qquad [20]$$

Dans cette méthode, la distance entre deux classes est la somme des carrés entre les deux classes, sommée sur toutes les variables. A chaque étape de la hiérarchie, la somme des carrés à l'intérieur d'une classe est minimisée par rapport à toutes les manières de fusionner 2 classes de la génération précédente. Cette méthode est biaisée vers l'obtention de classes contenant le même nombre d'individus et tend à fusionner les classes avec un faible nombre d'observations.

B.2.2.2 Méthode hiérarchique du CENTROID

La distance entre deux classes est définie par:

$$D_{KL} = || \overline{x_K} - \overline{x_L} ||^2 \qquad [21]$$

Si d (x, y) = $\| x- y\|^2$ alors:

$$D_{JM} = (N_K D_{JK} + N_L D_{JL}) / N_M - N_K N_L D_{KL} / N_M^2 \qquad [22]$$

Dans cette méthode, la distance entre deux classes est définie comme la distance euclidienne entre leur moyennes ("centroids"). Elle est moins performante que les méthodes AVERAGE et WARD mais moins sensibles également aux observations aberrantes. Elle est d'ailleurs souvent utilisée pour détecter ces derniers qui ne feront partie d'aucune classe et resteront des individus isolés.

B.2.3 Classification non-hiérarchique FASTCLUS

Le principe général de la classification est le suivant: un ensemble d'individus (dont le total est égal au nombre prédéterminé de classes) appelés "germes" est généré. Ces germes constituent la moyenne hypothétique de chaque classe. A l'étape suivante, les distances euclidiennes entre les germes et tous les individus de l'échantillon sont calculées et chaque individu est affecté à la classe dont le germe est le plus proche. Les germes initiaux sont ensuite remplacés par la moyenne réelle des classes ainsi formées temporairement. Le processus est répété jusqu'à ce qu'il n'y ait plus aucun changement dans les classes. Il n'y a aucune hiérarchie de partition dans un tel regroupement puisque l'appartenance des individus aux classes peut changer à chaque étape de la classification.

B.2.4 Exemple simple de regroupement

Nous reprendrons ici l'exemple choisi pour expliquer l'ACP. Nos 31 individus sont maintenant repérés dans l'espace de 2 composantes principales. En effet, cp1 et cp2 retiennent 81.06% de toute l'information contenue par les variables initiales. Ces deux composantes devraient donc permettre de regrouper correctement les individus. Après l'examen des indices statistiques, nous décidons que les individus peuvent être regroupés en 6 classes distinctes. Dans le cas du regroupement hiérarchique, nous pouvons présenter les dendogrammes qui montrent le regroupement entre classes pour chaque étape de la hiérarchie.

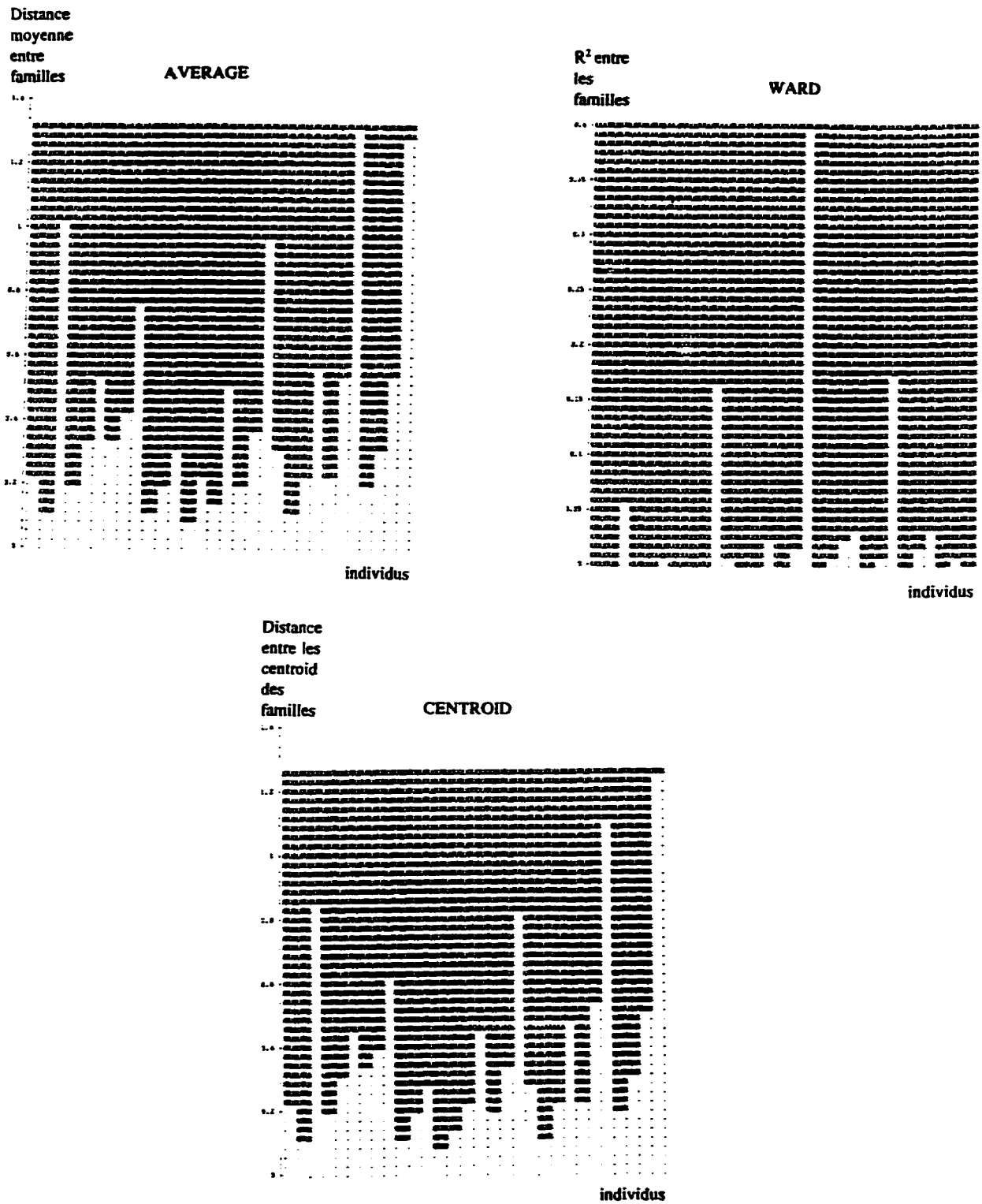


Figure 57. Dendrogramme indiquant les étapes de la classification hiérarchique pour les méthodes AVERAGE, WARD et CENTROID.

Nous pouvons constater que la manière d'affecter les individus aux classes diffère beaucoup selon la méthode utilisée. Cela traduit la manière différente de calculer les distances entre les individus ou entre les classes. Nous présentons les individus affectés aux classes à l'étape de regroupement en 6 classes qui est considérée optimale.

Tableau 32. Liste des individus affectés à chaque classe selon la méthode de classification employée.

FASTCLUS		AVERAGE		WARD				CENTROID				
CLA.	CONF.	CLA.	CONF.	CLA.	CONF.	CLA.	CONF.	CLA.	CONF.	N1	N2	N3
1	1	3	1	3	1			3	1	12.5	17.5	17.5
1	8	3	8	3	8			3	8	13.0	17.5	17.0
1	22	3	22	3	22			3	22	13.0	18.0	16.0
2	2	5	2	6	2			6	2	16.5	14.5	14.5
								6	5			
2	9	5	9	6	9			6	9	16.0	15.5	12.5
2	12									18.0	17.5	14.5
2	13	5	13	6	13			6	13	17.0	16.5	14.5
2	16	5	16	6	16			6	16	18.0	15.5	14.0
		5	19	6	19			6	19			
2	23	5	23	6	23			6	23	16.0	15.0	15.0
3	3	1	3			4	1	1	3	14.0	16.0	15.5
3	10	1	10	1	10			1	10	12.0	15.5	14.5
3	11	1	11			4	11	1	11	15.5	15.0	17.0
3	14	1	14	1	14			1	14	14.0	15.0	14.0
3	15	1	15			4	15	1	15	13.0	16.0	15.0
3	18	1	18			4	18	1	18	16.0	14.5	16.5
3	26	1	26			4	26	1	26	15.0	16.0	16.5
3	28	1	28			4	28	1	28	13.5	17.0	14.0
				1	29							
3	30	1	30	1	30			1	30	12.5	16.5	13.5
3	31	1	31			4	31	1	31	15.0	15.5	16.0
4	4	4	4	5	4			4	4	12.0	13.0	14.5
4	19									16.0	13.5	13.5
4	24	4	24	5	24			4	24	12.5	15.5	10.0
4	25	4	25	5	25			4	25	11.5	13.5	13.0
4	29	4	29					4	29	13.5	12.5	15.5
5	5	6	5					6	5	18.0	14.0	11.5
6	6	2	6	2	6			2	6	19.0	17.0	16.5
6	7	2	7	2	7			2	7	16.0	17.5	16.0
		2	12	2	12			2	12			
6	17	2	17	2	17			2	17	18.0	17.5	15.0
6	20	2	20	2	20			2	20	17.0	17.5	15.0
6	21	2	21	2	21			2	21	18.0	18.5	17.0
6	27	2	27	2	27			2	27	16.0	17.0	17.5

Nous constatons que les individus se trouvent regroupés approximativement de la même façon bien que le chemin conduisant à cette partition soit très différent selon les méthodes ce qui apparaît sur les dendogrammes.

L'individu 5 apparaît comme aberrant et est classé comme tel par les méthodes FASTCLUS, AVERAGE et CENTROID en formant une classe à un seul membre. En revanche, pour la méthode WARD, cet individu est placé dans une classe avec d'autres individus. La classification par cette méthode est réputée être perturbée par la présence d'individus aberrants. Nous avons donc retiré cet individu de l'échantillon et procédé à une nouvelle classification. Dans l'échantillon, 4 individus sont affectées à des classes différentes selon les méthodes: les individus 5, 12, 19 et 29. Pour cet échantillon de 31 individus, l'examen visuel permet de découvrir facilement les individus problématiques et éventuellement de les retirer de l'échantillon avant de procéder à une nouvelle analyse.

Pour un échantillon de grande taille, il est difficile d'échapper au problème des "outliers". En effet, les méthodes de classification exigeant des temps de calcul élevés (notamment les méthodes hiérarchiques), il sera parfois impossible de procéder à deux classifications successives, la première servant à détecter les "outliers". Les figures suivantes montrent la projection des individus sur les variables canoniques ce qui permet de voir l'éloignement relatif des classes les unes par rapport aux autres.

Nous constatons que les méthodes AVERAGE et CENTROID produisent des classes compactes et bien séparées alors que les méthodes FASTCLUS et WARD présentent des classes plus floues. Cette observation n'est pas surprenante: en effet, lors de la prévision du nombre de classe optimale par l'étude des indices statistiques, la partition en 6 classes a été suggérée par 2 sur 3 des indices statistiques pour AVERAGE et 3 sur 3 pour CENTROID. En revanche, 1 sur 3 indices statistique conseillait le choix de 6 classes comme optimal pour les méthodes FASTCLUS et WARD.

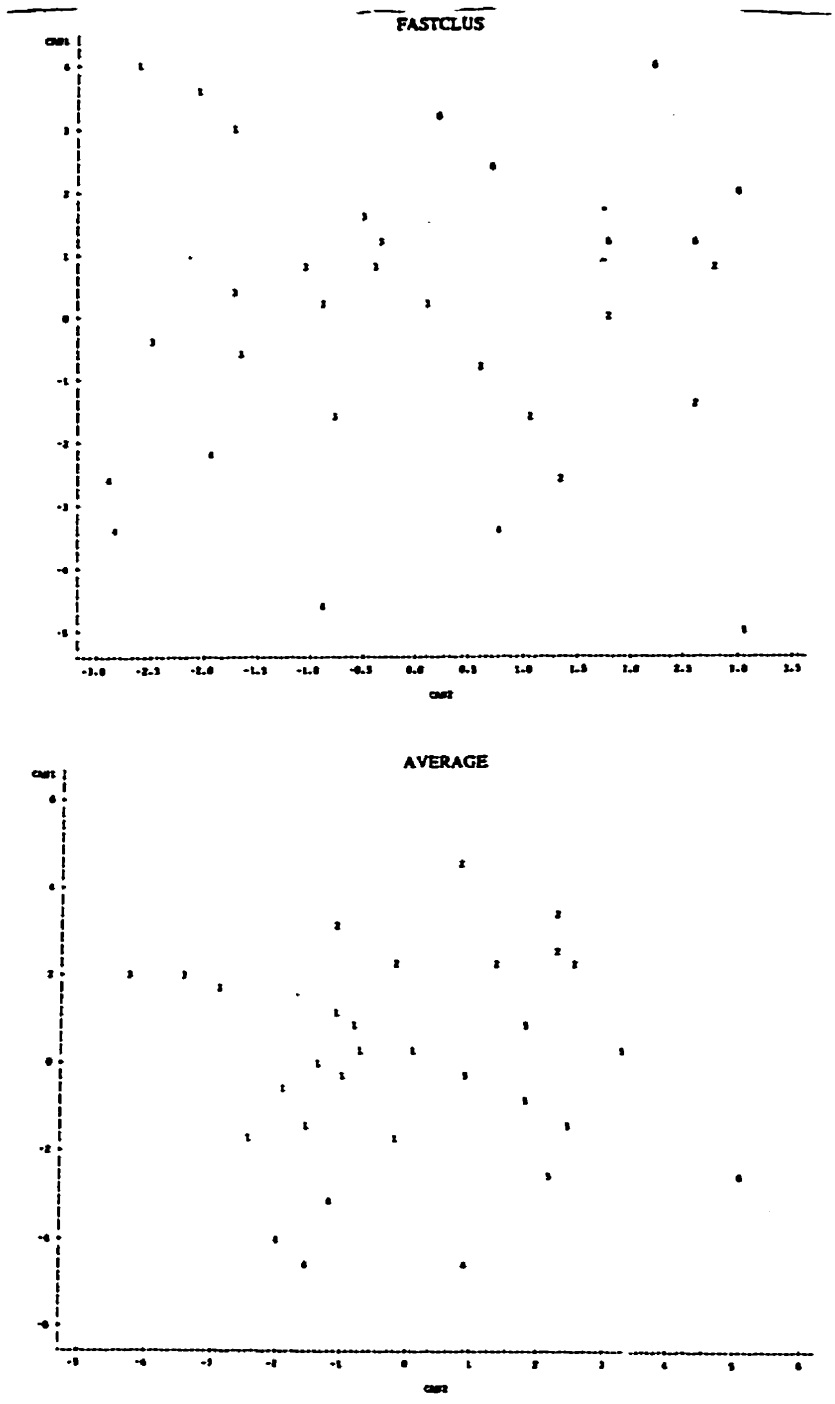


Figure 58. Projection des individus sur les variables canoniques pour les méthodes FASTCLUS et AVERAGE.

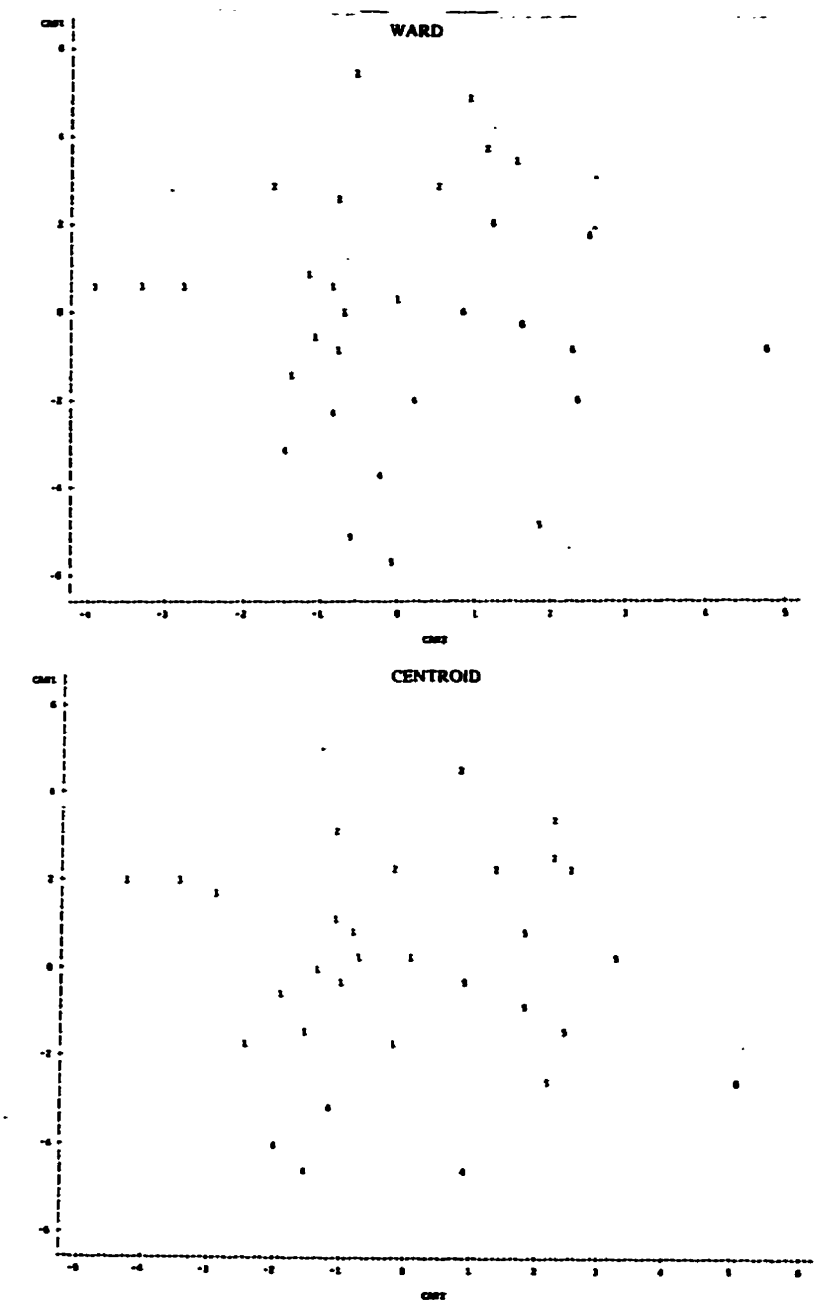


Figure 59. Projection des individus sur les variables canoniques pour les méthodes WARD et CENTROID.

Nous avons donc cohérence entre les prédictions des indices statistiques et la qualité de la séparation finale. Nous avons aussi observé la distorsion quant aux résultats du à la présence d'individus aberrants. Nous constatons également que la méthode choisie a une influence importante sur la partition obtenue. En effet, 4 des 31 individus ont été affectés à des classes différentes selon les méthodes choisies ce qui représente 13% des individus de l'échantillon. Ces difficultés de classification peuvent être reliées à la présence d'individus aberrants déjà mentionnée ce qui est une caractéristique propre à l'échantillon ou à des problèmes relatifs à la classification elle-même (la méthode choisie), ou au choix des variables initiales. En effet, les individus de départ sont caractérisés par leurs cordonnées dans l'espace des composantes principales dont certaines sont négligées. Une partie de la variabilité n'est donc plus prise en compte et peut nuire à l'étape de la classification.

ANNEXE C

FIGURES ET TABLEAUX DU CHAPITRE 3

Tableau 33. Résultats de l'ACP pour glmap: matrice des corrélations entre distances initiales.

Matrice des corrélations

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15
D1	1.0000	0.1098	0.0377	0.9243	0.1099	0.0458	0.6309	0.6759	0.0361	0.0337	-.0058	-.0091	-.0539	-.0498	0.0013
D2	0.1098	1.0000	0.1931	0.0591	0.8898	0.2587	0.0019	0.0335	0.6613	0.7361	0.1107	0.1499	-.0935	-.0195	0.0046
D3	0.0377	0.1931	1.0000	0.0153	0.1751	0.9400	-.0208	-.0068	0.0988	0.1130	0.8376	0.7659	0.1419	0.1524	0.1200
D4	0.9243	0.0591	0.0153	1.0000	0.0575	0.0175	0.7008	0.6256	-.0013	-.0003	-.0181	-.0242	-.0940	-.0073	-.0233
D5	0.1099	0.8898	0.1751	0.0575	1.0000	0.2388	0.0024	0.0345	0.7507	0.6564	0.0951	0.1315	-.0556	-.1064	0.0017
D6	0.0458	0.2587	0.9400	0.0175	0.2388	1.0000	-.0261	-.0110	0.1418	0.1570	0.7764	0.8055	0.1860	0.1751	0.0015
D7	0.6309	0.0019	-.0208	0.7008	0.0024	-.0261	1.0000	0.9326	0.0504	0.0501	-.0148	-.0139	-.0866	-.0801	-.0069
D8	0.6759	0.0335	-.0068	0.6256	0.0345	-.0110	0.9326	1.0000	0.1001	0.0991	0.0059	0.0118	-.0525	-.0462	0.0155
D9	0.0361	0.6613	0.0988	-.0013	0.7507	0.1418	0.0504	0.1001	1.0000	0.8983	0.1976	0.2640	-.0483	-.0884	-.0076
D10	0.0337	0.7361	0.1130	-.0003	0.6564	0.1570	0.0501	0.0991	0.8983	1.0000	0.2111	0.2794	-.0835	-.0133	-.0016
D11	-.0058	0.1107	0.8376	-.0181	0.0951	0.7764	-.0148	0.0059	0.1976	0.2111	1.0000	0.9463	0.1395	0.1512	0.1045
D12	-.0091	0.1499	0.7659	-.0242	0.1315	0.8055	-.0139	0.0118	0.2640	0.2794	0.9463	1.0000	0.1579	0.1716	-.0032
D13	-.0539	-.0935	0.1419	-.0940	-.0554	0.1860	-.0866	-.0525	-.0483	-.0835	0.1395	0.1579	1.0000	0.7716	-.0610
D14	-.0498	-.0195	0.1524	-.0873	-.1064	0.1751	-.0801	-.0462	-.0884	-.0133	0.1512	0.1716	0.7716	1.0000	-.0443
D15	0.0013	0.0046	0.1200	-.0233	0.0017	0.0015	-.0069	0.0155	-.0076	-.0016	0.1045	-.0032	-.0610	-.0443	1.0000
D16	-.0114	-.0056	0.0172	-.0433	-.0135	-.0450	-.0231	0.0057	-.0175	-.0076	0.0104	-.0468	-.0426	-.0220	0.9249
D17	-.0469	-.0774	0.1467	-.0564	-.0225	0.1705	-.0523	-.0454	-.0200	-.0697	0.1402	0.1594	0.9032	0.8602	-.1057
D18	-.0424	0.0148	0.1589	-.0501	-.0883	0.1831	-.0473	-.0397	-.0731	0.0175	0.1536	0.1747	0.6594	0.9097	-.0869
D19	0.0086	0.0100	0.1776	0.0008	0.0073	0.0463	0.0124	0.0201	-.0045	0.0012	0.1568	0.0368	-.1005	-.0836	0.9556
D20	-.0026	0.0001	0.0712	-.0115	-.0080	0.0028	0.0036	0.0122	-.0146	-.0045	0.0594	-.0042	-.1009	-.0791	0.8889
D21	0.0023	-.0198	0.1005	0.0007	-.0202	0.0350	-.0026	-.0026	-.0175	-.0151	0.0863	0.0263	-.0854	-.0672	0.1195
D22	0.0004	-.0459	0.0444	-.0014	-.0448	0.0142	0.0004	0.0010	-.0345	-.0348	0.0163	0.0083	-.1129	-.0940	0.0496
D23	0.0045	-.0126	0.1028	0.0017	-.0139	0.0382	-.0040	-.0043	-.0138	-.0285	0.0889	0.0291	-.0760	-.1150	0.1158
D24	0.0019	-.0660	0.0417	-.0009	-.0395	0.0144	-.0009	-.0012	-.0128	-.0557	0.0351	0.0092	-.1006	-.1594	0.0615
D25	-.1121	-.1475	-.0690	-.0960	-.1397	-.0889	-.0744	-.0865	-.1174	-.1241	-.0604	-.0772	-.0148	-.0310	-.0112
D26	0.0570	-.1192	-.0778	0.0660	-.1070	-.0972	0.0595	0.0525	-.0894	-.0991	-.0749	-.0894	-.0643	-.0879	-.0738
D27	0.0919	-.2186	0.1190	0.1063	-.2210	0.1320	0.0940	0.0821	-.1881	-.1836	0.1144	0.1274	0.1182	0.1110	-.1859
D28	0.0682	-.2053	0.0740	0.0819	-.2120	0.1187	0.0729	0.0605	-.1789	-.1727	0.0753	0.1160	0.1847	0.1805	-.2884
D29	-.0296	0.1501	0.0586	-.0284	0.1614	0.0757	-.0314	-.0394	0.1380	0.1263	0.0525	0.0673	-.1624	-.1784	-.1099
D30	-.0319	-.0037	0.2866	-.0176	-.0189	0.3366	-.0301	-.0433	-.0172	-.0029	0.2660	0.3007	0.0293	0.0509	-.1696
D31	-.0276	-.0379	0.1514	-.0151	-.0514	0.2203	-.0291	-.0414	-.0440	-.0324	0.1452	0.2040	0.0787	0.0985	-.2723
D32	0.0215	0.1304	0.0405	0.0131	0.1238	0.0400	0.0200	0.0276	0.1179	0.1232	0.0373	0.0827	-.0171	-.0189	0.0065
D33	0.0097	-.0285	0.0389	0.0136	-.0559	0.0597	0.0205	0.0189	-.0469	-.0204	0.0384	0.0561	-.0433	0.0154	0.1241
D34	0.0077	-.0146	-.0360	0.0106	-.0322	-.0086	0.0075	0.0054	-.0261	-.0091	-.0331	-.0083	-.0360	0.0076	-.0282
D35	0.2149	-.3505	-.1923	0.2237	-.3357	-.2420	0.2021	0.1949	-.3186	-.3304	-.1834	-.2277	0.0243	0.0160	0.0719
D36	0.2219	-.3350	-.1863	0.2339	-.3253	-.2332	0.2024	0.1899	-.3023	-.3093	-.1797	-.2210	0.0239	0.0113	0.0369
D37	-.0095	0.0001	-.0582	-.0103	0.0047	-.0428	-.0101	-.0088	0.0092	0.0114	-.0503	-.0348	0.0038	0.0114	-.0842
D38	-.0267	-.0693	0.1046	-.0260	-.0627	0.1320	-.0312	-.0331	-.0520	-.0592	0.1065	0.1333	0.2006	0.1630	-.1284
D39	0.1231	-.1558	0.0752	0.1295	-.1507	0.0978	0.0973	0.0878	-.1397	-.1445	0.0750	0.0878	0.1880	0.1549	-.1343
D40	0.2220	-.2194	-.0632	0.2284	-.2115	-.0791	0.1912	0.1805	-.1950	-.2013	-.0592	-.0713	0.1212	0.0923	-.2124
D41	-.0119	-.0003	-.0616	-.0068	-.0158	-.0232	0.0004	-.0028	-.0091	0.0059	-.0567	-.0226	-.0028	0.0239	-.1014
D42	-.0216	0.0903	0.1039	-.0110	-.1158	0.1768	-.0122	-.0216	-.0957	-.0719	0.1018	0.1660	0.1464	0.1706	-.2247
D43	0.0627	-.1841	0.1456	0.0767	-.2023	0.2093	0.0562	0.0424	-.1789	-.1619	0.1430	0.1989	0.2277	0.2288	-.3157
D44	0.1661	-.2278	0.0143	0.1782	-.2351	0.0406	0.1472	0.1349	-.2110	-.2032	0.0178	0.0019	0.1813	0.1717	-.2528

Tableau 34. Résultats de l'analyse en composantes principales pour glmap: matrice des corrélations entre distances initiales (suite).

	D16	D17	D18	D19	D20	D21	D22	D23	D24	D25	D26	D27	D28	D29	D30
D1	-.0114	-.0469	-.0424	0.0086	-.0026	0.0023	0.0004	0.0045	0.0019	-.1121	0.0570	0.0919	0.0682	-.0296	-.0319
D2	-.0056	-.0774	0.0148	0.0100	0.0001	-.0198	-.0459	-.0326	-.0660	-.1475	-.1192	-.2186	-.2053	0.1501	-.0037
D3	0.0172	0.1467	0.1589	0.1776	0.0712	0.1005	0.0444	0.1028	0.0417	-.0690	-.0779	0.1190	0.0740	0.0586	0.2866
D4	-.0433	-.0564	-.0501	0.0008	-.0115	0.0007	-.0014	0.0017	-.0009	-.0960	0.0660	0.1063	0.0819	-.0264	-.0176
D5	-.0135	-.0225	-.0883	0.0073	-.0080	-.0202	-.0448	-.0139	-.0395	-.1397	-.1070	-.2218	-.2120	0.1634	-.0189
D6	-.0450	0.1705	0.1831	0.0463	0.0028	0.0350	0.0142	0.0382	0.0144	-.0889	-.0972	0.1320	0.1187	0.0757	0.3366
D7	-.0231	-.0523	-.0473	0.0124	0.0036	-.0026	0.0004	-.0040	-.0009	-.0744	0.0595	0.0940	0.0729	-.0314	-.0301
D8	0.0057	-.0454	-.0397	0.0201	0.0122	-.0026	0.0010	-.0043	-.0012	-.0865	0.0525	0.0821	0.0605	-.0354	-.0433
D9	-.0175	-.0200	-.0731	-.0045	-.0146	-.0175	-.0345	-.0138	-.0328	-.1174	-.0894	-.1881	-.1789	0.1380	-.0172
D10	-.0076	-.0697	0.0175	0.0012	-.0045	-.0151	-.0348	-.0285	-.0557	-.1241	-.0991	-.1836	-.1727	0.1263	-.0029
D11	0.0104	0.1402	0.1536	0.1568	0.0594	0.0863	0.0363	0.0889	0.0351	-.0604	-.0749	0.1144	0.0753	0.0525	0.2680
D12	-.0468	0.1594	0.1747	0.0368	-.0042	0.0263	0.0083	0.0291	0.0092	-.0772	-.0894	0.1274	0.1168	0.0673	0.3087
D13	-.0426	0.9032	0.6594	-.1005	-.1009	-.0854	-.1129	-.0760	-.1006	-.0148	-.0643	0.1182	0.1847	-.1624	0.0293
D14	-.0220	0.6602	0.9097	-.0836	-.0791	-.0672	-.0940	-.1150	-.1554	-.0318	-.0879	0.1110	0.1805	-.1784	0.0509
D15	0.9249	-.1057	-.0869	0.9556	0.8889	0.1195	0.0496	0.1158	0.0415	-.0112	-.0738	-.1859	-.2884	-.1099	-.1696
D16	1.0000	-.1034	-.0798	0.8527	0.9380	0.0461	0.0185	0.0395	0.0084	-.0081	-.0896	-.2053	-.2984	-.1382	-.1815
D17	-.1034	1.0000	0.7203	-.0917	-.0870	-.0984	-.1306	-.0865	-.1150	-.0140	-.0608	0.1252	0.1917	-.1556	0.0331
D18	-.0798	0.7203	1.0000	-.0734	-.0632	-.0754	-.1075	-.1270	-.1730	-.0336	-.0889	0.1152	0.1865	-.1731	0.0581
D19	0.8527	-.0917	-.0734	1.0000	0.9174	0.1474	0.0662	0.1444	0.0582	-.0114	-.0689	-.1665	-.2786	-.0999	-.1463
D20	0.9380	-.0870	-.0632	0.9174	1.0000	0.0656	0.0326	0.0606	0.0233	-.0084	-.0860	-.1862	-.2909	-.1308	-.1563
D21	0.0461	-.0984	-.0754	0.1474	0.0656	1.0000	0.9121	0.8527	0.7361	-.0005	-.0047	-.0323	0.0674	-.0189	-.0541
D22	0.0185	-.1306	-.1075	0.0662	0.0326	0.9121	1.0000	0.7476	0.7763	0.0043	-.0012	-.0301	0.0755	-.0180	-.0572
D23	0.0395	-.0865	-.1270	0.1444	0.0606	0.8527	0.7476	1.0000	0.9141	-.0039	-.0118	-.0431	0.0554	-.0142	-.0641
D24	0.0084	-.1150	-.1730	0.0582	0.0233	0.7361	0.7763	0.9141	1.0000	0.0033	-.0070	-.0424	0.0631	-.0137	-.0739
D25	-.0081	-.0140	-.0336	-.0114	-.0084	-.0005	0.0043	-.0039	0.0033	1.0000	0.6872	0.4428	0.2734	0.0214	0.0131
D26	-.0896	-.0608	-.0889	-.0689	-.0860	-.0047	-.0012	-.0118	-.0070	0.6872	1.0000	0.7814	0.5237	0.2887	0.1783
D27	-.2053	0.1252	0.1152	-.1665	-.1862	-.0323	-.0301	-.0431	-.0424	0.4428	0.7814	1.0000	0.8410	0.2498	0.4854
D28	-.2984	0.1917	0.1865	-.2786	-.2909	0.0674	0.0755	0.0554	0.0631	0.2734	0.5237	0.8410	1.0000	0.1761	0.4733
D29	-.1382	-.1556	-.1731	-.0999	-.1308	-.0189	-.0180	-.0142	-.0137	0.0214	0.2887	0.2498	0.1761	1.0000	0.6778
D30	-.1815	0.0331	0.0581	-.1463	-.1563	-.0541	-.0572	-.0641	-.0739	0.0131	0.1783	0.4854	0.4733	0.6778	1.0000
D31	-.2682	0.0810	0.1038	-.2644	-.2613	0.0820	0.0862	0.0690	0.0689	0.0144	0.1163	0.4080	0.6613	0.4277	0.7933
D32	0.0077	-.0155	-.0176	0.0074	0.0101	-.0049	-.0064	-.0009	-.0015	-.1004	-.0398	-.0413	-.0334	-.0013	-.0193
D33	0.1862	-.0530	0.0156	0.1277	0.2019	-.1147	-.1349	-.1461	-.1762	-.0018	-.0085	0.1167	0.1370	-.0206	0.2481
D34	0.0234	-.0417	0.0072	-.0348	0.0216	0.0159	0.0027	-.0017	-.0183	0.0120	0.0093	0.0952	0.3053	0.0032	0.1974
D35	0.0943	0.0201	0.0102	0.0637	0.0866	-.0213	-.0160	-.0219	-.0163	0.1036	-.0341	0.0573	0.0518	-.0335	0.0460
D36	0.0512	0.0207	0.0062	0.0315	0.0460	-.0215	-.0178	-.0289	-.0241	0.3196	0.2928	0.2994	0.2169	-.0423	0.0011
D37	-.0758	-.0027	0.0054	-.0951	-.0889	-.0325	-.0393	-.0185	-.0179	0.0109	-.0016	0.0169	0.1363	0.0140	0.0492
D38	-.3905	0.2082	0.1642	-.3196	-.3950	-.0125	-.0026	-.0049	0.0122	0.0115	0.1702	0.1189	0.1625	0.0901	-.0822
D39	-.4016	0.1948	0.1564	-.3190	-.3965	-.0132	-.0055	-.0100	0.0018	0.0565	0.4093	0.4758	0.4602	0.3300	0.2480
D40	-.2522	0.1251	0.0914	-.2035	-.2497	-.0149	-.0080	-.0192	-.0094	0.1891	0.5527	0.5844	0.5020	0.1763	0.1227
D41	-.0609	-.0019	0.0289	-.1047	-.0620	0.1823	0.2017	0.1538	0.1604	-.0233	-.0232	0.1051	0.1045	-.0207	0.1612
D42	-.1958	0.1481	0.1770	-.2187	-.1879	0.1683	0.1855	0.1401	0.1471	-.0169	0.0420	0.3618	0.4864	0.0124	0.3792
D43	-.3218	0.2323	0.2343	-.3015	-.3097	0.1009	0.1170	0.0839	0.0953	0.0072	0.2034	0.6003	0.7105	0.1699	0.5341
D44	-.2598	0.1847	0.1737	-.2419	-.2514	0.0595	0.0726	0.0463	0.0579	0.1062	0.3744	0.6860	0.7238	0.1150	0.3520

Tableau 35. Résultats de l'analyse en composantes principales pour glmap: matrice des corrélations entre distances initiales (suite).

	D31	D32	D33	D34	D35	D36	D37	D38	D39	D40	D41	D42	D43	D44
D1	-.0276	0.0215	0.0097	0.0077	0.2149	0.2219	-.0095	-.0267	0.1231	0.2220	-.0119	-.0216	0.0627	0.1661
D2	-.0379	0.1304	-.0295	-.0146	-.3505	-.3350	0.0081	-.0693	-.1558	-.2194	-.0003	-.0903	-.1841	-.2278
D3	0.1514	0.0405	0.0399	-.0360	-.1923	-.1863	-.0582	0.1046	0.0752	-.0632	-.0616	0.1039	0.1456	0.0143
D4	-.0151	0.0131	0.0134	0.0106	0.2237	0.2339	-.0103	-.0260	0.1295	0.2296	-.0068	-.0110	0.0767	0.1782
D5	-.0514	0.1238	-.0559	-.0322	-.3357	-.3253	0.0047	-.0627	-.1507	-.2115	-.0158	-.1158	-.2023	-.2351
D6	0.2203	0.0600	0.0597	-.0086	-.2420	-.2332	-.0428	0.1320	0.0878	-.0791	-.0232	0.1768	0.2093	0.0406
D7	-.0291	0.0200	0.0205	0.0075	0.2021	0.2024	-.0101	-.0312	0.0973	0.1912	0.0004	-.0122	0.0562	0.1472
D8	-.0414	0.0276	0.0189	0.0054	0.1949	0.1899	-.0088	-.0331	0.0878	0.1805	-.0028	-.0216	0.0424	0.1349
D9	-.0440	0.1179	-.0469	-.0261	-.3186	-.3023	0.0082	-.0520	-.1397	-.1950	-.0091	-.0957	-.1789	-.2110
D10	-.0324	0.1232	-.0204	-.0091	-.3304	-.3093	0.0114	-.0592	-.1445	-.2013	0.0059	-.0719	-.1619	-.2032
D11	0.1452	0.0373	0.0394	-.0331	-.1834	-.1797	-.0503	0.1065	0.0750	-.0592	-.0567	0.1018	0.1430	0.0178
D12	0.2040	0.0527	0.0561	-.0083	-.2277	-.2210	-.0348	0.1333	0.0878	-.0713	-.0226	0.1660	0.1989	0.0419
D13	0.0797	-.0171	-.0433	-.0340	0.0243	0.0239	0.0038	0.2006	0.1880	0.1212	-.0028	0.1464	0.2277	0.1813
D14	0.0985	-.0189	0.0154	0.0076	0.0160	0.0113	0.0114	0.1630	0.1549	0.0923	0.0239	0.1706	0.2288	0.1717
D15	-.2723	0.0065	0.1243	-.0282	0.0719	0.0369	-.0842	-.3264	-.3343	-.2124	-.1014	-.2247	-.3157	-.2528
D16	-.2682	0.0077	0.1862	0.0234	0.0943	0.0512	-.0758	-.3905	-.4016	-.2522	-.0609	-.1958	-.3218	-.2598
D17	0.0810	-.0155	-.0530	-.0417	0.0201	0.0207	-.0027	0.2082	0.1948	0.1251	-.0019	0.1481	0.2323	0.1847
D18	0.1038	-.0176	0.0156	0.0072	0.0102	0.0062	0.0054	0.1642	0.1564	0.0914	0.0289	0.1770	0.2343	0.1737
D19	-.2644	0.0074	0.1277	-.0348	0.0637	0.0315	-.0951	-.3196	-.3190	-.2035	-.1047	-.2187	-.3015	-.2419
D20	-.2613	0.0101	0.2019	0.0216	0.0866	0.0460	-.0889	-.3950	-.3965	-.2497	-.0620	-.1879	-.3097	-.2514
D21	0.0820	-.0049	-.1147	0.0159	-.0213	-.0215	-.0325	-.0125	-.0132	-.0149	0.1823	0.1683	0.1009	0.0595
D22	0.0862	-.0064	-.1349	0.0027	-.0160	-.0178	-.0393	-.0026	-.0055	-.0080	0.2017	0.1855	0.1170	0.0726
D23	0.0690	-.0009	-.1461	-.0017	-.0219	-.0289	-.0185	-.0049	-.0100	-.0192	0.1538	0.1401	0.0839	0.0463
D24	0.0689	-.0015	-.1762	-.0183	-.0163	-.0241	-.0179	0.0122	0.0018	-.0094	0.1604	0.1471	0.0953	0.0579
D25	0.0144	-.1004	-.0018	0.0120	0.1036	0.3196	0.0109	0.0115	0.0565	0.1891	-.0233	-.0169	0.0072	0.1062
D26	0.1163	-.0398	-.0085	0.0093	-.0341	0.2928	-.0016	0.1702	0.4093	0.5527	-.0232	0.0420	0.2034	0.3744
D27	0.4080	-.0413	0.1167	0.0952	0.0573	0.2994	0.0169	0.1189	0.4758	0.5844	0.1051	0.3618	0.6003	0.6860
D28	0.6613	-.0334	0.1370	0.3053	0.0518	0.2169	0.1363	0.1625	0.4602	0.5020	0.1045	0.4864	0.7105	0.7238
D29	0.4277	-.0013	-.0206	0.0032	-.0335	-.0423	0.0140	0.0901	0.3300	0.1763	-.0207	0.0124	0.1699	0.1150
D30	0.7933	-.0193	0.2481	0.1974	0.0460	0.0011	0.0492	-.0822	0.2480	0.1227	0.1612	0.3792	0.5341	0.3520
D31	1.0000	-.0176	0.2558	0.4949	0.0301	-.0021	0.2163	-.0281	0.2154	0.1053	0.0927	0.4342	0.5565	0.3700
D32	-.0176	1.0000	0.0103	0.0049	-.2686	-.0467	0.0030	-.0102	-.1104	-.0102	0.0054	-.0060	-.0627	-.0055
D33	0.2558	0.0103	1.0000	0.7225	0.0302	0.0188	0.0636	-.1073	-.1260	-.0767	0.1170	0.3213	0.1365	0.0744
D34	0.4949	0.0049	0.7225	1.0000	0.0107	0.0078	0.3428	-.0441	-.0844	-.0579	-.1047	0.2108	0.0851	0.0426
D35	0.0301	-.2686	0.0302	0.0107	1.0000	0.7750	0.0090	-.2540	0.0140	0.2160	0.0010	-.0914	0.0380	0.1960
D36	-.0021	-.0467	0.0188	0.0078	0.7750	1.0000	-.0002	-.1141	0.1737	0.5289	0.0007	-.0383	0.1199	0.4093
D37	0.2163	0.0030	0.0636	0.3428	0.0090	-.0002	1.0000	-.0262	-.0273	-.0257	-.0991	-.0911	-.0584	-.0384
D38	-.0281	-.0102	-.1073	-.0441	-.2540	-.1141	-.0262	1.0000	0.7628	0.4790	-.2890	-.0585	0.1585	0.1772
D39	0.2154	-.1104	-.1260	-.0844	0.0140	0.1737	-.0273	0.7628	1.0000	0.8442	-.1476	0.0931	0.4821	0.5775
D40	0.1053	-.0102	-.0767	-.0579	0.2160	0.5289	-.0257	0.4790	0.8442	1.0000	-.0780	0.0776	0.4414	0.7222
D41	0.0927	0.0054	0.1170	-.1047	0.0010	0.0007	-.0991	-.2890	-.1476	-.0780	1.0000	0.7910	0.5095	0.3350
D42	0.4342	-.0060	0.3213	0.2108	-.0914	-.0383	-.0911	-.0585	0.0931	0.0776	0.7910	1.0000	0.8330	0.5915
D43	0.5565	-.0627	0.1365	0.0851	0.0380	0.1199	-.0584	0.1585	0.4821	0.4414	0.5095	0.8330	1.0000	0.8709
D44	0.3700	-.0055	0.0744	0.0426	0.1960	0.4093	-.0384	0.1772	0.5775	0.7222	0.3350	0.5915	0.8709	1.0000

Tableau 36. Résultats de l'analyse en composantes principales pour glmap: valeurs propres associées à chaque composante principale.

Valeurs propres

	Valeurs propres	Pourcentages de variance	Pourcentages cumulés
cp1	7.21491	0.163975	0.16398
cp2	4.97756	0.113126	0.27710
cp3	4.03987	0.091815	0.36892
cp4	3.76236	0.085508	0.45443
cp5	3.57921	0.081346	0.53577
cp6	3.07217	0.069822	0.60559
cp7	2.63257	0.059831	0.66542
cp8	2.09720	0.047664	0.71309
cp9	1.79903	0.040887	0.75397
cp10	1.42200	0.032318	0.78629
cp11	1.38187	0.031406	0.81770
cp12	1.06326	0.024165	0.84186
cp13	0.98905	0.022479	0.86434
cp14	0.83889	0.019066	0.88341
cp15	0.75566	0.017174	0.90058
cp16	0.60936	0.013849	0.91443
cp17	0.56762	0.012900	0.92733
cp18	0.47770	0.010857	0.93819
cp19	0.34517	0.007845	0.94603
cp20	0.33393	0.007589	0.95362
cp21	0.23666	0.005379	0.95900
cp22	0.20694	0.004703	0.96370
cp23	0.19893	0.004521	0.96823
cp24	0.19334	0.004394	0.97262
cp25	0.18463	0.004196	0.97682
cp26	0.16326	0.003711	0.98053
cp27	0.14821	0.003368	0.98389
cp28	0.11678	0.002654	0.98655
cp29	0.10963	0.002491	0.98904
cp30	0.10356	0.002354	0.99139
cp31	0.07892	0.001794	0.99319
cp32	0.06253	0.001421	0.99461
cp33	0.05724	0.001301	0.99591
cp34	0.05368	0.001220	0.99713
cp35	0.04438	0.001009	0.99814
cp36	0.01884	0.000428	0.99857
cp37	0.01337	0.000304	0.99887
cp38	0.01124	0.000256	0.99913
cp39	0.01053	0.000239	0.99936
cp40	0.00994	0.000226	0.99959
cp41	0.00614	0.000140	0.99973
cp42	0.00561	0.000128	0.99986
cp43	0.00478	0.000109	0.99997
cp44	0.00146	0.000033	1.00000

Tableau 37. Résultats de l'analyse en composantes principales pour glmap: vecteurs propres associés à chaque composante principale.

	cp1	cp2	cp3	cp4	cp5	cp6	cp7	cp8	cp9	cp10	cp11
D1	0.041805	-.065115	0.070658	-.180730	0.372758	0.150209	0.121029	-.063139	0.036972	-.021241	-.000641
D2	-.110423	0.259638	-.039881	-.215206	0.079917	-.022645	0.046161	0.283666	0.021836	0.061344	0.015314
D3	0.041141	0.311245	0.144690	0.103805	0.125200	0.027829	-.174655	-.225432	-.052893	0.022638	-.111374
D4	0.049264	-.082452	0.071437	-.184875	0.365549	0.141151	0.121288	-.088433	0.029624	-.028179	-.005646
D5	-.113748	0.248363	-.040054	-.228440	0.073628	-.016585	0.027839	0.283444	0.020982	0.078368	0.018376
D6	0.061502	0.338553	0.109711	0.075213	0.114120	0.011458	-.125583	-.215182	-.057757	0.010233	-.110770
D7	0.039952	-.084830	0.069840	-.179148	0.363053	0.139726	0.132590	-.077012	0.032989	-.054162	-.035382
D8	0.032636	-.068477	0.068562	-.174804	0.368791	0.145822	0.132364	-.052973	0.038640	-.048384	-.036630
D9	-.103769	0.251346	-.036021	-.222871	0.074245	-.016450	0.024674	0.296433	0.020136	0.072013	-.035367
D10	-.100856	0.260139	-.034283	-.210249	0.079916	-.023018	0.041274	0.296978	0.020882	0.055659	-.036910
D11	0.041521	0.313209	0.136599	0.101882	0.120133	0.024503	-.177787	-.222366	-.057570	0.022954	-.148573
D12	0.059654	0.336629	0.103187	0.074318	0.109302	0.009583	-.133877	-.207900	-.061244	0.010479	-.147284
D13	0.117666	0.110995	-.126963	0.292731	0.031165	0.228020	0.098180	0.166458	0.107260	0.088245	0.014394
D14	0.112748	0.121763	-.120515	0.301296	0.042667	0.203280	0.134806	0.163481	0.114459	0.068531	0.005416
D15	-.201455	-.042414	0.220398	0.232765	0.171486	-.079574	-.105716	0.146109	-.003157	-.055419	0.164303
D16	-.206455	-.061409	0.192654	0.242778	0.163150	-.110788	-.043979	0.162057	-.004352	-.059901	0.140193
D17	0.120065	0.115700	-.133593	0.273400	0.041210	0.227492	0.098594	0.165914	0.096859	0.089198	0.007437
D18	0.114495	0.127923	-.125074	0.283185	0.052739	0.199465	0.140765	0.161645	0.105844	0.066981	-.002376
D19	-.195191	-.029049	0.236649	0.223803	0.177435	-.075128	-.120671	0.123841	-.009112	-.055023	0.153083
D20	-.203581	-.049532	0.213012	0.236191	0.175251	-.114198	-.056574	0.140164	-.013372	-.065400	0.130475
D21	0.002484	0.002755	0.371176	-.062439	-.182127	0.216554	-.002325	0.068101	0.141062	0.037784	-.004387
D22	0.010176	-.013098	0.354533	-.083141	-.198281	0.211793	0.008700	0.050329	0.116262	0.031751	-.011890
D23	-.002536	0.000241	0.368499	-.071530	-.185481	0.223032	-.021539	0.049486	0.144115	0.047314	-.002297
D24	0.004712	-.018531	0.347152	-.094471	-.202187	0.219183	-.017231	0.026407	0.121525	0.041079	-.010446
D25	0.080279	-.113835	0.019303	0.036766	-.018877	-.114457	-.236428	0.221584	0.012248	0.043583	-.524461
D26	0.174005	-.114719	0.036533	-.066705	0.052967	-.134678	-.307216	0.282134	0.000876	-.091272	-.246813
D27	0.287227	-.036762	0.071939	0.018386	0.083657	-.150580	-.153687	0.156816	-.041938	-.049404	-.165889
D28	0.308150	-.004731	0.084467	0.010731	0.013209	-.117344	-.006974	0.106052	0.111503	-.054013	-.089814
D29	0.095951	0.068840	0.030276	-.162920	-.006346	-.225021	-.163077	0.048108	0.023998	0.312838	0.389903
D30	0.199153	0.136478	0.088967	-.010882	0.023753	-.294581	0.020195	-.073605	0.004661	0.294327	0.222547
D31	0.217585	0.106969	0.099490	-.025455	-.048712	-.239498	0.128820	-.064288	0.238407	0.194739	0.118236
D32	-.027548	0.065983	-.001076	-.050278	0.017169	-.000877	0.023253	0.058268	-.008810	-.252879	0.014806
D33	0.029568	0.019506	0.065609	0.096523	0.085481	-.299751	0.213731	-.042561	0.213782	-.304260	-.012186
D34	0.053751	0.012282	0.058190	0.021276	0.009322	-.256099	0.203617	-.052395	0.500230	-.210322	-.065357
D35	0.047301	-.243114	0.041310	0.088478	0.139085	0.017154	0.076650	-.094880	0.015588	0.482374	-.033989
D36	0.109893	-.259461	0.042412	0.061741	0.160003	0.010801	-.044774	0.051395	-.011160	0.292818	-.140444
D37	0.017525	-.001501	-.037450	-.027475	-.020576	-.092048	0.066971	-.040161	0.412113	0.058163	-.130913
D38	0.141465	0.056563	-.140388	-.049467	-.052369	0.205942	-.264399	-.047719	0.147273	-.319578	0.219403
D39	0.260988	-.013194	-.066209	-.067220	0.039848	0.122514	-.261473	0.027201	0.051771	-.109061	0.322772
D40	0.251108	-.132701	-.020285	-.050359	0.117732	0.089986	-.226878	0.133124	0.002771	-.090727	0.163585
D41	0.079805	0.015044	0.152932	-.011269	-.102078	-.034864	0.333273	0.119061	-.419493	-.039426	-.062584
D42	0.207507	0.082035	0.162728	0.041441	-.092621	-.082858	0.319925	0.052041	-.249759	-.182213	-.016591
D43	0.309008	0.052986	0.116907	0.029042	-.024455	-.038465	0.151040	0.037480	-.198916	-.071167	0.128481
D44	0.306149	-.059815	0.089559	0.007598	0.058374	-.001739	0.043132	0.123088	-.159659	-.078320	0.098438

Tableau 38. Résultats de l'analyse en composantes principales pour glmap: vecteurs propres associés à chaque composante principale (suite).

	cp12	cp13	cp14	cp15	cp16	cp17	cp18	cp19	cp20	cp21	cp22
D1	-.079728	-.045847	0.006025	-.351910	-.001305	-.011503	0.369807	0.043963	0.016691	0.152316	0.046270
D2	0.116599	0.017434	0.006532	-.304028	0.021579	-.153121	-.277992	0.051711	-.066249	0.034565	-.195220
D3	0.043108	0.036417	-.001534	-.189351	0.059551	0.009991	-.226689	-.075470	-.011338	-.107145	0.219157
D4	-.095966	-.057038	0.010009	-.323826	-.002469	-.021200	0.387511	0.066967	0.029165	-.154628	0.016877
D5	0.112843	0.020563	-.001433	-.306113	-.028877	0.191125	-.266934	-.022373	0.059237	0.037475	-.222480
D6	0.033079	0.032483	0.001099	-.218319	0.043163	0.024156	-.243692	0.024681	0.001952	0.102247	0.355728
D7	-.168208	-.119539	0.025366	0.323698	0.060787	0.030720	-.336939	0.016599	0.001322	-.124400	-.056786
D8	-.151043	-.110778	0.020460	0.357313	0.059443	0.034739	-.335919	-.007215	-.008682	0.148287	-.020700
D9	0.170333	0.023883	-.025482	0.300793	-.104882	0.142382	0.232295	-.003330	0.065859	-.048158	0.252241
D10	0.174275	0.020793	-.018353	0.299767	-.057271	-.170198	0.217990	0.061441	-.051293	-.054842	0.262023
D11	0.093435	0.047802	-.025834	0.175197	0.024277	-.007727	0.196878	-.067461	-.005394	-.031599	-.410764
D12	0.091501	0.042561	-.028371	0.219408	0.001519	-.001176	0.237153	0.022134	0.006618	0.153339	-.272937
D13	-.098468	-.018347	0.043736	-.000616	-.003511	0.422588	0.047569	-.105795	0.092155	0.327220	0.002795
D14	-.074326	-.036571	0.070203	0.006456	0.108051	-.401982	0.006856	0.006524	-.146759	0.330627	-.010034
D15	-.019185	-.013707	-.098741	0.008140	-.006662	0.002563	-.004812	0.095689	0.039258	0.023860	-.061735
D16	-.014430	-.012903	-.070983	0.015550	-.006761	0.012528	0.015784	0.095333	0.021895	0.394056	0.077589
D17	-.097177	-.023165	0.044732	-.011252	-.024301	0.469096	0.034812	-.091931	0.121660	-.349525	-.013769
D18	-.069330	-.043337	0.076508	-.004450	0.101376	-.454649	-.008864	0.034448	-.151346	-.323731	-.028059
D19	-.014567	-.013917	-.087028	-.002323	-.008620	-.000409	-.014135	0.057580	0.029666	-.349245	-.075164
D20	-.010622	-.014657	-.059709	0.004755	-.012771	0.010881	0.007110	0.066337	0.011451	-.054748	0.080663
D21	0.010064	-.010845	0.077498	-.005647	0.045599	-.147444	-.010178	-.203713	0.375974	-.154218	-.060821
D22	0.003867	-.012876	0.077922	0.011982	0.036174	-.162898	-.005329	-.158972	0.484774	0.178139	0.126228
D23	-.007615	0.005827	0.036489	-.010157	-.007470	0.116027	0.012298	0.034853	-.428609	-.168786	-.070360
D24	-.017824	0.009197	0.022685	0.007520	-.032101	0.169114	0.021026	0.144095	-.514109	0.157443	0.104978
D25	-.076200	-.149612	0.165927	-.025190	0.237482	0.058120	-.002515	0.490331	0.125396	-.054655	0.000861
D26	-.164194	-.122957	0.052464	-.034958	0.112623	-.001451	0.058965	-.227136	-.065215	0.066770	-.064754
D27	-.144590	-.034520	-.045369	-.017734	-.170726	-.042044	0.005064	-.330615	-.107192	-.017839	0.171149
D28	-.094514	0.013945	-.126894	-.002006	-.395347	-.046415	-.051738	0.048204	0.034018	0.009749	0.020315
D29	-.218564	-.074504	0.155085	0.042942	0.363425	0.052171	0.075956	-.112715	-.050107	0.058849	-.154140
D30	-.199524	-.031404	0.119763	0.024874	0.078724	-.008049	0.026550	-.006285	-.014906	-.070307	0.196286
D31	-.144699	0.004802	0.028132	0.012712	-.216789	-.021345	-.018716	0.355364	0.140350	0.008828	-.077305
D32	-.355665	0.814671	0.254085	0.040448	0.100144	0.010683	0.015292	0.095478	0.030393	-.021491	0.044479
D33	0.291475	-.073460	0.332470	-.004468	0.216579	0.068778	0.026707	-.275622	-.118264	-.064936	0.209169
D34	0.200227	-.008910	0.144131	-.018683	-.037751	0.042930	-.004021	0.019724	0.006615	0.059380	-.265282
D35	0.339708	0.118229	0.086850	0.029224	0.032604	0.027729	-.040235	0.068814	0.013238	-.013799	0.137376
D36	0.330034	0.340567	0.116590	0.004247	0.090604	0.008457	-.049388	0.033679	0.013219	0.027450	-.053299
D37	-.064258	0.182698	-.756640	-.002953	0.381209	0.019867	0.008750	-.048990	0.007458	-.028481	0.088325
D38	0.165391	-.102814	0.055021	0.032365	0.216930	0.053585	-.000100	0.347715	0.111511	-.010577	0.152998
D39	0.162549	-.055875	-.052504	0.018872	0.119669	-.003378	0.000351	0.076808	0.019045	-.013097	0.021036
D40	0.231363	0.172156	-.053725	0.000222	0.036573	-.029122	-.026830	-.147402	-.041621	0.023994	-.124921
D41	0.032050	-.031326	-.106580	-.002626	0.425479	0.059890	0.051525	-.020211	0.002140	0.010910	-.057715
D42	0.071039	-.064310	-.030672	-.006252	0.118558	0.059566	0.007910	0.151915	0.038314	0.009916	-.051711
D43	0.064455	-.007796	-.136944	0.004335	-.090646	0.011415	-.030588	0.174446	0.041005	-.025992	0.031678
D44	0.167444	0.167805	-.157748	-.001899	-.158131	-.017314	-.059902	-.022341	-.021869	-.005535	-.055640

Tableau 39. Résultats de l'analyse en composantes principales pour glmap: vecteurs propres associés à chaque composante principale (suite).

	cp23	cp24	cp25	cp26	cp27	cp28	cp29	cp30	cp31	cp32	cp33
D1	-.093149	0.016050	0.019091	-.032271	0.006018	-.443405	-.149349	-.141827	-.027996	-.012966	-.009936
D2	0.082828	-.018985	0.160492	0.057065	-.484991	0.010435	0.037174	-.033348	0.022315	-.008256	0.004004
D3	-.236917	0.048519	-.254195	-.128852	0.025000	-.023727	0.169212	-.032202	-.075396	0.093181	0.272769
D4	0.060708	-.042382	-.029702	-.008495	0.003119	0.443801	0.131915	0.119205	-.001023	0.016764	0.008088
D5	0.063926	0.028311	0.219139	0.102565	0.444976	0.034756	-.027063	-.004727	0.045081	0.000478	-.001644
D6	0.095940	0.031525	-.220177	-.067904	0.019814	0.011460	-.114190	0.139872	-.001694	-.064838	-.235716
D7	0.078101	-.045943	-.016823	0.023000	-.003735	0.422996	0.116872	0.118162	0.011164	0.017032	0.004898
D8	-.068006	0.013216	0.027956	-.005703	-.001445	-.413717	-.141122	-.127170	-.004152	-.017244	-.002294
D9	-.093949	0.040050	-.103297	-.031141	0.470186	0.021654	-.008130	0.025661	0.008590	-.002041	0.012049
D10	-.072966	-.006834	-.144116	-.069105	-.447170	-.003858	0.049708	-.004816	-.006141	-.001202	0.008268
D11	-.038588	-.002969	0.149720	0.038686	-.011663	-.019450	0.125860	-.080392	-.034754	0.086367	0.255155
D12	0.255333	-.014079	0.171520	0.085087	-.020070	0.015849	-.137894	0.078779	0.034366	-.069395	-.246540
D13	-.230569	-.078508	0.021104	-.002191	-.194518	0.144398	0.075046	-.002852	-.017237	-.242484	0.041935
D14	-.216653	-.027057	0.047190	-.013709	0.196656	0.147649	0.053383	0.005739	0.000798	-.222878	0.042251
D15	-.308154	0.077439	0.054216	-.001470	0.002273	0.069405	0.138323	-.200343	0.034763	0.413166	-.231403
D16	0.138631	-.008686	-.022322	0.022102	-.021558	0.066047	-.159974	0.287560	-.001883	0.427763	0.050145
D17	0.248832	0.009952	-.080285	-.001002	-.149913	-.141660	-.037489	-.025171	0.000665	0.239719	-.036690
D18	0.246077	0.052955	-.055739	-.019044	0.177831	-.132778	-.049587	-.020905	0.005786	0.223535	-.023000
D19	-.165242	0.063751	0.030721	-.001042	0.020627	-.045340	0.164693	-.287804	0.018032	-.390517	-.065640
D20	0.398950	-.024752	-.059307	0.034162	0.001003	-.072570	-.146602	0.199266	-.019326	-.476399	0.208247
D21	-.231131	-.029939	0.066561	-.022977	-.045075	0.000446	-.282980	0.271042	0.017363	-.012953	-.029266
D22	0.294985	-.061183	0.008161	0.003070	0.016145	0.003234	0.265240	-.271962	-.002258	0.022898	0.004094
D23	-.249963	-.006884	0.045398	0.011063	-.023351	0.002437	-.268662	0.254506	0.009089	-.010752	-.019121
D24	0.266967	-.030353	-.015864	0.036216	0.045040	0.011129	0.246499	-.263125	-.010591	0.030157	0.013926
D25	-.049327	-.329613	0.082834	-.113889	0.034293	-.073856	0.078024	0.063626	-.156709	-.015927	-.143792
D26	0.021240	0.183378	-.152487	0.118396	-.016394	-.036000	0.035768	0.058319	0.337681	0.051269	0.392620
D27	0.024056	0.187623	0.216075	0.039077	-.029411	0.054113	-.018279	-.052251	0.193073	-.031624	-.242203
D28	0.025426	0.364350	0.155532	-.091724	-.019815	0.056322	-.020774	0.007696	-.551243	-.024569	-.075497
D29	0.086782	0.110173	-.067754	-.509227	0.011784	0.034129	-.032354	-.009720	-.227149	-.009574	-.002233
D30	-.046032	-.238664	0.122526	0.303499	-.000377	0.030268	-.070910	-.107975	0.259122	-.002960	-.199344
D31	-.068654	0.021015	-.183949	0.382525	0.001855	-.031927	-.020614	0.002085	-.090251	0.039751	0.328885
D32	-.013611	0.061639	0.105988	-.009399	0.003627	-.057185	0.108585	0.097342	0.042724	-.003648	-.001406
D33	-.004593	-.199399	0.306501	0.098054	-.000324	0.023305	-.038572	-.070604	-.286844	0.072788	0.160368
D34	0.002770	0.133673	-.339463	-.207683	0.013932	-.063407	0.157886	0.090973	0.283050	-.061006	-.223298
D35	0.005641	0.294901	0.320340	-.061655	-.020487	-.162411	0.334476	0.315057	0.144838	-.002532	0.053400
D36	0.023012	0.083885	-.208358	0.052569	-.025041	0.252724	-.425700	-.384539	0.002354	-.015465	0.000640
D37	0.020264	-.064632	0.100512	-.010628	0.000521	0.015073	-.032720	-.026097	0.009976	0.005554	0.024257
D38	0.015433	0.338811	0.215384	-.003021	-.031932	0.128225	-.202880	-.143127	0.132795	0.005628	0.148422
D39	-.025114	-.026858	-.010040	0.205592	0.002044	-.123112	0.204064	0.196884	-.034446	-.044163	-.153487
D40	-.004409	-.223126	-.285763	0.240742	0.029495	-.076336	0.115410	0.095014	-.209463	-.002839	-.188792
D41	-.036624	0.276995	-.134762	0.277690	-.003720	-.039048	0.101229	0.081913	-.195135	-.011836	-.004621
D42	-.007949	0.143293	-.072865	-.113045	-.006518	0.046457	-.054166	-.070463	0.142128	-.042480	-.144795
D43	-.011091	-.183801	0.124824	-.213453	0.006918	-.021234	-.009290	0.004429	0.267573	0.014578	0.119212
D44	0.017819	-.356962	0.049270	-.338806	0.039840	-.020293	-.009412	0.000150	0.033453	0.057752	0.224915

Tableau 40. Résultats de l'analyse en composantes principales pour glmap: vecteurs propres associés à chaque composante principale (suite).

	cp34	cp35	cp36	cp37	cp38	cp39	cp40	cp41	cp42	cp43	cp44
D1	0.026221	-.015627	0.001216	-.002777	-.004686	0.474407	-.028550	-.003515	-.001494	0.017367	-.001250
D2	0.013880	-.019331	0.009264	-.011488	0.495025	-.007254	0.039163	-.000639	-.043716	0.024463	-.000165
D3	0.223557	-.189422	0.009451	0.002525	-.009642	-.021030	0.007953	0.000879	0.001345	0.499966	-.031255
D4	-.002551	0.016043	0.006143	0.003320	0.002648	-.489196	0.030309	0.003116	0.000838	-.014288	0.001525
D5	0.021121	-.037464	-.002865	0.013838	-.485729	-.021380	-.036067	0.001526	0.043836	0.019147	-.002082
D6	-.277066	0.245385	-.020630	-.002896	0.000415	0.008002	-.006472	0.003693	-.001011	-.478721	0.028582
D7	-.011832	-.000058	-.005323	0.000416	-.003961	0.520920	-.033620	-.004980	0.000521	0.014611	0.006057
D8	0.019177	-.023380	0.004134	-.000067	0.005340	-.507905	0.032001	0.001766	0.000233	-.016613	-.005667
D9	0.007278	-.021527	0.000168	-.015237	0.500269	0.020054	0.038247	0.000451	0.005192	-.016507	0.000616
D10	0.001895	-.001642	0.000355	0.019847	-.509920	0.008414	-.038391	0.002321	-.005023	-.027167	-.000824
D11	0.241602	-.213626	0.013157	0.002224	0.006570	0.021483	0.002859	0.001928	-.003640	-.517169	-.005977
D12	-.247247	0.220281	-.013439	-.009864	0.000536	-.008664	-.004814	-.002099	0.003273	0.499503	0.006582
D13	-.001522	0.010669	0.003589	-.007119	0.020921	-.000059	0.024845	-.003812	0.506112	-.003574	-.016635
D14	-.002010	0.006200	0.006972	0.003310	-.027021	-.003016	-.023360	0.008070	-.527669	0.000718	-.019017
D15	-.170833	0.140943	-.013486	-.002085	0.005157	0.002258	-.001279	-.005049	0.006282	-.020105	-.536479
D16	0.169903	-.135388	-.003725	0.013324	0.000552	-.008983	-.001331	-.001338	0.004535	0.018777	0.472745
D17	-.007362	0.014160	-.006940	0.001123	-.007637	-.001551	-.021764	0.001749	-.467946	0.003694	0.015857
D18	-.016803	0.005155	-.003655	-.005499	0.011521	0.003956	0.022442	-.005635	0.490211	-.001261	0.018157
D19	-.143778	0.130937	0.006763	-.000726	-.001561	0.003173	0.005989	0.006363	-.005845	0.003118	0.530472
D20	0.169211	-.118705	0.016799	0.008958	-.002272	0.003267	-.002915	-.001920	-.002762	-.003021	-.451472
D21	-.008662	0.003939	0.009662	-.002985	-.040650	0.034080	0.540214	-.005765	-.023813	-.003925	-.010813
D22	0.004313	-.009223	-.013359	-.000986	0.032833	-.026440	-.430079	0.004266	0.019116	0.003076	0.008290
D23	0.001582	0.001089	-.000377	0.002239	0.045761	-.035261	-.555618	0.005743	0.023596	-.001348	-.005277
D24	0.000244	-.000800	-.003067	-.005087	-.036569	0.026147	0.449625	-.000484	-.020319	0.001996	0.005708
D25	0.102088	-.022449	-.002846	0.061086	0.000995	-.000918	0.000183	0.012359	0.000191	0.001733	-.000336
D26	-.332204	0.083497	0.030126	-.301318	-.009486	0.002500	-.005274	-.048205	-.002881	-.001427	0.001745
D27	0.246431	-.038746	-.053036	0.583956	0.021040	0.003518	0.008392	0.027840	0.008195	0.000839	-.002392
D28	0.031910	-.056680	-.013239	-.415695	-.013599	-.003966	-.003779	0.032918	-.004403	-.002236	0.001333
D29	-.074457	-.002277	0.015298	0.142190	0.003287	0.000297	-.000734	0.050866	-.000534	-.002303	-.001680
D30	0.313152	0.013285	0.112922	-.439547	-.011337	0.000151	0.000378	-.052760	-.000479	0.000125	0.001310
D31	-.182020	0.122371	0.034047	0.403672	0.009444	0.001584	0.000304	-.052178	0.000470	0.004672	0.000810
D32	-.027821	-.037627	-.027280	-.001957	-.000303	0.000310	0.000404	-.002634	0.000523	0.001381	0.000013
D33	-.209652	-.004624	-.031540	0.044921	0.002180	-.000630	0.002131	0.012643	-.003210	0.001441	0.001786
D34	0.252734	0.095884	-.114296	-.077748	-.001000	0.001847	0.001844	0.077673	0.002227	-.002793	-.001031
D35	-.079839	0.042258	0.178238	-.002006	-.000210	0.005633	-.001558	-.012982	0.001187	-.000933	-.000338
D36	-.000018	-.127279	-.245626	-.020432	0.001088	-.006009	0.001034	-.010529	-.001249	0.001483	-.000437
D37	-.051408	-.058036	0.036987	0.005349	0.000351	-.000421	-.001807	-.019676	0.000403	0.000221	0.000454
D38	0.211394	0.201042	0.209784	0.017616	0.000657	0.006613	-.004163	0.082485	0.004542	0.001141	-.001061
D39	-.085061	-.267457	-.555906	-.010405	0.001551	-.010836	0.005422	-.332291	-.004547	-.000177	0.002112
D40	0.001918	-.034541	0.520804	0.043404	-.001432	0.001359	-.000829	0.298855	0.002697	-.000749	-.001031
D41	0.239039	0.320472	-.195300	0.001124	0.000922	0.004008	0.005751	0.137602	0.002235	-.000932	-.000025
D42	-.189185	-.444603	0.395453	0.037259	-.001653	-.006427	-.007208	-.409149	-.003785	0.000238	0.001735
D43	-.193551	-.189392	-.229996	-.039983	0.001221	0.002068	0.005898	0.651220	0.005207	0.004812	-.004470
D44	0.157662	0.477869	-.001853	-.002904	-.000459	0.004459	-.006910	-.404219	-.002823	-.002911	0.002841

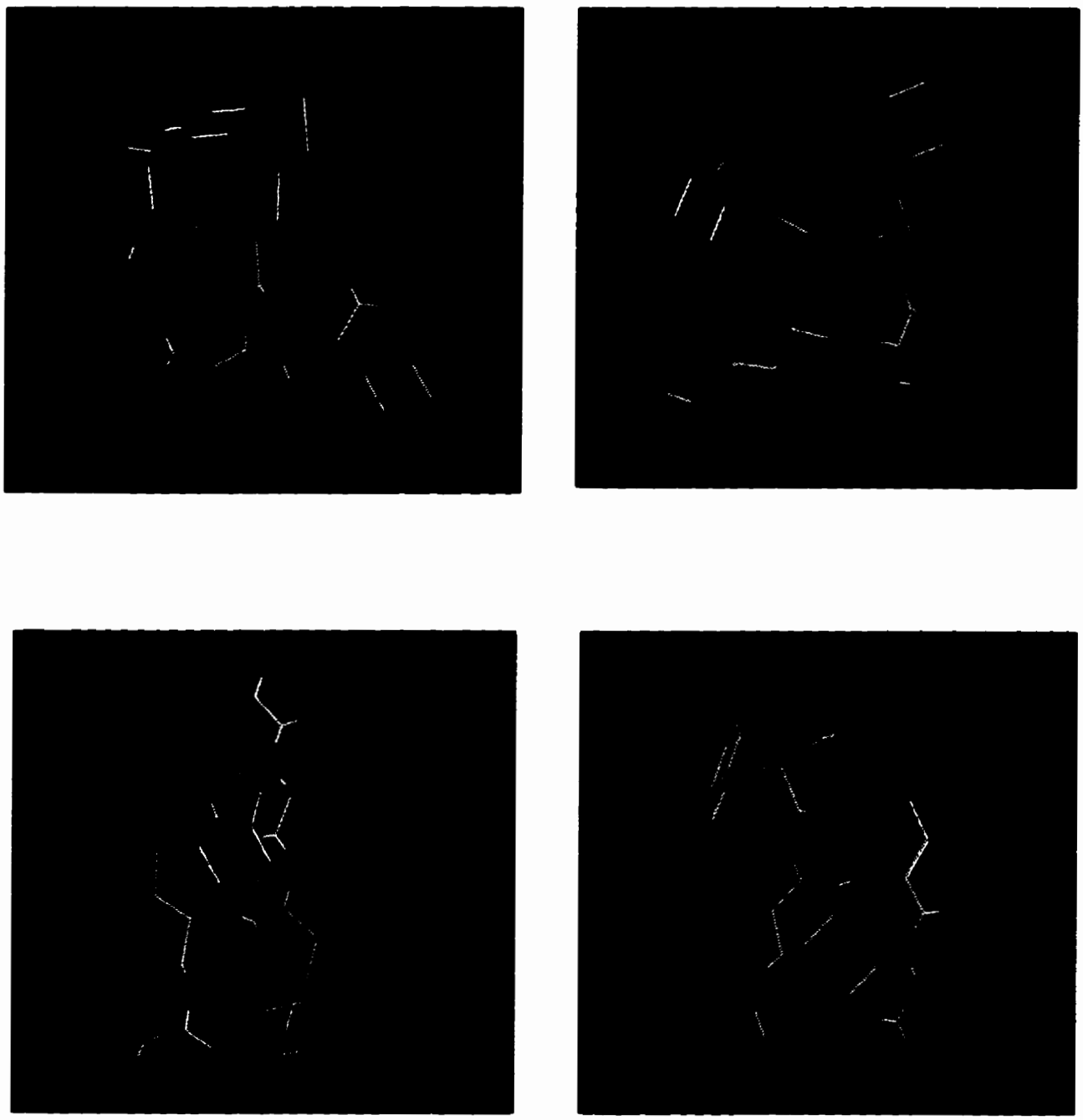


Figure 60. Conformation type pour chaque famille de glmap: de gauche à droite et de haut en bas, famille 1, 2, 3, 4.

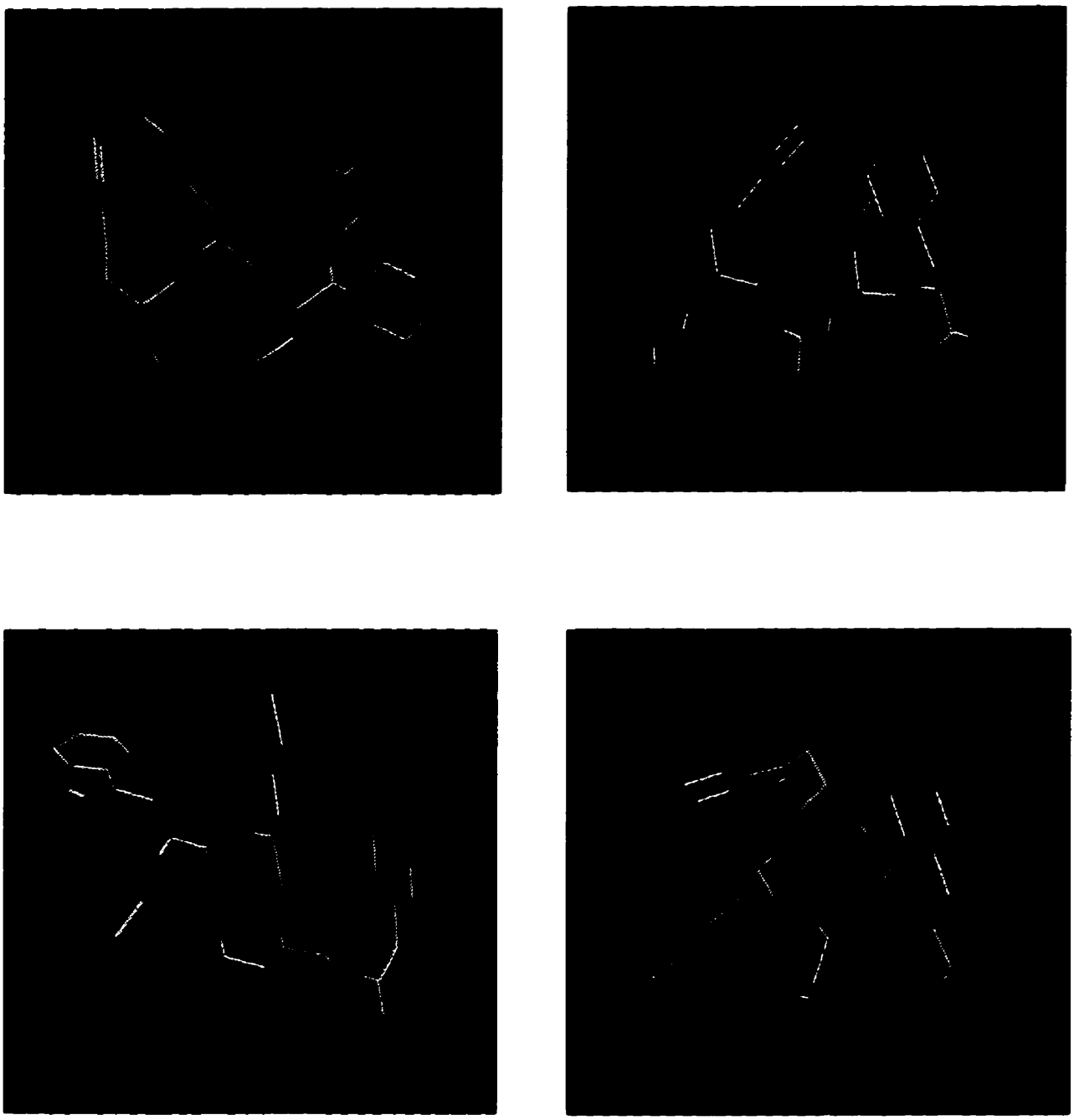


Figure 61. Conformation type pour chaque famille de glmap: de gauche à droite et de haut en bas, famille 5, 6, 7, 8.

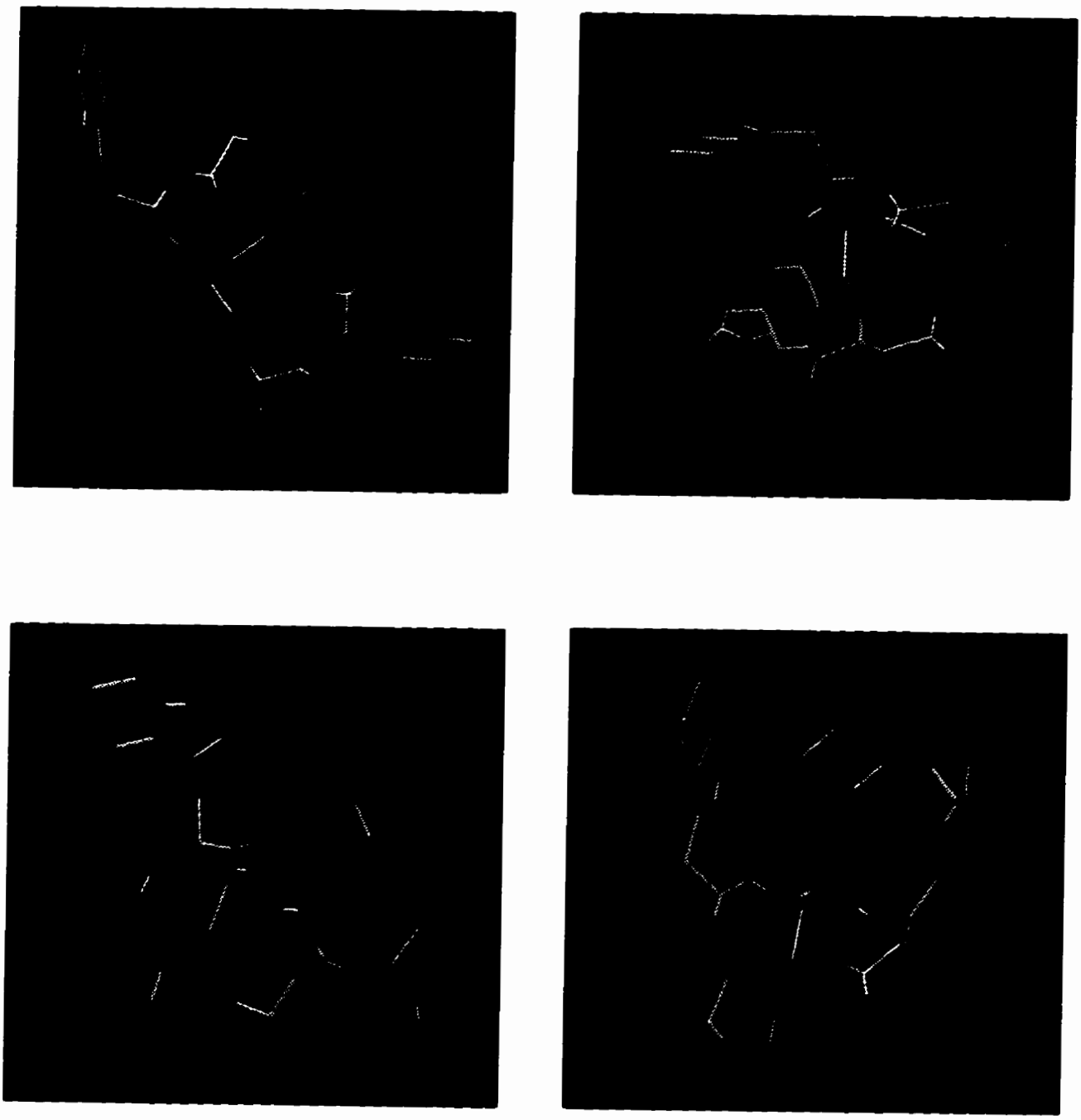


Figure 62. Conformation type pour chaque famille de gimap: de gauche à droite et de haut en bas, famille 9, 10, 11, 12.

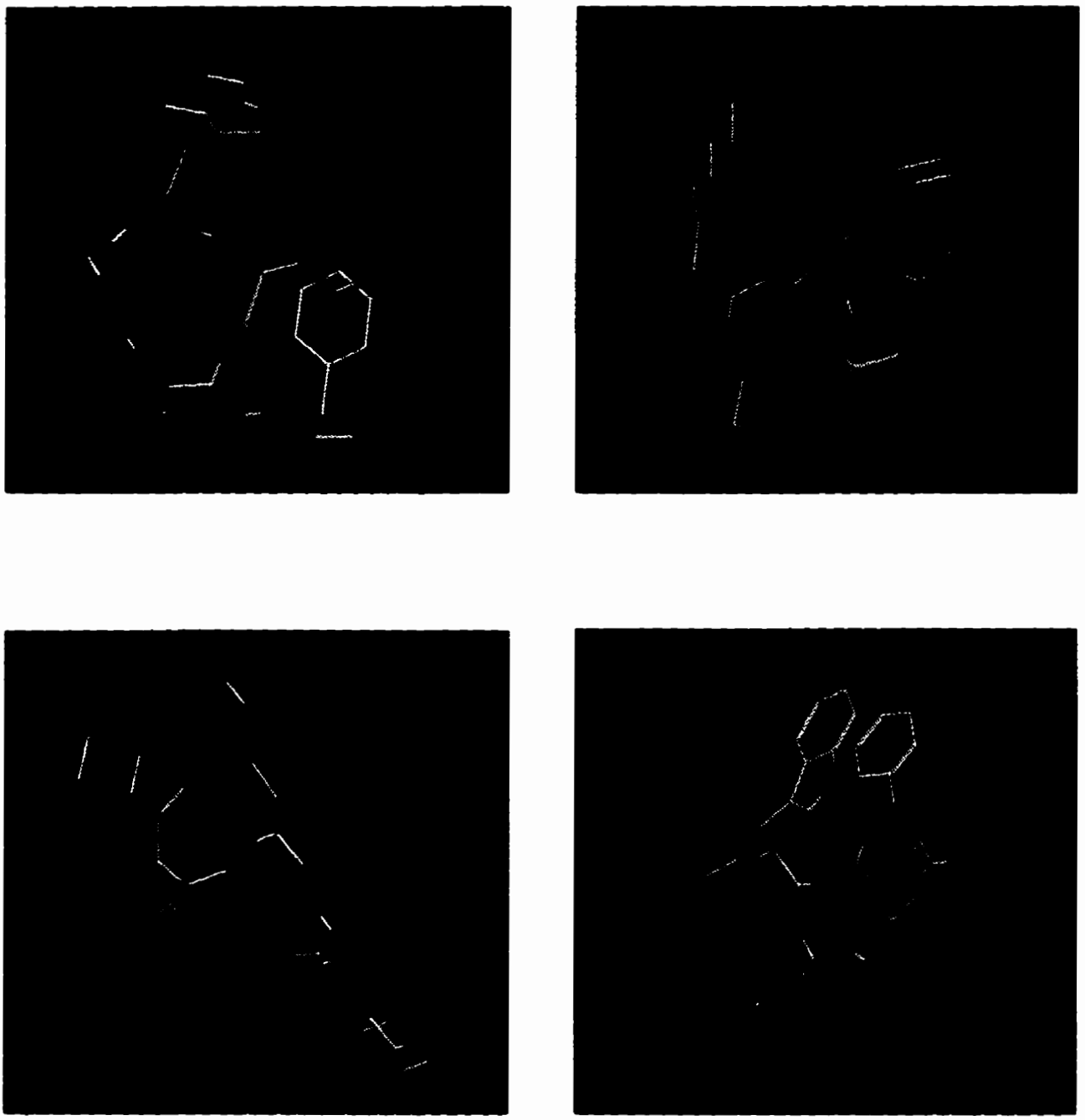


Figure 63. Conformation type pour chaque famille de glmap: de gauche à droite et de haut en bas, famille 13, 14, 15, 16.

Tableau 41. Résultats de l'analyse en composantes principales pour gdmap: matrice des corrélations entre distances initiales.

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15
D1	1.0000	0.6993	0.4581	0.2819	-.0103	-.0359	-.0280	-.0881	0.0051	0.0100	0.0859	0.3433	0.0007	0.0311	0.0715
D2	0.6993	1.0000	0.7866	0.5295	0.2470	0.1414	0.0936	0.0366	-.0172	0.0019	-.1336	0.2504	0.0076	0.1804	0.4044
D3	0.4581	0.7866	1.0000	0.8470	0.2161	0.4540	0.3962	0.0588	0.1086	0.1009	-.0675	0.2177	0.0348	0.1362	0.4686
D4	0.2819	0.5295	0.8470	1.0000	0.1544	0.4601	0.6540	0.0481	0.1269	0.3022	-.0513	0.1410	0.1417	0.1729	0.4586
D5	-.0103	0.2470	0.2161	0.1544	1.0000	0.6775	0.4293	0.0259	-.0149	0.0101	-.0618	-.0649	0.0159	0.0885	0.3068
D6	-.0359	0.1414	0.4540	0.4601	0.6775	1.0000	0.7967	0.0469	0.2550	0.2140	0.0025	-.0481	0.0520	-.0712	0.2610
D7	-.0280	0.0936	0.3962	0.6540	0.4293	0.7967	1.0000	0.0381	0.2515	0.4977	-.0004	-.0356	0.2091	-.0218	0.2392
D8	-.0881	0.0366	0.0588	0.0481	0.0259	0.0469	0.0381	1.0000	0.0106	0.0075	-.2056	0.0416	-.0136	-.0144	-.0412
D9	0.0051	-.0172	0.1086	0.1269	-.0149	0.2350	0.2515	0.0106	1.0000	0.7205	0.0372	0.0153	0.0655	-.1110	-.1223
D10	0.0100	0.0019	0.1009	0.3022	0.0101	0.2140	0.4977	0.0075	0.7205	1.0000	0.0233	0.0112	0.3389	-.0588	-.0847
D11	0.0859	-.1336	-.0675	-.0513	-.0618	0.0025	-.0004	-.2056	0.0372	0.0233	1.0000	0.7219	0.0251	-.2646	-.0995
D12	0.3433	0.2504	0.2177	0.1410	-.0649	-.0481	-.0356	0.0416	0.0153	0.0112	0.7219	1.0000	0.0061	-.0975	0.1097
D13	0.0007	0.0076	0.0348	0.1417	0.0159	0.0520	0.2091	-.0136	0.0655	0.3389	0.0251	0.0061	1.0000	-.0353	-.0156
D14	0.0311	0.1804	0.1362	0.1729	0.0885	-.0712	-.0218	-.0144	-.1110	-.0588	-.2646	-.0975	-.0353	1.0000	0.7783
D15	0.0715	0.4044	0.4686	0.4586	0.3068	0.2610	0.2392	-.0412	-.1223	-.0847	-.0995	0.1097	-.0156	0.7783	1.0000
D16	0.2217	0.5619	0.5859	0.5070	0.1528	0.1411	0.1400	0.0820	-.0753	-.0535	0.0676	0.4597	-.0079	0.5131	0.8487
D17	-.0128	-.0151	0.0945	0.0991	0.0002	0.1560	0.0918	0.0149	0.1076	-.1057	-.0165	-.0076	-.0939	-.2732	-.1384
D18	-.0111	0.0472	0.3581	0.4866	0.0325	0.3831	0.4390	0.0163	0.3157	0.2107	-.1197	-.0575	-.0861	-.0478	0.1115
D19	0.0076	0.2013	0.5934	0.7142	0.1715	0.5406	0.5753	0.0062	0.1333	0.0927	-.0596	0.0483	-.0465	0.1869	0.5026
D20	0.1234	0.3768	0.6900	0.7377	0.1066	0.3728	0.4080	0.0834	0.0710	0.0520	0.0529	0.3196	-.0239	0.2182	0.5917
D21	-.1570	-.0298	0.0055	0.0043	-.0505	-.0362	-.0214	0.1327	0.0142	0.0181	-.0710	-.0784	-.0029	-.0403	-.0326
D22	-.0654	-.0892	-.2142	-.2193	0.1642	0.0054	-.0407	-.0375	-.0301	-.0240	-.1710	-.1453	-.0016	-.0867	-.1581
D23	-.0665	-.0865	0.1075	0.0604	0.0560	0.2740	0.1313	-.0202	0.0364	-.0406	-.1149	-.1147	-.0749	0.1000	0.1205
D24	-.1419	-.0245	0.0102	0.0069	-.0574	-.0358	-.0205	0.1273	0.0208	0.0228	-.0707	-.0747	-.0068	-.0460	-.0369
D25	-.0589	-.0782	-.2222	-.2292	0.1803	-.0088	-.0532	-.0375	-.0596	-.0403	-.1566	-.1364	-.0068	-.0843	-.1606
D26	-.0762	-.0957	0.1289	0.1121	0.0825	0.3341	0.2075	-.0220	0.0559	-.0133	-.1409	-.1347	-.0527	0.1265	0.1468
D27	-.1088	-.0038	0.0187	0.0089	-.0515	-.0409	-.0296	0.1357	0.0233	0.0184	-.0572	-.0503	-.0200	-.0414	-.0237
D28	-.1202	-.0038	0.0178	0.0086	-.0446	-.0412	-.0311	0.1456	0.0180	0.0150	-.0588	-.0525	-.0164	-.0352	-.0170
D29	-.0288	-.0355	-.1780	-.1927	0.1682	-.0109	-.0503	-.0273	-.0628	-.0401	-.1404	-.1052	-.0046	-.0689	-.1287
D30	-.0313	-.0421	-.1680	-.1811	0.1565	0.0036	-.0378	-.0293	-.0339	-.0233	-.1545	-.1141	-.0010	-.0694	-.1243
D31	-.0431	-.0566	0.1215	0.0719	0.0621	0.2582	0.1258	-.0162	0.0311	-.0377	-.1167	-.1053	-.0676	0.1049	0.1254
D32	-.0486	-.0599	0.1428	0.1207	0.0850	0.3069	0.1913	-.0181	0.0443	-.0151	-.1395	-.1210	-.0473	0.1291	0.1509
D33	0.0006	-.0425	0.1293	0.1939	-.1735	0.0120	0.0644	0.0074	-.0424	-.0389	0.0016	0.0078	-.0106	0.1972	0.1909
D34	-.0156	-.0737	0.1122	0.1838	-.1904	0.0251	0.0787	0.0048	0.0131	0.0028	0.0024	0.0007	-.0057	0.1645	0.1622
D35	-.0036	-.0797	-.1871	-.2853	-.1163	-.1695	-.2752	-.0011	0.1172	-.0370	0.0754	0.0156	-.0841	-.3278	-.3457
D36	0.0000	-.0955	-.2099	-.2995	-.1486	-.1897	-.2761	0.0019	0.1761	0.0168	0.0979	0.0325	-.0659	-.3961	-.4190
D37	0.0022	-.0389	0.1360	0.1991	-.1679	0.0184	0.0685	0.0118	-.0491	-.0453	-.0002	0.0080	-.0097	0.1993	0.1938
D38	-.0174	-.0753	0.1176	0.1888	-.1862	0.0358	0.0869	0.0081	0.0146	0.0030	0.0049	0.0029	-.0030	0.1595	0.1617
D39	-.0076	-.0788	-.1738	-.2798	-.1113	-.1498	-.2694	0.0003	0.1219	-.0435	0.0632	0.0059	-.0926	-.3214	-.3327
D40	-.0048	-.0958	-.1965	-.2960	-.1450	-.1668	-.2703	0.0046	0.1936	0.0166	0.0869	0.0230	-.0759	-.4021	-.4160
D41	-.0120	-.0242	-.0417	0.0641	-.0013	-.0315	0.1025	-.0080	-.1160	0.0173	-.0058	-.0243	-.0333	-.0251	-.0222
D42	-.0093	-.0211	-.0370	0.0784	-.0002	-.0297	0.1175	-.0073	-.1317	0.0139	-.0029	-.0200	-.0323	-.0182	-.0190
D43	-.0045	-.0161	-.0397	0.0605	-.0017	-.0473	0.0845	-.0018	-.1458	-.0008	-.0068	-.0167	-.0245	-.0005	-.0039
D44	-.0009	-.0103	-.0353	0.0734	0.0011	-.0514	0.0943	-.0023	-.1728	-.0101	-.0032	-.0105	-.0190	0.0172	0.0077

Tableau 42. Résultats de l'analyse en composantes principales pour gdmap: matrice des corrélations entre distances initiales (suite).

	D16	D17	D18	D19	D20	D21	D22	D23	D24	D25	D26	D27	D28	D29	D30
D1	0.2217	-.0128	-.0111	0.0076	0.1234	-.1570	-.0654	-.0665	-.1419	-.0589	-.0762	-.1088	-.1202	-.0288	-.0313
D2	0.5619	-.0151	0.0472	0.2013	0.3768	-.0298	-.0892	-.0865	-.0245	-.0782	-.0957	-.0038	-.0038	-.0355	-.0421
D3	0.5859	0.0945	0.3581	0.5934	0.6900	0.0055	-.2142	0.1075	0.0102	-.2222	0.1289	0.0187	0.0178	-.1780	-.1680
D4	0.5070	0.0991	0.4666	0.7142	0.7377	0.0043	-.2193	0.0604	0.0069	-.2292	0.1121	0.0089	0.0086	-.1927	-.1811
D5	0.1528	0.0002	0.0325	0.1715	0.1066	-.0505	0.1642	0.0560	-.0574	0.1803	0.0825	-.0515	-.0446	0.1682	0.1565
D6	0.1411	0.1560	0.3831	0.5406	0.3728	-.0362	0.0054	0.2740	-.0358	-.0088	0.3341	-.0409	-.0412	-.0109	0.0036
D7	0.1400	0.0918	0.4390	0.5753	0.4080	-.0214	-.0407	0.1313	-.0205	-.0532	0.2075	-.0296	-.0311	-.0503	-.0378
D8	0.0820	0.0149	0.0163	0.0062	0.0834	0.1327	-.0375	-.0202	0.1273	-.0375	-.0220	0.1357	0.1456	-.0273	-.0293
D9	-.0753	0.1076	0.3157	0.1333	0.0710	0.0142	-.0301	0.0364	0.0208	-.0596	0.0559	0.0233	0.0180	-.0628	-.0339
D10	-.0535	-.1057	0.2107	0.0927	0.0520	0.0181	-.0240	-.0406	0.0228	-.0403	-.0133	0.0184	0.0150	-.0401	-.0233
D11	0.0676	-.0165	-.1197	-.0596	0.0529	-.0710	-.1710	-.1149	-.0707	-.1566	-.1409	-.0572	-.0588	-.1404	-.1545
D12	0.4597	-.0076	-.0575	0.0483	0.3196	-.0784	-.1453	-.1147	-.0747	-.1364	-.1347	-.0503	-.0525	-.1052	-.1141
D13	-.0079	-.0939	-.0861	-.0465	-.0239	-.0029	-.0016	-.0749	-.0068	-.0068	-.0527	-.0200	-.0164	-.0046	-.0010
D14	0.5131	-.2732	-.0478	0.1869	0.2182	-.0403	-.0867	0.1000	-.0460	-.0843	0.1265	-.0414	-.0352	-.0689	-.0694
D15	0.8487	-.1384	0.1115	0.5026	0.5917	-.0326	-.1581	0.1205	-.0369	-.1606	0.1468	-.0237	-.0170	-.1287	-.1243
D16	1.0000	-.0729	0.1009	0.4601	0.7230	0.0020	-.2119	0.0088	-.0005	-.2141	0.0116	0.0215	0.0279	-.1663	-.1627
D17	-.0729	1.0000	0.7903	0.4880	0.3148	-.0120	0.0051	-.0779	-.0082	-.0071	-.0363	-.0038	-.0056	-.0169	-.0028
D18	0.1009	0.7903	1.0000	0.8273	0.6002	-.0176	-.0958	0.0812	-.0109	-.1182	0.1570	-.0119	-.0174	-.1202	-.0959
D19	0.4601	0.4880	0.8273	1.0000	0.8812	-.0231	-.1891	0.1473	-.0212	-.2092	0.2204	-.0152	-.0156	-.1923	-.1708
D20	0.7230	0.3148	0.6002	0.8812	1.0000	0.0009	-.2317	0.0506	0.0001	-.2465	0.0912	0.0174	0.0211	-.2125	-.1964
D21	0.0020	-.0120	-.0176	-.0231	0.0009	1.0000	0.0742	0.0319	0.9257	0.0762	0.0359	0.6465	0.6905	0.0170	0.0139
D22	-.2119	0.0051	-.0958	-.1891	-.2317	0.0742	1.0000	0.1767	0.0261	0.8844	0.2448	-.0352	-.0081	0.6510	0.7536
D23	0.0088	-.0779	0.0812	0.1473	0.0506	0.0319	0.1767	1.0000	0.0071	0.1663	0.9373	-.0319	-.0197	0.0869	0.0955
D24	-.0005	-.0082	-.0109	-.0212	0.0001	0.9257	0.0261	0.0071	1.0000	0.0275	0.0035	0.7136	0.6404	-.0158	-.0179
D25	-.2141	-.0071	-.1182	-.2092	-.2465	0.0762	0.8844	0.1663	0.0275	1.0000	0.2311	-.0362	-.0072	0.7511	0.6519
D26	0.0116	-.0363	0.1570	0.2204	0.0912	0.0359	0.2448	0.9373	0.0035	0.2311	1.0000	-.0449	-.0281	0.1257	0.1373
D27	0.0215	-.0038	-.0119	-.0152	0.0174	0.6465	-.0352	-.0319	0.7136	-.0362	-.0449	1.0000	0.9345	0.0259	0.0264
D28	0.0279	-.0056	-.0174	-.0156	0.0211	0.6905	-.0081	-.0197	0.6404	-.0072	-.0281	0.9345	1.0000	0.0722	0.0705
D29	-.1663	-.0169	-.1202	-.1923	-.2125	0.0170	0.6510	0.0869	-.0158	0.7511	0.1257	0.0259	0.0722	1.0000	0.8905
D30	-.1627	-.0028	-.0959	-.1708	-.1964	0.0139	0.7536	0.0955	-.0179	0.6519	0.1373	0.0264	0.0705	0.8905	1.0000
D31	0.0228	-.0783	0.0749	0.1420	0.0546	0.0036	0.0890	0.8350	-.0135	0.0755	0.7684	-.0120	0.0103	0.1777	0.1906
D32	0.0292	-.0418	0.1400	0.2065	0.0930	-.0027	0.1262	0.7566	-.0241	0.1097	0.7953	-.0160	0.0133	0.2415	0.2577
D33	0.1413	-.0057	0.1399	0.2279	0.2041	-.0540	-.0880	0.1278	-.1020	-.0417	0.1466	-.0926	-.0513	-.0355	-.0782
D34	0.1170	0.0110	0.1582	0.2263	0.1929	-.0485	-.0099	0.1340	-.0960	-.0962	0.1554	-.0883	-.0487	-.0804	-.0037
D35	-.2371	-.1066	-.2333	-.3207	-.2656	0.0073	0.0100	0.1288	-.0147	0.0102	0.0065	-.0110	0.0098	0.0169	0.0155
D36	-.2816	-.0641	-.2052	-.3342	-.2811	-.0041	0.0126	0.0187	-.0349	0.0080	-.0439	-.0266	0.0021	0.0106	0.0134
D37	0.1414	-.0047	0.1420	0.2326	0.2060	-.0459	-.0760	0.1312	-.0585	-.0128	0.1507	-.0553	-.0452	-.0115	-.0682
D38	0.1156	0.0143	0.1629	0.2324	0.1962	-.0391	0.0177	0.1389	-.0538	-.0840	0.1608	-.0512	-.0405	-.0714	0.0207
D39	-.2310	-.1112	-.2292	-.3108	-.2598	0.0157	0.0124	0.1904	0.0135	0.0101	0.0552	0.0122	0.0159	0.0150	0.0160
D40	-.2819	-.0666	-.1992	-.3264	-.2780	0.0068	0.0140	0.0781	0.0029	0.0069	0.0085	0.0048	0.0102	0.0080	0.0133
D41	-.0302	0.1882	0.1708	0.1043	0.0581	-.0000	-.0279	0.0955	0.0009	-.0265	0.0344	0.0000	-.0037	-.0305	-.0315
D42	-.0283	0.2082	0.1903	0.1197	0.0684	-.0053	-.0486	0.0424	-.0065	-.0478	0.0192	-.0042	-.0047	-.0447	-.0460
D43	-.0123	0.1622	0.1464	0.0956	0.0613	-.0004	-.0475	0.0975	0.0001	-.0268	0.0332	0.0003	-.0020	-.0261	-.0435
D44	-.0030	0.1685	0.1540	0.1069	0.0720	-.0059	-.0795	0.0385	-.0060	-.0524	0.0124	-.0040	-.0041	-.0410	-.0644

Tableau 43. Résultats de l'analyse en composantes principales pour gdmap: matrice des corrélations entre distances initiales (suite).

	D31	D32	D33	D34	D35	D36	D37	D38	D39	D40	D41	D42	D43	D44
D1	-.0431	-.0486	0.0006	-.0156	-.0036	0.0000	0.0022	-.0174	-.0076	-.0048	-.0120	-.0093	-.0045	-.0009
D2	-.0566	-.0599	-.0425	-.0737	-.0797	-.0955	-.0389	-.0753	-.0788	-.0958	-.0242	-.0211	-.0161	-.0103
D3	0.1215	0.1428	0.1293	0.1122	-.1871	-.2099	0.1360	0.1176	-.1738	-.1965	-.0417	-.0370	-.0397	-.0353
D4	0.0719	0.1207	0.1939	0.1838	-.2853	-.2995	0.1991	0.1888	-.2798	-.2960	0.0641	0.0784	0.0605	0.0734
D5	0.0621	0.0850	-.1735	-.1904	-.1163	-.1486	-.1679	-.1862	-.1113	-.1450	-.0013	-.0002	-.0017	0.0011
D6	0.2582	0.3069	0.0120	0.0251	-.1695	-.1897	0.0184	0.0358	-.1498	-.1668	-.0315	-.0297	-.0473	-.0514
D7	0.1258	0.1913	0.0644	0.0787	-.2752	-.2761	0.0685	0.0869	-.2694	-.2703	0.1025	0.1175	0.0845	0.0943
D8	-.0162	-.0181	0.0074	0.0048	-.0011	0.0019	0.0118	0.0081	0.0003	0.0046	-.0080	-.0073	-.0018	-.0023
D9	0.0311	0.0443	-.0424	0.0131	0.1172	0.1761	-.0491	0.0146	0.1219	0.1936	-.1160	-.1317	-.1458	-.1728
D10	-.0377	-.0151	-.0389	0.0028	-.0370	0.0168	-.0453	0.0030	-.0435	0.0166	0.0173	0.0139	-.0008	-.0101
D11	-.1167	-.1395	0.0016	0.0024	0.0754	0.0979	-.0002	0.0049	0.0632	0.0869	-.0058	-.0029	-.0068	-.0032
D12	-.1053	-.1210	0.0078	0.0007	0.0156	0.0325	0.0080	0.0029	0.0059	0.0230	-.0243	-.0200	-.0167	-.0105
D13	-.0678	-.0473	-.0106	-.0057	-.0841	-.0659	-.0097	-.0030	-.0926	-.0759	-.0333	-.0323	-.0245	-.0190
D14	0.1049	0.1291	0.1972	0.1645	-.3278	-.3961	0.1993	0.1595	-.3214	-.4021	-.0251	-.0182	-.0005	0.0172
D15	0.1254	0.1509	0.1909	0.1622	-.3457	-.4190	0.1938	0.1617	-.3327	-.4160	-.0222	-.0190	-.0039	0.0077
D16	0.0228	0.0292	0.1413	0.1170	-.2371	-.2816	0.1414	0.1156	-.2310	-.2819	-.0302	-.0283	-.0123	-.0030
D17	-.0783	-.0418	-.0057	0.0110	-.1066	-.0641	-.0047	0.0143	-.1112	-.0666	0.1882	0.2082	0.1622	0.1685
D18	0.0749	0.1400	0.1399	0.1582	-.2333	-.2052	0.1420	0.1629	-.2292	-.1992	0.1708	0.1903	0.1464	0.1540
D19	0.1420	0.2065	0.2279	0.2263	-.3207	-.3342	0.2326	0.2324	-.3108	-.3264	0.1043	0.1197	0.0956	0.1069
D20	0.0546	0.0930	0.2041	0.1929	-.2656	-.2811	0.2060	0.1962	-.2598	-.2780	0.0581	0.0684	0.0613	0.0720
D21	0.0036	-.0027	-.0540	-.0485	0.0073	-.0041	-.0459	-.0391	0.0157	0.0068	-.0000	-.0053	-.0004	-.0059
D22	0.0890	0.1262	-.0880	-.0099	0.0100	0.0126	-.0760	0.0177	0.0124	0.0140	-.0279	-.0486	-.0475	-.0795
D23	0.8350	0.7566	0.1278	0.1340	0.1288	0.0187	0.1312	0.1389	0.1904	0.0781	0.0955	0.0424	0.0975	0.0385
D24	-.0135	-.0241	-.1020	-.0960	-.0147	-.0349	-.0585	-.0538	0.0135	0.0029	0.0009	-.0065	0.0001	-.0060
D25	0.0755	0.1097	-.0417	-.0962	0.0102	0.0080	-.0128	-.0840	0.0101	0.0069	-.0265	-.0478	-.0268	-.0524
D26	0.7684	0.7953	0.1466	0.1534	0.0065	-.0439	0.1507	0.1608	0.0552	0.0085	0.0344	0.0192	0.0332	0.0124
D27	-.0120	-.0160	-.0926	-.0883	-.0110	-.0266	-.0553	-.0512	0.0122	0.0048	0.0000	-.0042	0.0003	-.0040
D28	0.0103	0.0133	-.0513	-.0487	0.0098	0.0021	-.0452	-.0405	0.0159	0.0102	-.0037	-.0047	-.0020	-.0041
D29	0.1777	0.2415	-.0355	-.0804	0.0169	0.0106	-.0115	-.0714	0.0150	0.0080	-.0305	-.0447	-.0261	-.0410
D30	0.1906	0.2577	-.0782	-.0037	0.0155	0.0134	-.0682	0.0207	0.0160	0.0133	-.0315	-.0460	-.0435	-.0684
D31	1.0000	0.9439	0.1248	0.1333	0.1132	0.0129	0.1246	0.1356	0.1671	0.0646	0.0753	0.0304	0.0791	0.0303
D32	0.9439	1.0000	0.1417	0.1521	0.0067	-.0414	0.1408	0.1538	0.0486	0.0036	0.0198	0.0089	0.0221	0.0080
D33	0.1248	0.1417	1.0000	0.7738	-.0449	-.0260	0.9041	0.6592	-.0893	-.0902	-.0837	-.1148	-.0695	-.0955
D34	0.1333	0.1521	0.7738	1.0000	-.0327	-.0073	0.6629	0.9093	-.0756	-.0691	-.0718	-.0969	-.1099	-.1450
D35	0.1132	0.0067	-.0449	-.0327	1.0000	0.9229	-.0918	-.0779	0.9566	0.8883	0.1157	0.0388	0.1117	0.0318
D36	0.0129	-.0414	-.0260	-.0073	0.9229	1.0000	-.0875	-.0659	0.8516	0.9382	0.0427	0.0085	0.0355	-.0013
D37	0.1246	0.1408	0.9041	0.6629	-.0918	-.0875	1.0000	0.7201	-.0842	-.0786	-.0993	-.1363	-.0839	-.1150
D38	0.1356	0.1538	0.6592	0.9093	-.0779	-.0659	0.7201	1.0000	-.0690	-.0550	-.0852	-.1158	-.1282	-.1692
D39	0.1671	0.0486	-.0893	-.0756	0.9566	0.8516	-.0842	-.0690	1.0000	0.9166	0.1447	0.0573	0.1398	0.0486
D40	0.0646	0.0036	-.0902	-.0691	0.8883	0.9382	-.0786	-.0550	0.9166	1.0000	0.0664	0.0266	0.0577	0.0142
D41	0.0753	0.0198	-.0837	-.0718	0.1157	0.0427	-.0993	-.0852	0.1447	0.0664	1.0000	0.9105	0.8506	0.7312
D42	0.0304	0.0089	-.1148	-.0969	0.0388	0.0085	-.1363	-.1158	0.0573	0.0266	0.9105	1.0000	0.7450	0.7742
D43	0.0791	0.0221	-.0695	-.1099	0.1117	0.0355	-.0839	-.1282	0.1398	0.0577	0.8506	0.7450	1.0000	0.9125
D44	0.0303	0.0080	-.0955	-.1450	0.0318	-.0013	-.1150	-.1692	0.0486	0.0142	0.7312	0.7742	0.9125	1.0000

Tableau 44. Résultats de l'analyse en composantes principales pour gdmap: valeurs propres associées à chaque composante principale.

Valeurs propres

	Valeurs propres	Pourcentages de variance	Pourcentages cumulez
cp1	7.26402	0.165091	0.16509
cp2	4.48200	0.101864	0.26696
cp3	4.02236	0.091417	0.35837
cp4	3.73699	0.084932	0.44330
cp5	3.31703	0.075387	0.51869
cp6	3.08253	0.070057	0.58875
cp7	2.76155	0.062763	0.65151
cp8	2.29085	0.052065	0.70358
cp9	1.77016	0.040231	0.74381
cp10	1.64744	0.037442	0.78125
cp11	1.38261	0.031423	0.81267
cp12	1.19081	0.027064	0.83974
cp13	1.00221	0.022777	0.86251
cp14	0.86524	0.019664	0.88218
cp15	0.74518	0.016936	0.89911
cp16	0.61507	0.013979	0.91309
cp17	0.57035	0.012962	0.92605
cp18	0.47715	0.010844	0.93690
cp19	0.34690	0.007884	0.94478
cp20	0.33611	0.007639	0.95242
cp21	0.24212	0.005503	0.95792
cp22	0.22021	0.005005	0.96293
cp23	0.20718	0.004709	0.96764
cp24	0.20108	0.004570	0.97221
cp25	0.19071	0.004334	0.97654
cp26	0.16689	0.003793	0.98034
cp27	0.15335	0.003485	0.98382
cp28	0.11893	0.002703	0.98652
cp29	0.11072	0.002516	0.98904
cp30	0.10708	0.002434	0.99147
cp31	0.07690	0.001748	0.99322
cp32	0.06227	0.001415	0.99464
cp33	0.05641	0.001282	0.99592
cp34	0.05430	0.001234	0.99715
cp35	0.04310	0.000980	0.99813
cp36	0.01949	0.000443	0.99858
cp37	0.01316	0.000299	0.99887
cp38	0.01145	0.000260	0.99913
cp39	0.01035	0.000235	0.99937
cp40	0.00994	0.000226	0.99960
cp41	0.00624	0.000142	0.99974
cp42	0.00549	0.000125	0.99986
cp43	0.00462	0.000105	0.99997
cp44	0.00145	0.000033	1.00000

Tableau 45. Résultats de l'analyse en composantes principales pour gdmap: vecteurs propres associés à chaque composante principale.

	cp1	cp2	cp3	cp4	cp5	cp6	cp7	cp8	cp9	cp10	cp11
D1	0.073152	-.098506	0.002123	0.071279	-.072599	0.174591	0.223186	0.223668	0.021300	0.040256	-.422032
D2	0.164890	-.107952	0.016190	-.017548	-.003912	0.218995	0.305958	0.221456	0.015688	-.166955	-.224033
D3	0.278325	-.036944	0.047379	0.044385	0.076969	0.202938	0.146662	0.122705	-.026495	-.087192	-.144815
D4	0.304167	-.029885	0.080548	0.025446	0.061599	0.123288	-.004206	0.095005	0.110083	-.058708	-.091193
D5	0.094235	0.076077	0.047023	-.174424	-.093955	0.224997	0.047674	-.028614	0.045561	-.110113	0.477945
D6	0.200412	0.140756	0.091471	-.052330	0.011654	0.253645	-.141702	-.096926	0.019115	-.000843	0.328771
D7	0.222705	0.079861	0.120499	-.064821	0.002967	0.169886	-.224076	-.054058	0.234311	0.022436	0.185714
D8	0.015350	-.010004	0.013558	-.036541	0.119683	-.009527	0.002952	0.017639	-.011581	-.163022	0.071887
D9	0.026997	0.028814	0.047724	0.074298	0.127665	0.262012	-.263082	-.030698	0.184078	-.016719	-.189256
D10	0.054566	-.001737	0.066471	-.000894	0.072582	0.207245	-.262705	-.031892	0.473787	0.020396	-.197820
D11	-.021677	-.132329	0.012416	0.130385	-.013111	0.070275	-.009884	0.050297	0.018964	0.645468	0.204854
D12	0.058039	-.157779	0.000675	0.115453	-.005661	0.120963	0.158032	0.181213	-.014331	0.546109	0.066369
D13	0.019331	-.028207	-.022477	-.037380	-.007156	0.076368	-.098051	-.025406	0.400102	0.062539	-.093443
D14	0.157856	0.014298	-.130327	-.047798	-.062774	-.176501	0.252465	-.121152	0.153077	-.205283	0.033699
D15	0.264888	-.008851	-.070228	-.032399	-.027534	-.047980	0.275085	-.050398	0.063604	-.105920	0.224244
D16	0.249182	-.097689	-.039914	0.019821	0.029495	0.029756	0.310873	0.086349	0.010628	0.034066	0.138140
D17	0.077848	-.012747	0.169492	-.023648	-.014917	-.032447	-.282444	0.180550	-.451669	-.001845	-.083049
D18	0.213205	0.041381	0.172959	0.013633	0.031767	0.003262	-.313312	0.095709	-.283351	-.060857	-.120684
D19	0.311710	0.032489	0.115798	0.025988	0.038739	0.011985	-.136468	0.047846	-.202653	-.039422	0.079368
D20	0.305650	-.050703	0.077132	0.049379	0.064819	0.042246	0.025246	0.123861	-.156100	0.032456	0.087130
D21	-.022353	0.020653	0.034451	-.195819	0.429177	-.093071	0.041106	0.050358	0.026130	0.048919	0.015087
D22	-.096147	0.266881	-.066690	-.201537	-.107221	0.058541	-.015329	0.296655	0.022510	0.012799	0.025236
D23	0.056742	0.364025	0.086240	0.092600	0.040260	0.003014	0.140613	-.216409	-.048914	0.118462	-.103382
D24	-.021027	-.001561	0.039719	-.201794	0.431789	-.088385	0.038019	0.027606	0.017387	0.048736	-.002142
D25	-.101112	0.257498	-.063527	-.210697	-.115426	0.054569	0.001866	0.295847	0.023039	0.018854	0.037765
D26	0.082887	0.377753	0.057801	0.054817	0.026601	0.009959	0.091204	-.197193	-.056076	0.119482	-.107210
D27	-.017509	-.014889	0.038964	-.196364	0.435744	-.085640	0.050864	0.048879	0.012736	0.052586	-.001368
D28	-.018130	0.004127	0.034075	-.190499	0.432188	-.087063	0.055633	0.071069	0.019234	0.050373	0.012872
D29	-.090016	0.256745	-.061930	-.204398	-.105422	0.062361	0.033599	0.319245	0.029305	0.034602	0.014554
D30	-.084387	0.265302	-.064156	-.196530	-.098676	0.067144	0.017775	0.321046	0.028012	0.027730	0.001626
D31	0.060684	0.361732	0.077631	0.088305	0.040099	0.010758	0.154747	-.210783	-.050069	0.125237	-.135305
D32	0.083776	0.371705	0.051496	0.054017	0.027596	0.017723	0.113854	-.187130	-.055994	0.124862	-.139388
D33	0.121134	0.108192	-.164365	0.263860	0.052222	-.235073	-.076461	0.177794	0.108115	-.026707	0.062986
D34	0.115287	0.119780	-.163699	0.268684	0.063934	-.219172	-.111915	0.178643	0.114120	-.017454	0.044869
D35	-.207869	0.055807	0.158763	0.276348	0.120438	0.147044	0.109484	0.089239	-.001910	-.137814	0.140121
D36	-.216008	0.029777	0.132706	0.279538	0.120944	0.164043	0.043731	0.117857	-.004729	-.128182	0.111931
D37	0.123194	0.111330	-.172312	0.243580	0.060740	-.230933	-.074430	0.179509	0.098445	-.016660	0.057357
D38	0.117044	0.123883	-.170236	0.249266	0.074117	-.211853	-.114685	0.178836	0.105094	-.002967	0.038495
D39	-.202828	0.072221	0.174497	0.266772	0.126195	0.144085	0.124985	0.065418	-.007065	-.126643	0.124274
D40	-.214596	0.046505	0.152734	0.272289	0.133239	0.169933	0.057343	0.093551	-.012737	-.118791	0.095708
D41	0.005483	0.004117	0.396558	-.012988	-.104803	-.221265	0.041872	0.088463	0.141168	0.009696	0.007071
D42	0.013820	-.017162	0.387001	-.037516	-.115957	-.219922	0.026635	0.072917	0.120772	0.019096	-.008209
D43	0.005474	-.001628	0.394209	-.017628	-.109114	-.227356	0.068638	0.077309	0.144873	0.010652	0.009458
D44	0.013897	-.027039	0.379486	-.043063	-.120707	-.226953	0.061398	0.056655	0.126935	0.018866	-.001444

Tableau 46. Résultats de l'analyse en composantes principales pour gdmap: vecteurs propres associés à chaque composante principale (suite).

	cp12	cp13	cp14	cp15	cp16	cp17	cp18	cp19	cp20	cp21	cp22
D1	-.259062	-.106055	-.203394	0.015604	0.243838	0.000443	-.024554	0.334135	0.372187	0.060071	-.171934
D2	-.179935	-.043179	-.076162	0.022495	0.123365	-.014995	0.030805	-.151845	-.170759	-.069393	0.017211
D3	-.180704	-.013852	0.033256	0.028937	-.150145	-.013436	0.000971	-.191987	-.241730	0.056205	0.345196
D4	-.103540	-.004231	0.121748	0.018129	-.376304	0.007498	-.028054	0.033621	0.071120	0.001617	0.285156
D5	-.219334	-.099986	-.178016	-.034901	0.386813	-.002991	0.062090	-.062958	-.106729	-.082614	0.003107
D6	-.243184	-.050891	-.127123	-.010779	0.087175	-.018250	0.003957	0.014075	-.015074	0.095821	-.003529
D7	-.165691	-.015017	-.022212	-.008349	-.215148	0.003367	-.024618	0.218636	0.293775	-.040692	-.216932
D8	-.080702	0.920457	-.134310	0.039114	0.083596	0.002267	0.006275	0.078204	0.108761	0.046200	0.134562
D9	0.394903	-.011076	-.335943	0.014100	0.209410	0.039970	0.037182	-.172371	-.239402	0.116456	0.164138
D10	0.257428	0.030824	-.139097	0.005994	-.040932	0.049029	0.005747	0.005771	0.005657	-.117627	-.206156
D11	0.014692	-.074203	-.060965	-.001696	0.025633	0.020166	0.020268	0.051281	0.074556	0.080566	0.427713
D12	0.108350	0.212311	-.059767	0.041593	0.082502	0.003355	0.001804	0.015383	0.024300	-.071215	-.164221
D13	-.095967	0.086905	0.772013	-.046331	0.395826	-.021050	-.015107	-.024049	-.002914	0.051277	0.068162
D14	0.368014	-.079161	-.087147	-.002550	0.204482	0.029629	0.016531	0.246198	0.325191	0.047333	0.364353
D15	0.281964	-.101911	0.042147	-.008730	0.110138	-.019298	0.005440	0.027665	0.058051	0.017025	-.024473
D16	0.285079	0.094617	0.085839	0.016399	0.031591	-.032148	0.002119	-.132185	-.145102	-.077576	-.355962
D17	0.051026	-.012250	0.105359	-.028648	0.422157	0.005276	0.050956	-.011349	0.000275	-.033083	0.008462
D18	0.160131	-.034785	0.021568	-.007204	0.117261	0.040202	0.012679	0.105168	0.121302	-.017405	0.027562
D19	0.129650	-.049707	0.133551	-.007512	-.111247	0.018255	-.028948	0.112774	0.121566	0.053094	0.011172
D20	0.198723	0.078061	0.181624	0.011060	-.186982	0.003512	-.038630	-.022088	-.052072	0.012673	-.143403
D21	-.030014	-.074055	0.025795	0.325912	0.015119	-.013348	0.379227	0.025503	0.030995	-.150169	0.063440
D22	0.094338	-.001770	0.026239	0.318408	-.018391	-.179113	-.239682	0.065056	-.044373	-.058670	-.048965
D23	-.019088	0.025843	0.009542	0.202191	0.057858	0.005767	-.211797	-.038578	-.057451	0.156121	0.056927
D24	-.037374	-.083074	0.012663	0.300916	0.012606	-.015322	0.389883	0.030728	0.046434	0.148444	-.072876
D25	0.069686	-.002703	0.026292	0.328741	-.015937	0.193904	-.226453	-.046520	0.031423	-.068283	-.029224
D26	-.008077	0.022924	0.021114	0.236050	0.045046	0.010410	-.233490	0.030153	0.013604	-.071037	0.123517
D27	-.030191	-.062480	-.052774	-.295479	0.053444	0.014423	-.384534	0.011222	0.006978	0.113360	-.069025
D28	-.020698	-.046880	-.041264	-.330517	0.052160	0.020270	-.381625	0.005303	-.006205	-.150844	0.058636
D29	0.082854	0.032372	0.023571	-.303096	-.100265	0.178544	0.222337	-.046530	0.059767	0.095473	0.112518
D30	0.106743	0.030381	0.021709	-.311809	-.099268	-.165796	0.209064	0.062525	-.013502	0.106598	0.087774
D31	-.024864	0.048000	0.007615	-.195624	0.029967	-.004333	0.195812	-.057079	-.037118	-.011634	-.148219
D32	-.013268	0.047999	0.020329	-.248079	0.005543	0.000225	0.233221	0.006080	0.031483	-.209598	-.086827
D33	-.114648	-.019158	-.048774	-.000771	0.062097	0.416536	0.022196	-.123971	0.014632	-.297178	0.043759
D34	-.071943	-.018031	-.073043	-.015734	0.060079	-.410519	0.009079	0.081415	-.113808	-.293610	0.080959
D35	0.062501	-.038018	0.083804	-.005408	-.009369	0.009061	-.000981	0.061303	0.088251	0.003042	0.053152
D36	0.063671	-.024939	0.069138	-.013085	-.002761	0.010833	0.005401	0.069396	0.064190	-.385211	0.063776
D37	-.120212	-.019164	-.044082	0.014813	0.052978	0.463683	0.015441	-.140059	0.046046	0.302299	-.125740
D38	-.071624	-.017525	-.069780	-.003845	0.049216	-.468810	-.000937	0.092511	-.104781	0.287751	-.083866
D39	0.061186	-.036259	0.075402	0.007283	-.008875	0.000617	-.009769	0.028792	0.061321	0.338218	-.032867
D40	0.066180	-.023868	0.061357	0.000718	-.005522	0.002755	-.002126	0.038227	0.041602	0.016890	-.044066
D41	-.009775	-.005296	-.069598	0.017971	0.026240	-.138804	-.016228	-.389379	0.205109	0.163784	-.032113
D42	-.016031	0.000114	-.067592	0.000018	0.006601	-.164250	-.008059	-.401538	0.309195	-.187469	0.060322
D43	-.012656	0.008592	-.037222	0.006818	0.017968	0.113729	0.010642	0.264285	-.332732	0.184012	-.036195
D44	-.020350	0.014086	-.024218	-.015837	-.005925	0.153642	0.031629	0.411276	-.354089	-.155664	0.050364

Tableau 47. Résultats de l'analyse en composantes principales pour gdmap: vecteurs propres associés à chaque composante principale (suite).

	cp23	cp24	cp25	cp26	cp27	cp28	cp29	cp30	cp31	cp32	cp33
D1	-.183187	-.034742	-.245040	-.145318	0.035086	0.148985	0.026531	0.011831	-.146245	0.000393	0.028599
D2	0.079636	0.009598	0.204658	0.162922	0.003632	0.093515	0.003767	0.045688	0.301849	-.009398	-.104305
D3	0.068792	0.045840	-.025336	0.070518	-.010154	-.116370	-.005419	-.012796	0.212116	-.000341	0.086292
D4	0.240247	-.015604	0.118262	-.025346	-.039428	-.089239	-.021798	0.002632	-.556796	0.008094	-.000672
D5	0.222256	-.009907	0.100560	-.487717	-.096370	0.023380	-.001882	0.012588	-.219217	-.010627	-.060572
D6	-.264867	0.049943	-.213199	0.257091	0.059475	-.159696	-.017843	-.067923	0.298811	0.008140	0.145372
D7	-.053577	-.038809	0.150010	0.356541	0.036352	-.038640	-.023837	-.032016	-.119866	0.038647	0.012753
D8	0.023303	-.007429	-.043078	0.019191	0.014757	0.146425	0.037835	0.067212	0.013109	-.010622	-.022726
D9	-.179094	0.065878	-.324785	0.078230	0.063456	-.094970	0.010630	-.059405	-.305504	0.060150	-.090778
D10	0.185347	-.090910	0.317266	-.176842	-.064625	0.217836	0.063535	0.059417	0.301063	-.048673	0.084973
D11	0.131234	-.008700	-.040851	0.087222	0.027755	0.426383	0.094898	0.179229	0.070856	-.014867	-.016088
D12	0.043940	0.010779	0.144206	-.054593	-.077708	-.567201	-.126816	-.192335	0.046895	-.011892	-.009299
D13	-.055602	0.026154	-.113061	-.025358	0.015169	-.046127	-.000115	-.010542	0.007713	-.000312	-.022950
D14	0.157439	0.027485	0.033598	0.028682	-.036956	-.270956	-.082336	-.069439	0.140411	0.035108	0.212257
D15	-.062739	0.011587	0.016260	0.228905	0.055153	0.243666	0.054020	0.111445	-.054259	-.044148	-.102630
D16	-.138461	0.002574	0.084434	0.177912	0.042814	0.138024	0.038917	0.035692	-.191408	0.003048	-.060939
D17	0.087266	-.046533	0.262892	0.305260	0.008132	0.054671	0.027693	0.004924	-.184537	0.041031	0.184883
D18	0.117460	-.032254	0.130241	-.101660	-.039273	-.064069	-.011967	0.002120	0.157518	-.077793	-.200196
D19	-.054216	-.005396	-.205373	-.216751	0.012581	0.071517	0.003947	0.026112	0.243620	-.026081	-.098542
D20	-.095005	-.011280	-.269291	-.381544	0.002497	0.082988	0.024387	-.017340	0.026868	0.066127	0.205364
D21	-.071024	-.089904	-.013231	-.011457	0.007592	0.175372	-.241115	-.380497	-.001574	-.022978	0.004090
D22	0.219315	0.017931	-.248536	0.158591	-.406088	0.033700	0.055960	-.012760	0.017193	0.000803	0.010381
D23	-.226429	-.180086	0.275450	-.119130	-.051635	0.067340	0.131039	-.058183	-.048819	0.153887	0.324477
D24	-.014132	0.079593	0.020264	-.031365	-.023216	-.178255	0.232344	0.375442	-.001582	0.027431	0.002783
D25	0.228244	-.002612	-.115997	0.025632	0.491389	-.035229	0.025004	0.017928	0.019264	-.010074	0.022088
D26	-.290004	0.185202	0.226294	-.094721	-.049597	0.001359	-.163965	0.119249	-.021388	-.110880	-.348075
D27	0.063394	0.076260	-.004106	-.000313	-.011860	-.180787	0.218924	0.354460	-.000063	0.031345	-.004917
D28	0.003716	-.073828	-.024314	0.018620	0.020787	0.155255	-.219492	-.355391	0.004492	-.032448	0.003144
D29	-.236302	-.021475	0.204199	-.120080	0.416884	-.034304	-.016130	0.028018	0.000340	-.004586	0.026053
D30	-.228976	0.000024	0.067632	0.010085	-.479734	0.044452	0.015423	-.000311	-.001252	0.015686	0.006420
D31	0.328188	-.153888	-.163931	0.060872	0.041982	0.040602	0.148567	-.077147	-.012625	0.142535	0.334250
D32	0.264501	0.172241	-.207398	0.080799	0.039435	-.024615	-.127437	0.086590	0.008289	-.133189	-.330959
D33	-.125962	-.226411	-.089941	0.036798	-.207288	-.060029	0.117838	0.101455	-.022901	-.226781	0.061133
D34	-.093800	-.237161	-.008175	-.038811	0.187141	-.097269	0.103942	0.117700	-.013299	-.216690	0.056542
D35	0.021129	-.318444	0.004321	0.028156	0.009973	-.069864	0.207681	-.105505	0.056739	0.349107	-.361774
D36	-.063268	0.135550	-.004794	-.002418	-.010204	0.005053	-.212464	0.241919	0.024431	0.475054	0.119748
D37	0.099829	0.265612	0.035001	0.014972	-.157961	0.084979	-.074518	-.124400	0.015198	0.223017	-.073429
D38	0.128162	0.238238	0.103838	-.053581	0.184339	0.059939	-.086116	-.115444	0.020345	0.204127	-.065460
D39	0.087772	-.157286	0.033352	0.031302	0.007846	-.026125	0.223448	-.230606	-.005356	-.429615	-.117011
D40	0.024201	0.415031	0.039506	-.000405	-.008631	0.060783	-.212318	0.102220	-.044562	-.418044	0.358933
D41	0.010018	-.227671	-.079352	0.002636	-.021689	-.019315	-.340010	0.188790	0.008392	-.008145	-.006270
D42	-.080009	0.299689	-.019315	-.035271	0.004159	0.002757	0.337477	-.192549	0.009752	0.025474	-.000977
D43	0.030502	-.239612	-.059633	0.039034	-.000080	-.018263	-.332418	0.178589	-.003279	-.011066	-.006854
D44	-.052255	0.277965	0.008397	0.000987	0.034035	-.000262	0.313830	-.174978	-.000017	0.022406	-.009028

Tableau 48. Résultats de l'analyse en composantes principales pour gdmap: vecteurs propres associés à chaque composante principale (suite).

	cp34	cp35	cp36	cp37	cp38	cp39	cp40	cp41	cp42	cp43	cp44
D1	0.163912	-.082941	-.000804	0.066044	-.000651	0.003178	-.001113	0.013633	-.000001	0.000435	-.000323
D2	-.473620	0.244809	0.033452	-.317338	0.013462	-.012930	0.009357	-.055402	0.000075	0.001579	-.001097
D3	0.281639	-.112955	-.070634	0.592089	-.023719	0.024518	-.021620	0.029570	0.000811	-.001920	0.002349
D4	0.104957	-.108446	-.000401	-.417025	0.015349	-.018353	0.020092	0.036710	0.001427	0.000176	-.000811
D5	-.041354	0.009977	0.007734	0.131394	-.001464	0.008345	-.008259	0.047394	0.000999	-.000775	0.000582
D6	0.320063	-.085504	0.120482	-.418608	0.014537	-.020948	0.020075	-.041414	-.001733	0.003524	-.001214
D7	-.332643	0.227414	0.023947	0.391821	-.015500	0.020531	-.021366	-.058416	-.002335	-.001145	0.000853
D8	-.014767	-.031572	-.022855	0.002610	-.000523	-.002575	0.001820	-.002219	0.000198	0.000780	0.000093
D9	-.217203	0.048233	-.028741	0.041056	-.003153	0.004756	0.000658	0.010444	0.000828	0.001719	0.000441
D10	0.322549	0.020558	-.116788	-.073829	0.004334	-.009571	0.006249	0.074359	0.001379	-.001224	0.002234
D11	-.047729	0.066327	0.164435	0.004666	0.004312	0.002634	-.003409	-.008628	-.003300	0.001771	-.000448
D12	-.061050	-.104753	-.236699	-.034645	-.004378	-.003086	0.005288	0.001788	0.001788	-.001056	0.000433
D13	-.066485	-.039170	0.036809	0.003809	-.000778	0.001188	-.001512	-.018724	-.000408	0.000951	-.000836
D14	0.030384	0.232123	0.198903	0.009688	0.001226	0.002953	-.006542	0.085542	0.004480	0.005350	-.001722
D15	0.018227	-.315910	-.548221	-.008680	-.003306	-.011499	0.012112	-.341214	-.009650	-.012842	0.003866
D16	0.157953	-.096620	0.521478	0.069713	0.001705	0.011657	-.010097	0.309043	0.007360	0.011664	-.002311
D17	0.237222	0.283542	-.207187	-.002758	-.002207	-.006726	0.006415	0.132552	0.003738	-.001489	0.001349
D18	-.132815	-.436002	0.402983	0.049123	0.006418	0.008871	-.007421	-.392885	-.011898	-.002134	0.000021
D19	-.270468	-.113412	-.244343	-.046973	-.004643	-.001632	0.003431	0.646844	0.023618	0.016927	-.002790
D20	0.050108	0.496484	0.010372	-.018604	0.004109	-.003914	-.002354	-.411441	-.016644	-.013746	0.002122
D21	-.003373	-.025121	0.002305	0.023500	0.024067	-.470457	0.043171	-.000878	-.002680	0.027883	-.004042
D22	0.000240	-.019365	0.001851	0.019555	0.493425	0.047768	0.045159	0.006106	-.041830	0.032225	-.001932
D23	-.155997	-.117403	0.002545	-.007633	-.013222	0.031738	-.011157	-.010063	0.020006	0.506221	-.030409
D24	0.002150	0.019972	-.010881	-.027406	-.025420	0.485380	-.046453	-.000086	0.003453	-.024560	0.004816
D25	0.014253	-.023986	0.010965	-.015105	-.491224	-.018066	-.043624	-.002428	0.042988	0.011588	-.000851
D26	0.114165	0.187945	-.013072	0.001903	0.007494	-.018335	0.009349	0.012058	-.017571	-.475445	0.027914
D27	-.006239	0.017036	0.010329	0.026972	0.028534	-.518366	0.044713	0.002671	0.005232	0.026042	0.003027
D28	0.003172	-.021790	-.009558	-.027564	-.027453	0.506745	-.041805	-.004058	-.006276	-.028748	-.003124
D29	0.001159	-.015815	-.004162	0.024524	0.498502	0.016148	0.043628	0.004586	0.003458	-.011414	0.001838
D30	-.008136	-.010499	0.009873	-.016141	-.501251	-.044803	-.042303	-.000315	-.003726	-.031478	-.001493
D31	-.145390	-.146070	0.017986	-.008266	0.012266	-.029223	0.019723	0.015534	-.021693	-.516914	-.004913
D32	0.132338	0.166648	-.020343	-.000840	-.005822	0.017452	-.016796	-.016470	0.019038	0.491212	0.004928
D33	-.024744	0.010786	0.001500	0.002639	0.021592	0.004791	0.017402	-.014254	0.510182	-.022644	-.008377
D34	-.030096	0.012684	-.000189	-.001294	-.028456	-.001369	-.017632	0.018826	-.525555	0.017862	-.025852
D35	0.124778	0.086515	-.012463	0.001995	0.003934	0.003324	0.001587	-.006326	0.012934	-.021855	-.537241
D36	-.022693	-.106294	-.000305	-.000364	-.000138	0.006735	-.004127	-.002374	-.002179	0.023967	0.468971
D37	0.019443	0.005219	-.005688	-.006478	-.007414	-.003663	-.015033	0.013293	-.468456	0.020138	0.008536
D38	0.013459	0.011043	0.000442	-.003456	0.013796	0.002126	0.016042	-.018074	0.486242	-.016262	0.024220
D39	0.039258	0.094933	0.006254	0.008052	-.002331	-.006219	0.004567	0.005460	-.012933	-.001225	0.534115
D40	-.098293	-.077723	0.011121	0.012854	-.002048	-.005037	-.003171	0.001473	0.002995	-.004731	-.450221
D41	0.011483	0.005753	0.012184	0.021400	-.052500	0.048976	0.537616	-.000095	-.015584	0.010280	-.008356
D42	0.004003	-.018130	-.010519	-.018883	0.039588	-.038087	-.429266	0.002232	0.013280	-.006904	0.007399
D43	0.011438	-.000412	-.007946	-.020104	0.054874	-.046610	-.552578	-.000152	0.015607	-.016697	-.006000
D44	-.006153	-.008754	0.001786	0.011946	-.042226	0.037639	0.448934	0.002528	-.013215	0.013115	0.005224

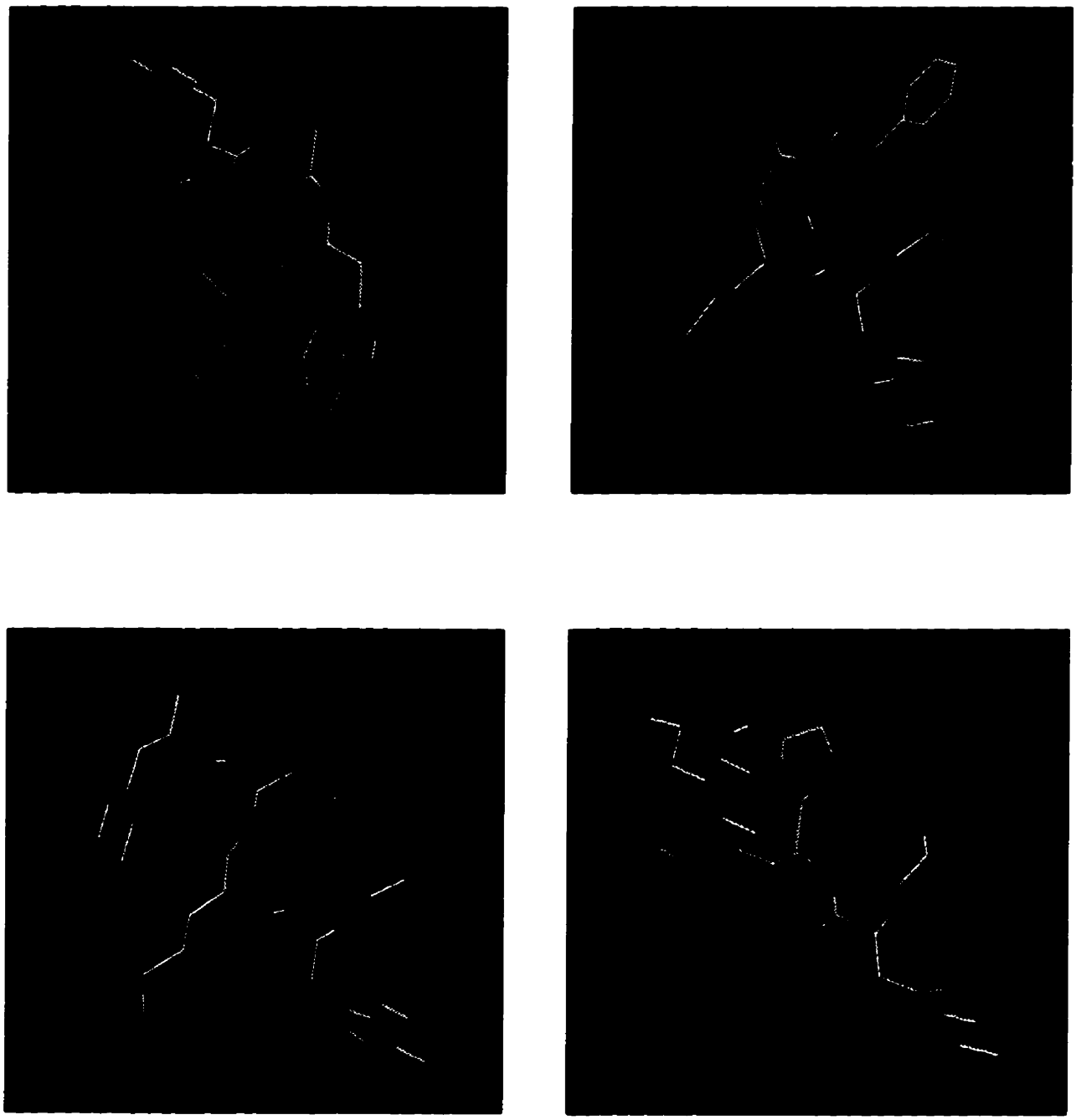


Figure 64. Conformation type pour chaque famille de gdmap: de gauche à droite et de haut en bas, famille 1, 2, 3, 4.

Figure 65. Conformation type pour chaque famille de gdmap: de gauche à droite et de haut en bas, famille 5, 6, 7, 8.

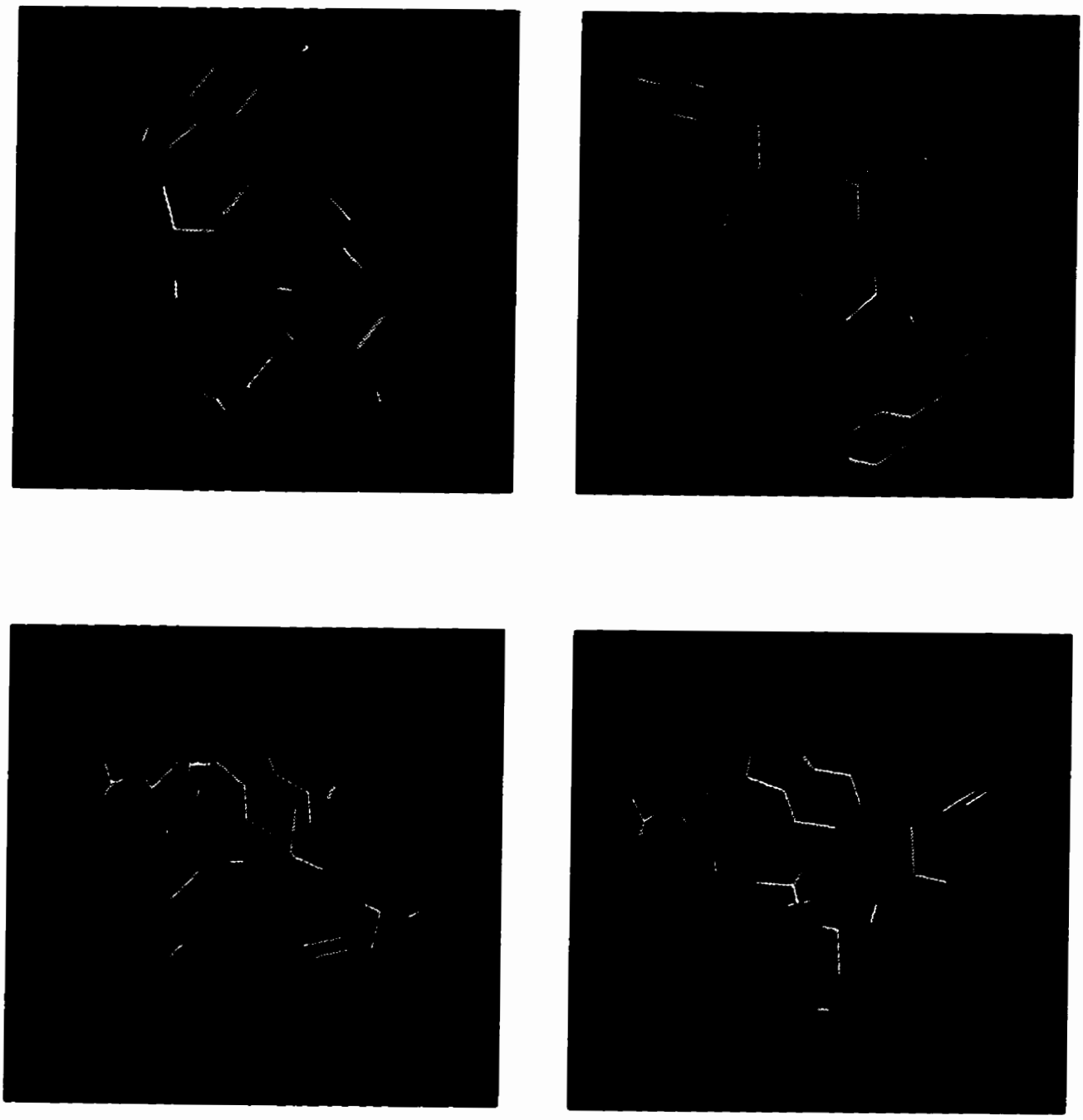


Figure 66. Conformation type pour chaque famille de gdmap: de gauche à droite et de haut en bas, famille 9, 10, 11, 12.

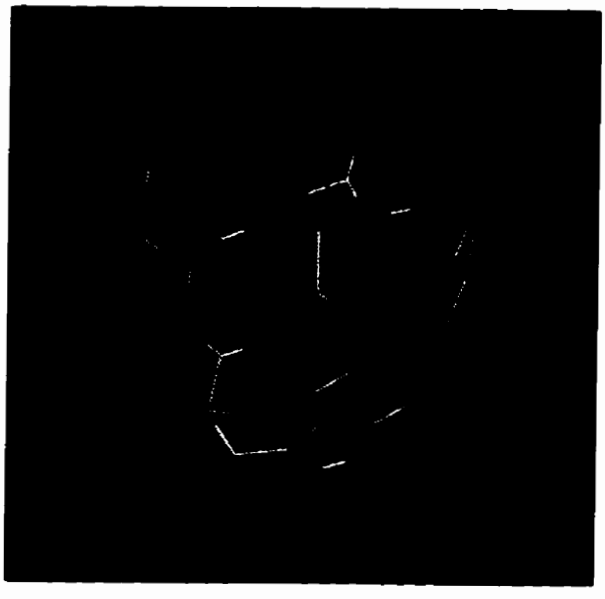


Figure 67. Conformation type pour la dernière famille de gdmap: famille 13.

BIBLIOGRAPHIE

1. F.R.N. GURD et T.M. ROTHGEB. Adv. Prot. Chem. 33, 73 (1979).
2. Z. Li et H.A. SCHERAGA. Proc. Natl. Acad. Sci. USA. 84, 6611 (1987).
3. Dans Acta Crystallographica, Section D.
4. D.L. OXENDER et C.F. FOX. Protein Engineering. A.R. Liss Inc. New-York, 1987.
5. B. ROUX. Dans Simulations numériques en physique. Vol 1, Edité par L. Lewis, J. Lopez, G. Slater et A-M.S. Tremblay. CRPS, Sherbrooke, 1993. Chap.3.
6. M. SAUNDERS, K.N. HOUK, Y-D. WU, W.C. STILL, M. LIPTON, G. CHANG et W.C. GUIDA. J. Am. Chem. Soc. 112, 1419 (1990).
7. C. LANDIS et V.S. ALLURED. J. Am. Chem. Soc. 113, 9493 (1991).
8. A. DI NOLA, H.J.C. BERENDSEN, O. EDHOLM, Macromolecules. 17, 2044 (1984).
9. R. FUSCO, L. CACCIANOTTI et C. TOSI. Dans Strategies for computer chemistry. Kluwer, Amsterdam, 1989.
10. R. SCORDAMAGLIA et L. BARINO dans Strategies for computer chemistry. Kluwer, Amsterdam, 1989.
11. J. M. GOODFELLOW et D.S. MOSS. Computer modelling of biomolecular process. Ellis, Horwood, 1992.
12. C.B. ANFINSEN. Science. 181, 223 (1973).
13. D.B. WETLAUFER et S. RISTOW. Ann. Rev. Biochem. 42, 135 (1973).
14. R. BRUSCHWEILER, M BLACKLEDGE et R.R. ERNST. J. Biomol. NMR. 1, 3 (1991).
15. O. EDHOLM et H.J.C. BERENDSEN. Molecular physics. 51, 1011 (1984).
16. M. KARPLUS et J.N. KUSHICK. Macromolecules. 14, 325 (1981).

17. H. MEIROVITCH, D. KITSON et A. HAGLER. rapport # FSU-SRCI-92-09, Supercomputer computation research institute, Tallahassee, FL 1992.

18. A.E. HOWARD et P.A. KOLLMAN. J. Med. Chem. 31, 1669 (1988).

19. F. LEMAY, C. AMEZIANE-HASSANI et A.G. MICHEL. Can. J. Chem. 68, 1186 (1990).

20. N. GO et H.A. SCHERAGA. Macromolecules. 9, 535 (1976).

21. A.T. HAGLER, P.S. STERN, R. SHARON, J.M. BECKER et F. NAIDER. J. Am. Chem. Soc. 101, 6842 (1979).

22. D.L. BEVERIDGE et F.M. DICAPUA. Annu. Rev. Biophys., Biophys. Chem. 18, 431 (1989).

23. J.P. STOESSEL et P. NOWAK. Macromolecules. 23, 1961 (1990).

24. S.S. ZIMMERMAN et H.A. SCHERAGA. Biopolymers. 16, 811 (1977).

25. J. LEFBVRE. Introduction aux statistiques multidimensionnelles. Masson, Paris, 1976.

26. M. JAMBU, M.O. LEBEAUX. Classification automatique pour l'analyse des données. Dunod, Paris, 1983.

27. J.P. BENZECRI. Cahier de l'analyse des données. 1, 9 (1976).

28. G. PHILIPPEAU. Rapport intitulé Comment interpréter les résultats d'une analyse en composantes principales. I.T.C.F. Paris, 1986.

29. D.L. MASSART et L. KAUFMAN. The interpretation of analytical chemical data by the use of cluster analysis. John Wiley and Sons, New-York, 1983.

30. B.S. EVERITT. Cluster analysis. Heineman Educational books Ltd. London, 1980.

31. L. LEBART, A. MORINEAU et J-P. FENELON. Traitement des données statistiques. Dunod, Paris 1982.

32. B.S. EVERITT. Biometrics. 35, 169 (1979).

33. W.S. SARLE.The cubic clustering criterion. Rapport technique SAS # A-108. SAS Institute Inc., Cary N-C, 1983.

34. T. CALINSKI et J. HARABASZ. Communications in statistics. 3, 1 (1974).

35. R.O. DUDA et P.E. HART. Pattern classification and scene analysis. John Wiley and sons, New-York, 1973.

36. W.S. SARLE. Proceedings of the seventh annual SAS users group international conference. SAS Institute Inc., Cary N-C, 1982.

37. Dans SAS User's guide: Statistics Version 5. SAS Institute Inc., Cary N-C, 1985. p. 48.

38. G. W. MILLIGAN. Multivariate Behavioral Research. 16, 379 (1981).

39. J.E. MEZZICH et H. SOLOMON. Taxonomy and behavioral science. Academic Press, New-York, 1980.

40. C. AMEZIANE-HASSANI. Thèse de doctorat. Université de Sherbrooke, Québec 1991.

41. SAS version 6.08. Sas Institute Inc., Cary N-C, 1993.

42. W.H. PRESS, B.P. FLANNERY, S.A. TEUKOLSKY et W.T. VETTERLING. Numerical Recipes. Cambridge University Press, Cambridge, 1986.

43. J-F. COLONNA. Pour la science. 179, 104 (1992).

44. B.W. KLENZ. Rapport technique #230. SAS Institute Inc., Cary NC, 1994.

45. R.D. LANGSTON. Proceedings of the Twelfth Annual SAS Users Group International Conference. SAS Institute Inc., Cary N-C, 1987.

46. SAS User's guide: Basics Version 5. SAS Institute Inc., Cary N-C, 1985.

47. J. ZUPAN. Algorithms for chemists. John Wiley and sons, New-York, 1989.

48. G. FIAMENBAUM, E. ABILLON et J.P. BENZECRI. les cahiers de l'analyse des données. 4, 339 (1979).

49. B. MAIGRET et S. PREMILLAT. Biochem. and Biophys. Res. Comm. 104, 971 (1982).

50. C. MARCHIONINI, B. MAIGRET et S. PREMILLAT. Biochem. and Biophys. Res. Comm. 112, 339 (1983).

51. M. BENKOULOUCHE, M. COTRAIT et B. MAIGRET. Journal of Computer-Aided Molecular Design. 6, 79 (1992).

52. M. KREISSLER, M. PESQUER, B. MAIGRET, M.C. FOURNIE-ZALUSKI et B.P. ROQUES. Journal of Computer-Aided Molecular Design. 3, 85 (1989).

53. A. PERCZEL, J.G. ANGYAN, M. KATJAR, W. VIVIANI, J-L. RIVAIL, J-F. MARCOCCIA et I.G. CSIZMADIA. J. Am. Chem. Soc. 113, 6256 (1991).

54. M. VASQUEZ, G. NEMETHY et H.A. SCHERAGA. Macromolecules. 16, 1043 (1983).

55. S.S. ZIMMERMAN, M.S. POTTLE, G. NEMETHY et H.A. SCHERAGA. Macromolecules. 10, 1 (1977).

56. P.N. LEWIS, F.A. MOMANY et H.A. SCHERAGA. Israel journal of Chemistry. 11, 121 (1973).

57. D. NEUHAUS et M. WILLIAMSON. The Nuclear Overhauser Effect in structural and conformational analysis. VCH publishers Inc., New-York,.

58. F. LEMAY. Mémoire de Maîtrise. Université de Sherbrooke, Québec, 1991.

59. V.S. ANANTHANARAYANAN. Proc. 12th Am. Pep. Sym. 82 (1991).

60. A.G. MICHEL, C. AMEZIANE-HASSANI et N. BREDIN. Can. J. Chem. 70, 569 (1992).

61. A.G. MICHEL et C. JEANDENANS. Computers Chem. 17, 49 (1993).

62. M.C. BEINFELD. Neuropeptides. 3, 411 (1983).

63. V. MUTT. Dans Cholecystokinin: isolation, structure, and functions in Gastrointestinal hormones. Raven Press, New-York, 1980. p. 169.

64. J.E. MORLEY. Life Science. 27, 355 (1980).

65. P.C. EMSON, S.P. HUNT, J.P. REHFELD, N. GOLTERMAN et J. FAHRENKRUG, Dans Neural peptides and neuronal communication. Raven Press, New-York, 1980. p. 63.

66. J.S. KELLY et J. DODD. Dans Neurosecretion and brain peptides. Edité par J.B. Martin, S. Reichlin et K.L. Bick. Raven Press, New-York, 1981. p. 133.

67. G.J. DOCKRAY. Dans Gut peptides. Edité par A. Miyoshi. Elsevier/North, Amsterdam, 1980. p. 237.

68. S. PAUWELS. Biochim. Biophys. Acta. 996, 82 (1989).

69. T.H. MORAN, P.H. ROBINSON, M.S. GOLDRICH et P.R. McHUGH. Brain Research. 362, 175 (1986).

70. J.Y. MELVIN, K.J. TRASHER, J.R. McCOWAN, N.R. MASON et L.G. MENDELSOHN. J. Med. Chem. 34, 1508 (1991).

71. M.G. BOCK, R.M. DIPARDO, B.E. EVANS, K.E. RITTLE, W.L. WHITTER, D.F VEBER, P.S. ANDERSON et R.M. FREIDINGER. J. Med. Chem. 32, 16 (1989).

72. W.B.T. CRUSE, E. EGERT, M.A. VISWAMITRA et O. KENNARD. Acta Cryst. B38, 1758 (1982).

73. R.A. HUGHES et P.R. ANDREWS. Pept.: Chem. Biol., Proc. Am. Pept. Symp. 10th, Escom Sci. Leiden, Netherlands, 1988. p. 115.

74. G. GRECO, E. NOVELLINO, C. SILIPO et A. VITTORIA. Quant. Struct.-Act. Rela. 11, 461 (1992).

75. M. RUIZ-GAYO, V. DAUGE, I. MENANT, D. BEGUE, G. GACEL et B.P. ROQUES. Peptides. 6, 415 (1985).

76. M.C. FOURNIE-ZALUSKI, C. DURIEUX, B. LUX, J. BELLENEY, P. PHAM, D. GERARD et B.P. ROQUES. Biopolymers. 24, 1663 (1985).

77. S. WOLFE, S. BRUDER, D.F. WEAVER et K. YANG. Can. J. Chem. 66, 2703 (1988).

78. C. JEANDENANS. Poster. 76ème congrès de la société canadienne de chimie. Sherbrooke, Québec, Mai 1992

79. Y-M LEE, M. BEINBORN, E.W. McBRIDE, M. LU, L.F. KLAKOWSKY et A.S. KOPIN. J. Mol. Biol. 268, 8164 (1993).

80. S.A. WANK, R. HARKINS, R.T. JENSEN, H. SHAPIRA, A. de WEERTH et T. SLATTERY. Proc. Natl. Acad. Sci. USA. 89, 3125 (1992).

81. M. KURODA, K. YAMAZAKI et T. TAGA. Int. J. Peptides Prot. Res. 44, 499 (1994).

82. M. KURODA, K. YAMAZAKI, T. KOBAYASHI, N. FUJI et T. TAGA. Bull. Chem. Soc. Jpn. 67, 648 (1994).

83. D. PATTOU, B. MAIGRET, M.-C. FOURNIE-ZALUSKI et B.P. ROQUES. Int. J. Peptide Protein Res. 37, 440 (1991).

84. S.B. KALINDJIAN, M.J. BODKIN, I.M. BUCK, D.J. DUNSTONE, C.M.R. LOW, I.M. McDONALD, M.J. PETHER et K.I.M. STEEL. J. Med. Chem. 37, 3671 (1994).

85. J.A. LOWE, D.L. HAGEMAN, S.E. DROZDA, S. McLEAN, D.K. BRYCE, R.T. CRAWFORD, S. ZORN, J. MORRONE et J. BORDNER. J. Med. Chem. 37, 3789 (1994).

86. G. GRECO, E. NOVELLINO, C. SILIPO et A.VITTORIA, Dans Proceeding of the XI European Symposium on Quantitative Structure-Activity Relationships. Edité par C. Silipo et V. Vittoria. Elsevier, Amsterdam 1993. p. 293.

87. S. RAULT, R. BUREAU, J.C. PILO et M. ROBBA. J. Computer-Aided Mol. Design. 6, 553 (1992).

88. S. RAULT, R. BUREAU, J.C. PILO et M. ROBBA, Dans Proceeding of the XI European Symposium on Quantitative Structure-Activity Relationships. Edité par C. Silipo et V. Vittoria. Elsevier, Amsterdam 1993. p. 295.