Deuxième session M2 Biologie informatique

Année 2010/2011

Classification de ligands

Table des matières

Données sources : Descripteurs_Molecules.dat
Fichier joint : ANALYSE_FERNANDEZ.r

Guillaume Fernandez

guillaume.fernandez@alumni.u-psud.fr

Introduction

Le jeu de données étudié est composé de 478 ligands décrits par 31 descripteurs. L'ensemble des ligands a été classé en plusieurs groupes, suivant le type de protéines que le ligand fixe. Le but du projet est d'étudier si l'ensemble des descripteurs permet de retrouver les groupes de ligands.

Considérant :
1/ qu'il existe une connaissance structurée *a priori* des données, mais que rien ne dit que les descripteurs ont un lien avec la relation ligand-protéine (on ne connaît pas la nature des descripteurs) ;
2/ qu'il s'agit de confirmer ou d'infirmer ce lien ;
on adoptera une méthode de classification non supervisée (c'est-à-dire qui n'incorpore aucun présupposé).

Une première approche pour commencer à caractériser le groupe de données est de l'analyser variable après variable. Dans cette analyse univariée, chaque critère est analysé sans tenir compte des autres. Les ligands seront classés sur la base de leurs similarités en utilisant un classement hiérarchique (permet de générer des suites de classes emboîtées). Un algorithme ascendant semble répondre correctement à la question posée (construction des groupes par agrégations successives des éléments les plus proches deux à deux pour fournir la hiérarchie de partitions). La hiérarchie obtenue peut être représentée sous la forme d'un dendrogramme. Les individus qui se ressemblent le plus se regroupent dans le bas de l'arbre. La longueur des branches témoigne de leur éloignement. En choisissant une résolution (c'est-à-dire la plus petite distance visible entre entités), il est possible d'estimer le nombre de classes de ligands. Il est possible à l'inverse de forcer la partition des molécules en 72 groupes (ce qui conditionne la résolution). On observera si les groupes formés sont cohérents et superposables avec le classement sur la base de la relation ligand-protéine.

Pour compléter l'analyse, ou bien si les résultats proposés à cette étape sont incohérents, il sera possible de mener une analyse multi-critères. Une analyse multivariée permet de tenir compte de l'interaction des descripteurs. Il paraît élégant de commencer par réduire la complexité du problème (en perdant le minimum d'information). La première étape sera une analyse en composantes principales. Elle permettra d'identifier les descripteurs contribuant majoritairement à la classification des données. Elle peut également fournir une classification. La deuxième étape sera de réitérer la classification hiérarchique ascendante sur le jeu restreint de contributeurs majeurs. En dernier recours, si aucun classement ne peut être donné par cette approche, une dernière approche basée sur l'ensemble des descripteurs pourra être proposée.

Méthode

Classification hiérarchique

Les algorithmes de classification travaillent à partir des matrices de distances issues des matrices de données. Les distances calculées reflètent la proximité entre deux entités. Une première matrice de distance est donc calculée à partir des données initiales (fonction `dist()` de R). On a fait le choix d'utiliser des distances euclidiennes (la distance entre deux entités est la racine carrée de la somme des carrés des distances entre descripteurs $d(A,B)=\sqrt{\sum_{i=1}^{n}(a_i-b_i)^2} \quad a\in A, b\in B$).

L'algorithme hiérarchique agglomératif est ensuite appliqué (fonction `hclust()`). A la première étape du traitement, la distance calculée est donc celle qui sépare deux ligands. Un premier groupe est formé et la matrice de distance est mise à jour pour incorporer la distance entre ce groupe et les ligands restants. Et ainsi de suite jusqu'à ce qu'il n'y ait plus qu'un seul groupe. A chaque itération, les distances sont calculées avec la méthode du lien moyen[1] (approche la plus utilisée). La distance entre deux groupes est la moyenne entre toutes les paires d'objets (ici de ligands) qui les composent $d(A,B)=moyenne(d(a,b)) \quad a\in A, b\in B$. C'est une approche sûre sans connaissance supplémentaire sur la nature des descripteurs. La hiérarchie des partitions est ensuite tracée sous forme d'arbre (fonction `plot()`). Cela permet de comparer les structure d'arbre pour différentes données initiales.

On coupe ensuite l'arbre à une hauteur permettant d'obtenir le nombre de classes testé (par exemple 72). Cela peut être fait à l'aide de l'objet créé par la fonction `hclust()` qui contient la description de la hiérarchie des partitions et les distances associées (notamment le tableau du nombre de classes en fonction de la distance au bas de l'arbre). Pour savoir à quelle distance couper l'arbre pour obtenir 72 classes il faut déterminer combien de pas d'agglomération sont nécessaires pour atteindre ce nombre. Au départ il y a 478 classes composées de 1 entité. La première opération d'agglomération réduit ce nombre à 477. Chaque nouvelle opération réduit d'une unité le nombre de classes. Celui-ci atteint donc 72 à la 406ème opération. La distance correspondante se lit au rang 406 dans la liste des distances. On place le seuil légèrement au dessus (fonction `cutree()`).

1 ou UPGMA (Unweighted Pair Group Method of Agregation).

Plan de travail

Analyse univariée

La méthode qui vient d'être exposée est appliquée une première fois d'ans l'analyse univariée de deux descripteurs choisis au hasard (d9 et d26).

Analyse multivariée

☞ Réduction de la complexité du problème : déterminer les descripteurs influençant le plus la classification des ligands

Pour déterminer l'influence des descripteurs sur la classification on utilise une méthode d'analyse factorielle, l'analyse en composantes principales (ACP).

Face au nuage de points représentant la variation statistique des descripteurs en fonction des ligands, l'analyse en composantes principales dégage les combinaisons de variables (= les composantes principales) qui rendent le mieux compte de la dispersion des points.

Par construction, la première composante principale est celle qui explique le mieux la variation des descripteurs, le pouvoir explicatif de la seconde lui est inférieur et ainsi de suite. L'ACP fournit une évaluation chiffrée du pouvoir explicatif de chaque composante principale par la proportion de la variance du nuage de points dont cette composante rend compte. La somme cumulée des premières valeurs décroissantes permet en principe de se satisfaire d'un nombre de composantes principales inférieurs au nombre total de descripteurs en se fixant un seuil de pouvoir explicatif considéré comme satisfaisant. Ainsi, l'ACP permet de réduire la dimension du problème traité et d'en faciliter l'analyse, d'autant plus qu'elle met en évidence la corrélation entre descripteurs.

☞ Réaliser la classification des ligands

On utilise une deuxième fois la méthode exposée plus haut.

Résultats

Analyse univariée

La figure 1 présente les hiérarchies de partitions obtenues avec le descripteur 9 ou le descripteur 26.

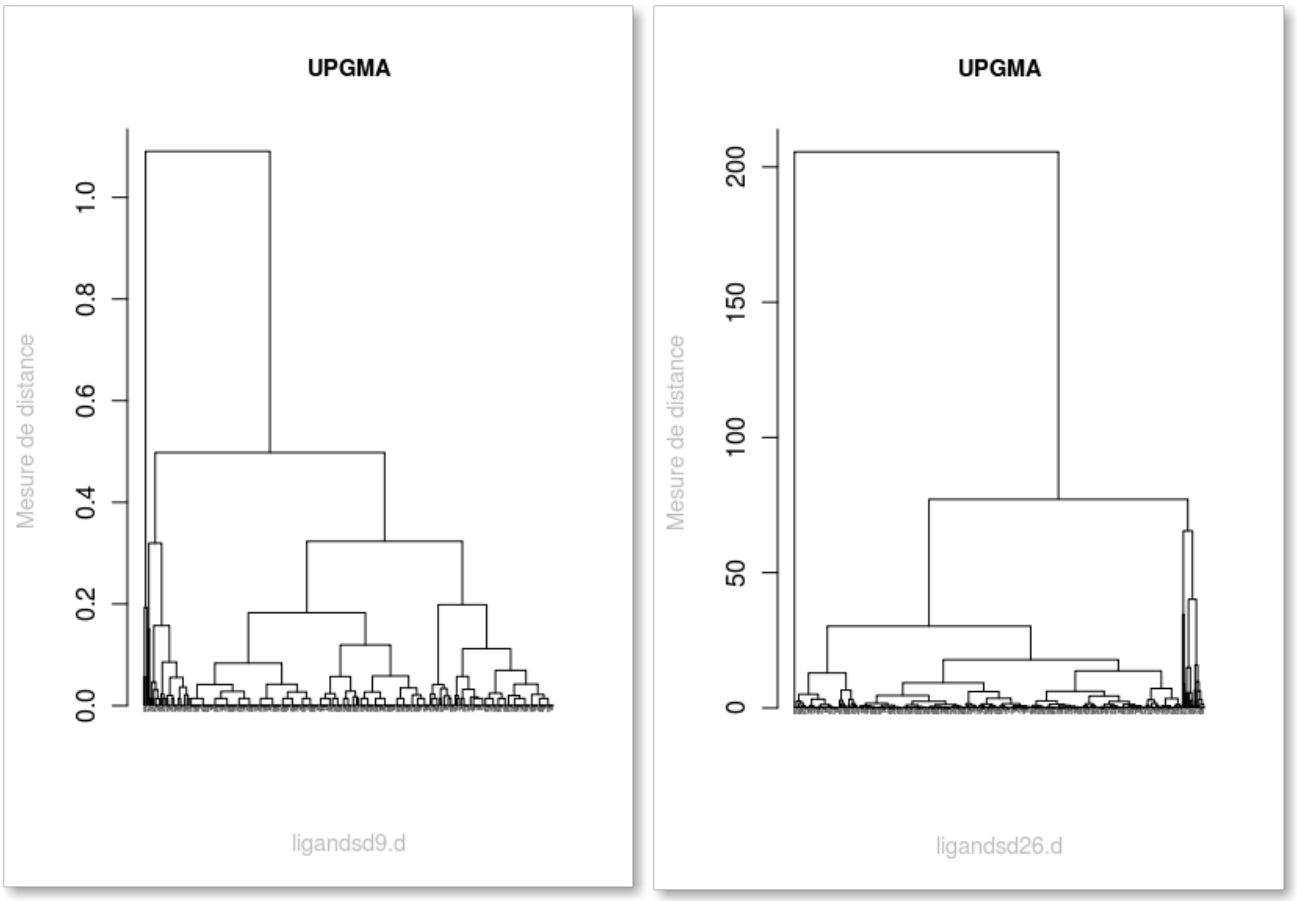


Figure 1 : structures de la population de ligands obtenues par classification hiérarchique ascendante (CAH) par lien moyen sur le descripteur 9 (à gauche), ou bien sur le descripteur 26 (à droite).

Même en considérant les rotations possibles au niveau des nœuds, la structure des nœuds semble différente (surtout à l'inspection des distances).

Pour la classification selon le ligand 9, à la 406ème opération d'agglomération la distance est de 0,01414214. Cependant, il suffit de faire varier de manière infinitésimale la distance à laquelle on coupe l'arbre pour obtenir une grande variation du nombre de classes de ligands (entre 48 et 80, la distance associée est toujours 0.01414214). Le nombre de classes n'est pas robuste.

Cela confirme l'impression donnée par l'existence des deux arbres de topologies différentes.

Pour la classification selon le ligand 26, la distance à la 406ème opération est de 0,877. Pour regrouper les ligands en 72 classes, je coupe l'arbre à une distance de 0,878. La robustesse du

nombre de classes est encore une fois limitée (la distance varie peu entre la 406ème et la 410ème opération – de 0,877 à 0,905).

La figure 2 représente la répartition des ligands en 72 classes.

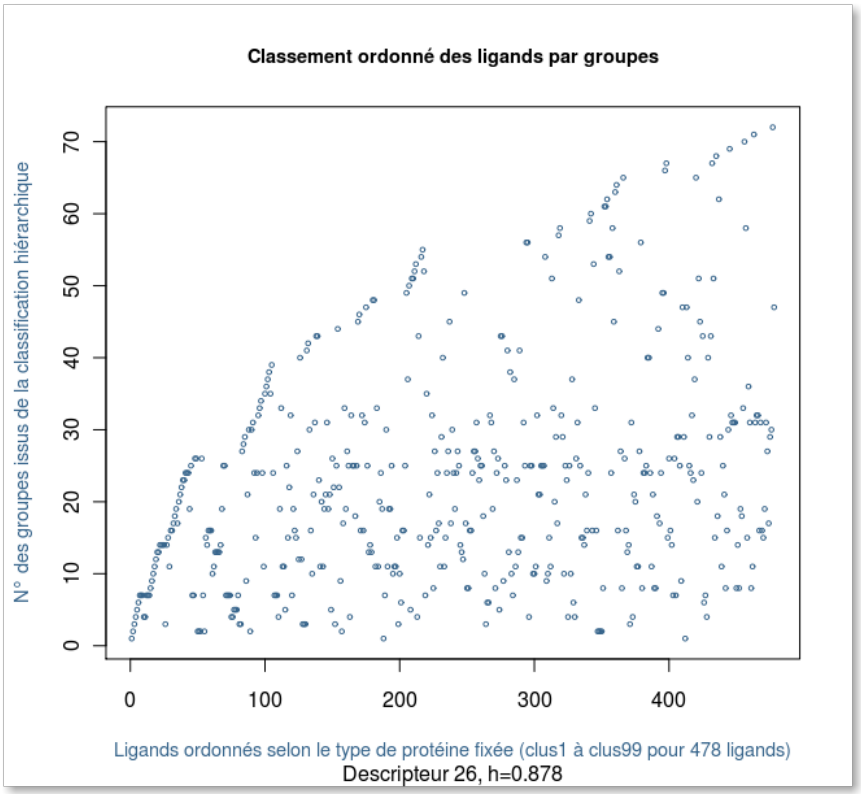


Figure 2 : classement des ligands en 72 groupes selon le descripteur 26. En abscisse, les ligands sont ordonnés selon le type de protéine fixée.

1/ Quelques ensembles de ligands alignés horizontalement dans la même classe montrent qu'il est possible de trouver une structure à la population de ligand sur la base du seul descripteur utilisé. (même résultat pour la classification sur le descripteur 9, non montré).

2/ Sur la base d'un seul descripteur, on ne peut retrouver de manière évidente la relation ligands-protéine. Si cela avait été le cas, on aurait eu le résultat présenté figure 3 où tous les ligands s'associant à une même protéine seraient alignés dans le même groupe. (même résultat pour la classification sur le descripteur 9, non montré).

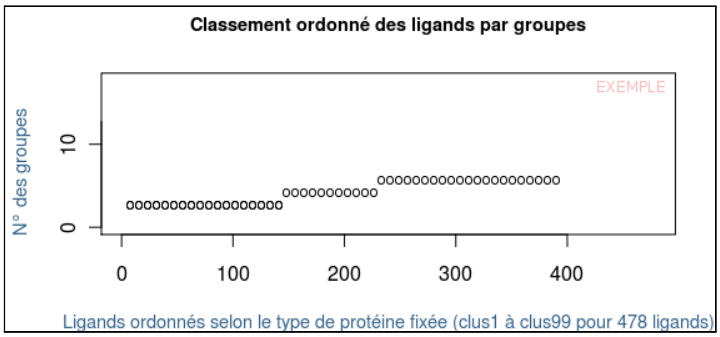


Figure 3: classement hypothétique des ligands en 72 groupes équivalents aux groupes basés sur la relation ligands-protéine

Ces résultats sont-ils dus à la prise d'un seul critère à chaque fois, quand d'autres parmi 31 devraient aussi être pris en compte ?

Analyse multivariée

Analyse en composantes principales

Le calcul des valeurs propres à partir des écarts-types montre une décroissance relativement lente du pouvoir explicatif cumulé de chaque composante. Cependant, les deux premières composantes ont une contribution cumulée importante. Elle s'élève à 45 % à la corrélation des données (tableau 1).

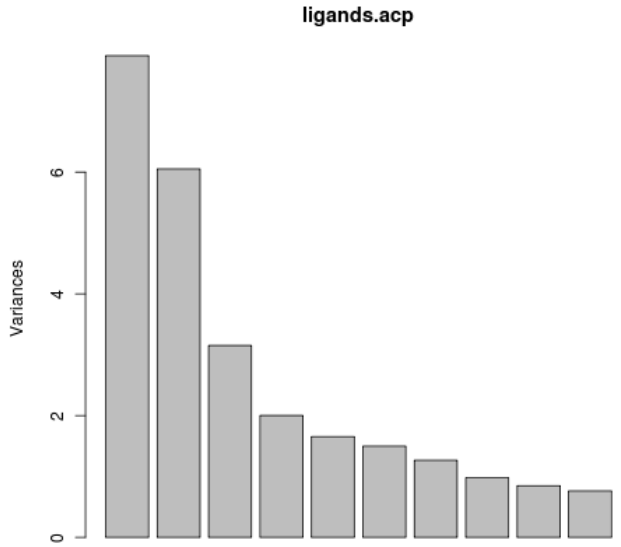


Figure 4 : représentation graphique des écarts-types (qui sont les racines carrées des valeurs propres) en fonction des composantes principales.

Visuellement, les deux ou les trois premières composantes principales sont significatives (figure 4). On a représenté en figure 5 la projection du nuage de points dans le plan des deux premières composantes principales.

Tableau 1

Composante principale	1	2	3	4	5
SD	2,81	2,46	1,78	1,41	1,29
Var	7,912	6,051	3,152	2,002	1,655
Pouvoir explicatif	25,5%	19,5%	10,2%	6,5%	5,3%
PE cumulé	25,5%	45,0%	55,2%	61,7%	67,0%

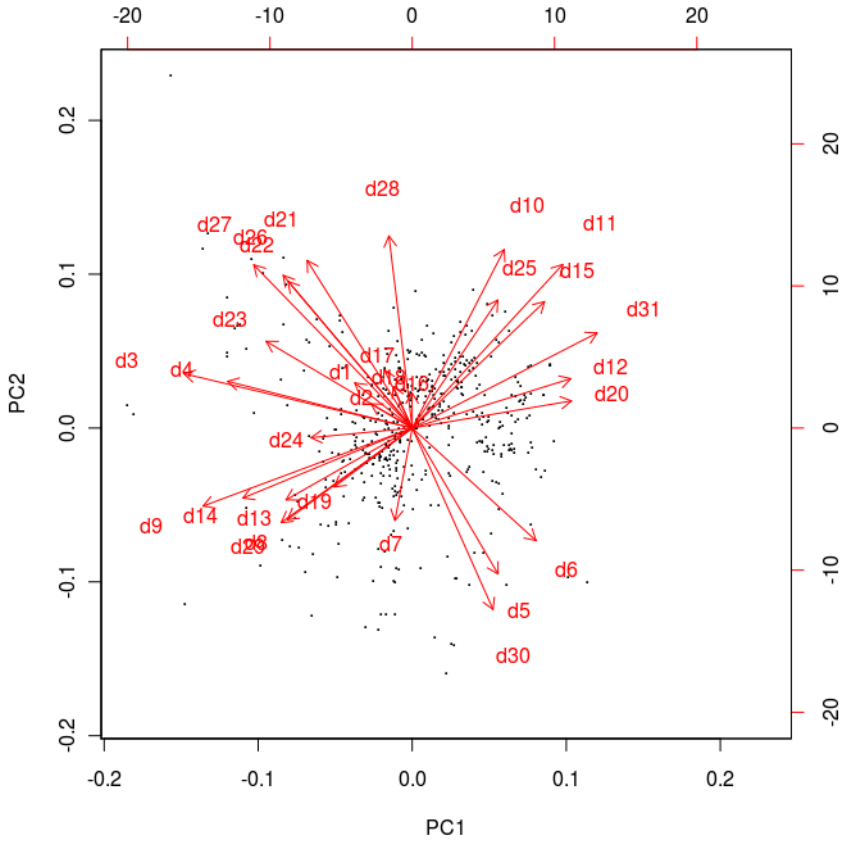


Figure 5 : représentation du nuage de points et des axes initiaux dans les deux premières composantes principales. La proximité des axes rouges des variables initiales traduit le degré de corrélation qui existe entre ces variables. Des axes de même direction indiquent des variables très corrélées. Des axes perpendiculaires indiquent des variables indépendantes.

1/ Les descripteurs 14, 13, 19, 29, 25, 11, 15, 31, 12, 20, 10 sont assez corrélés (premier groupe de corrélation). D'autre part, les descripteurs 1, 2, 17, 18, 16, 27, 22, 26, 21, 23, 4, 28, 7, 6, 5, 30 semblent

également corrélés (deuxième groupe de corrélation).

2/ La meilleure variable explicative (PC1) est assez proche de l'axe formé par le premier groupe de corrélation. La seconde meilleure variable explicative (PC2) est proche de l'axe formé par le deuxième groupe de corrélation.

3/ Les descripteurs 1, 2, 16, 17, 18 contribuent peu à la deuxième composante principale.

3/ Ce mode de représentation des données ne permet pas de dégager des groupes de ligands. (478 points se superposent dans la représentation, empêchant toute lecture, et la décroissance en valeur propres n'est pas assez rapide pour voir des groupes avec deux composantes principales).

4/ On décide qu'une classification peut être menée en se restreignant à la prise en compte des descripteurs contribuant le plus à chacune des cinq premières composantes principales (67,0 % du pouvoir explicatif cumulé). En se basant sur le tableau 2, on conservera les descripteurs d3 (contributeur en CP1), d28 (CP2), d7 (CP3), d19 (CP4) et d21 (CP5).

Tableau 2 : vecteurs propres des 31 descripteurs pour les 5 premières composantes principales (contenu de la variable ligands.acp)

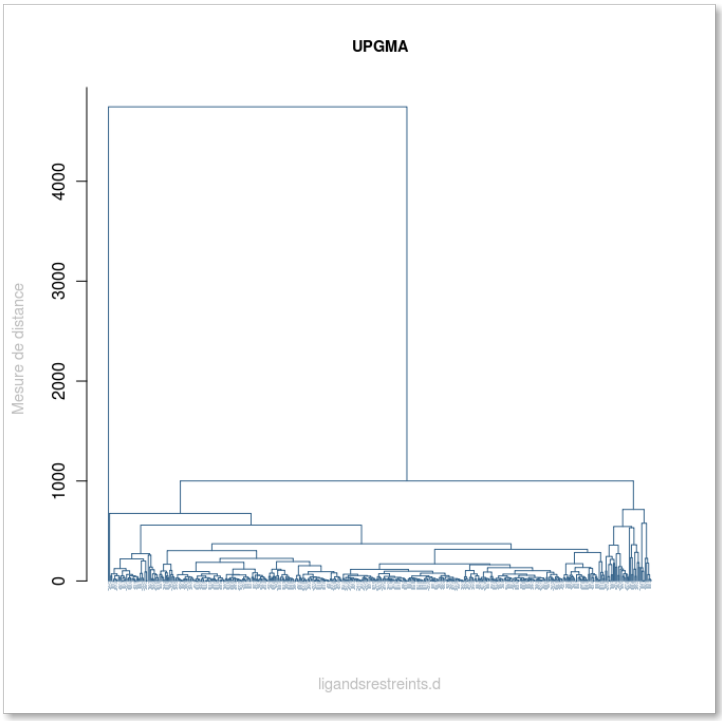
	PC1	PC2	PC3	PC4	PC5
d1	-0,08	0,07	-0,06	0,02	-0,31
d2	-0,06	0,04	0,09	0,00	-0,35
d3	-0,33	0,09	0,05	-0,07	-0,04
d4	-0,26	0,08	0,00	-0,04	-0,20
d5	0,12	-0,24	-0,07	-0,10	0,24
d6	0,18	-0,18	0,07	-0,24	0,05
d7	-0,02	-0,15	0,45	-0,19	-0,02
d8	-0,18	-0,15	0,39	-0,14	-0,08
d9	-0,30	-0,13	0,18	-0,05	-0,03
d10	0,13	0,29	-0,24	-0,09	0,10
d11	0,22	0,27	0,03	-0,20	0,01
d12	0,23	0,08	0,23	-0,20	-0,02
d13	-0,18	-0,12	-0,14	-0,37	0,15
d14	-0,24	-0,11	-0,21	-0,18	0,05
d15	0,19	0,21	0,09	-0,22	-0,05
d16	0,00	0,06	-0,01	0,23	-0,05
d17	-0,04	0,10	0,00	0,27	0,04
d18	-0,03	0,07	0,00	0,30	-0,14
d19	-0,11	-0,10	-0,25	-0,44	0,19
d20	0,23	0,04	0,23	0,01	-0,03
d21	-0,15	0,27	0,10	0,05	0,38
d22	-0,18	0,24	0,26	0,02	0,25
d23	-0,21	0,14	0,30	0,04	0,22
d24	-0,14	-0,02	-0,14	-0,21	-0,50
d25	0,12	0,21	0,18	-0,19	-0,19
d26	-0,18	0,25	-0,03	-0,04	0,04
d27	-0,23	0,27	-0,09	-0,06	0,02
d28	-0,03	0,31	0,00	-0,18	-0,10
d29	-0,19	-0,15	-0,15	0,16	0,08
d30	0,12	-0,30	0,16	0,04	0,03
d31	0,26	0,16	-0,15	-0,04	0,05

Classification des ligands

Nombre réduit de descripteurs

La distance associée à la 406ème itération est 80,679861. Je choisis donc initialement de couper l'arbre à une distance égale à 80,68.

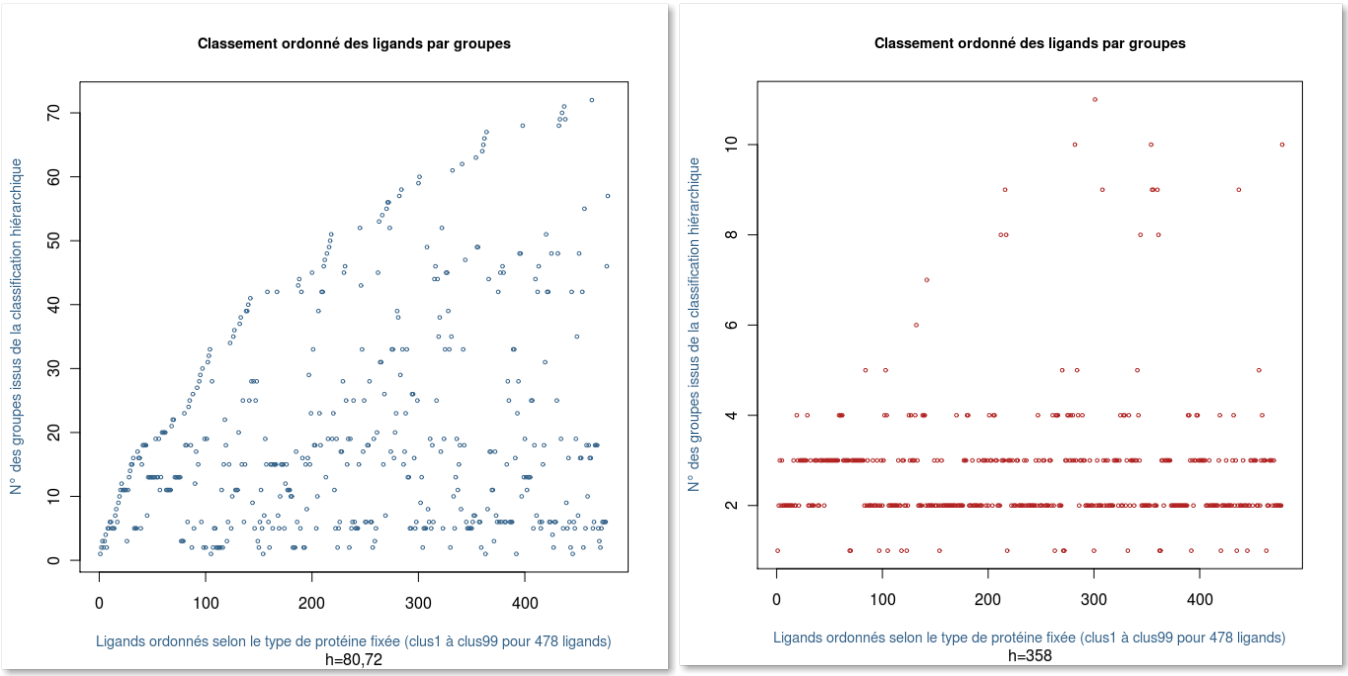
Figure 6 : partition obtenue par CAH par lien moyen de 5 descripteurs.



La figure 7-A représente la répartition des ligands selon les 72 classes (voir le fichier de résultat Cut_tree_height=80.csv).

Observations

Ici, ligne à ligne (classe par classe), on observe des regroupements (11 ou 12 assez nets et une importante dispersion) qui ne recoupent pas du tout les groupes basés sur le type de protéine auquel le ligand se fixe. (Bien que localement on puisse observer moins de 8 regroupement de ligands fixant les mêmes protéines alignés horizontalement les uns à côté des autres). Par ailleurs, de nombeuses classes n'ont qu'une seule entité.



A *B*

Figure 7 : A - classement des ligands en 72 groupes selon le descripteur 26. Partition coupée à la hauteur 80,72. B - partition coupée à la hauteur 358.

Cependant, rien ne dit que le nombre de groupe optimal sur la base des descripteurs soit de 72. Libéré de cette contrainte, nous avons cherché à réduire le nombre de classes pour voir si une classification pertinente pouvait être trouvée. Pour regrouper en moins de groupes (sans faire cette fois d'*a priori* sur le nombre de groupe), on a essayé d'avancer dans l'arbre à la recherche d'une grande variation de distance qui ferait peu varier le nombre de classes (zone de robustesse).

Je fais un saut important de 41,44 (changement de zone de robustesse) entre 12 (478-466) et 11 classes (478-467) – de 316.533819 à 357.975133. Choisir de couper l'arbre à une distance de 358 est donc une démarche raisonnée. Le résultat des regroupements peut être retrouvée dans le fichier Cut_tree_height=358.csv et dans la figure 7-B.

Observations (figure 7-B)

En regroupant les ligands en 11 classes, on observe que classer les ligands en catégories est pertinent puisque 8 lignes horizontales peuvent être facilement distingués (dont 3 groupes particulièrement visibles). Pour répondre à certaines questions (à définir avec l'expérimentateur[2]),

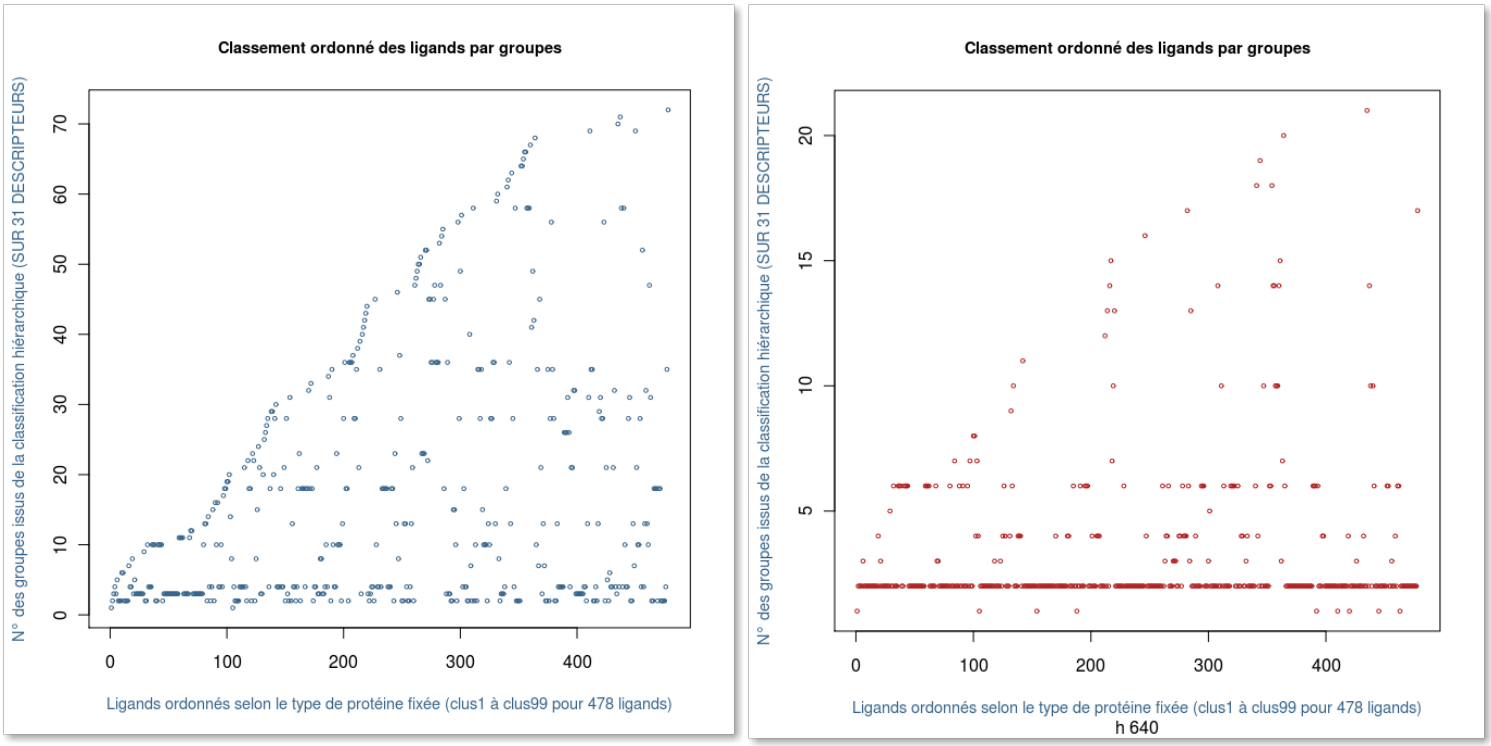
2 Les données de l'exercice ne précisent pas la nature des descripteurs et ne permet pas de prendre de décision sans discuter avec l'expérimentateur.

utiliser les classes basées sur les descripteurs des composantes principales pourrait rester utile, en se libérant de l'*a priori* qui est fait avec la relation ligand-protéine.

Pour tenter d'améliorer le résultat, nous avons utilisé l'ensemble des 31 descripteurs.

Complexité entière (31 descripteurs)

Pour obtenir 72 classes il faut couper l'arbre (figure 9) à une distance égale à 335 (figure 8-A).



A *B*

Figure 8 A - classement des ligands ordonnés selon la protéine associée (en abcisses), en fonction des classes issues de la partition hiérarchique (en ordonnées). Partition coupée à la hauteur 335. B - Partition coupée à la hauteur 640.

Observations (figure 8-A)

Même en prenant l'ensemble des descripteurs les 72 groupes décrits par l'association avec une protéine ne sont pas superposables avec les classes obtenues selon les descripteurs (bien que quelques îlots de ligands de même type soient alignés dans la même classe). Je ne change pas significativement le résultat obtenu avec l'analyse réduite aux composantes principales.

La démarche de classification a été poursuivie à la recherche d'une grande variation de distance qui ferait peu varier le nombre de classes. Pour passer de 22 classes (478-456) à 21classes (478-457), je fais un saut important de 36,455941. (changement de zone de robustesse) – de 603,4926 à 639,9486. Pour obtenir les regroupements de la figure 8-B, l'arbre a donc été coupé à une distance de 640.

Observations (figure 8-B)

Comme avec les données à complexité réduite, en regroupant les ligands en 21 classes, 8 lignes horizontales peuvent être facilement distingués (dont 3 groupes particulièrement visibles). Cette situation est similaire à celle menée avec les données à complexité réduites.

Conclusion

On a souvent été confronté à une faible robustesse du nombre de classe autour de la valeur recherchée (72), voir la figure 9.

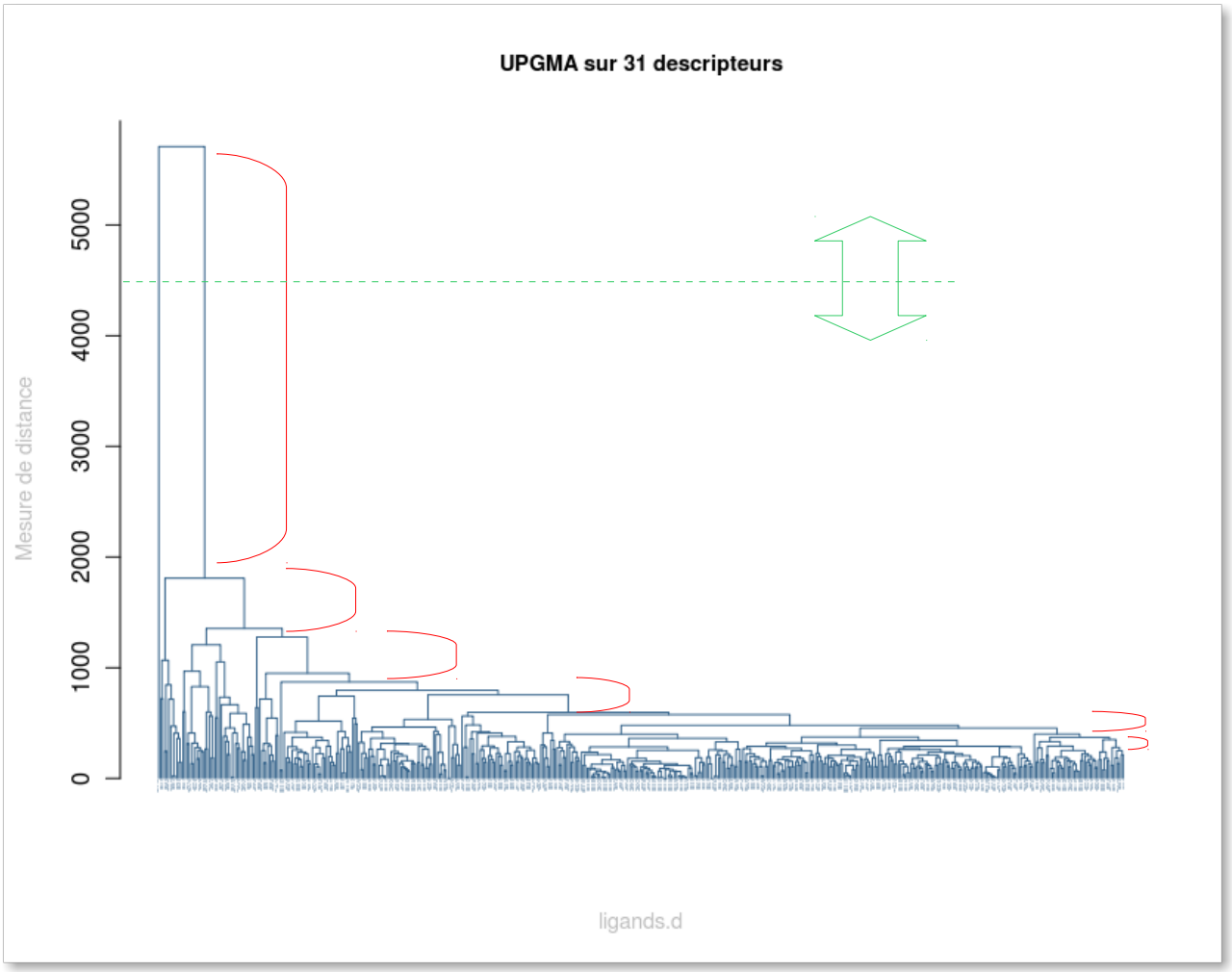


Figure 9 : CAH sur 31 descripteurs, les arcs rouges définissent les zones de variation de la distance sans changement significatif du nombre de classes (zones de robustesse). Au fur et à mesure que l'on progresse vers des groupes de plus en plus proches les uns des autres la taille des zones de robustesse diminue.

Avec la méthode de classification employée, nous n'avons pas retrouvé les associations ligand-protéines sur la base des descripteurs donnés. La classification basée sur les descripteurs ne peut donc pas être utilisée pour prédire les associations entre les protéines et les ligands. Cependant, en se basant sur un plus faible nombre de classes (dans des zones de robustesses plus grandes), une structuration des ligands a pu être détectée, même si elle n'a pas de lien avec les protéines. On pourrait rechercher un lien entre celle-ci et une information biologique.