Étude sur les cystéines modifiées du jeu de données bdd_models_byCys

1) Modification du jeu de données

J'ai choisi de travailler uniquement sur la composition en acides aminés de l'environnement à – de 10 Å de la cystéine étudiée.
On retient donc uniquement les descripteurs 23 à 202.
On modifie le jeu de données pour regrouper sous un même descripteurs les acides aminés d'un type sans prendre en compte la distance (les descripteurs ALA, A4, A5, A6, …, A11 sont additionnés et regroupés sous le descripteur A).
On transforme ensuite les valeurs absolues dénombrant le nombre d'acides aminés observés en fréquence.
On ajoute ensuite des descripteurs indiquant les propriétés des résidus suivants :

- aromatic : 'F', 'Y', 'H', 'W'
- polar : 'C','D','E','H','K','N','Q','R','S','T','W','Y'
- aliphatic : 'I', 'L', 'V'
- charged : 'D', 'E', 'R', 'K', 'H'
- negative : 'D', 'E'
- positive : 'H', 'K', 'R'
- hydrophobic : 'C','G','A','T','V','L','I','M','F','W','Y','H', 'K'
- small : 'C','V','T','G','A','S','D','N','P'
- tiny : 'A', 'C', 'G', 'S'

en fréquence également.

2) Utilisation de la méthode k-means et visualisation par cmdscale

On essaye de déterminer naïvement des groupes entre les cystéines étudiées à partir de la méthode k-means. On représente les individus en passant les données à 2 dimensions à partir de cmdscale. On choisit de calculer 4 groupes qu'on représente sur le graphique précédent (figure 1).

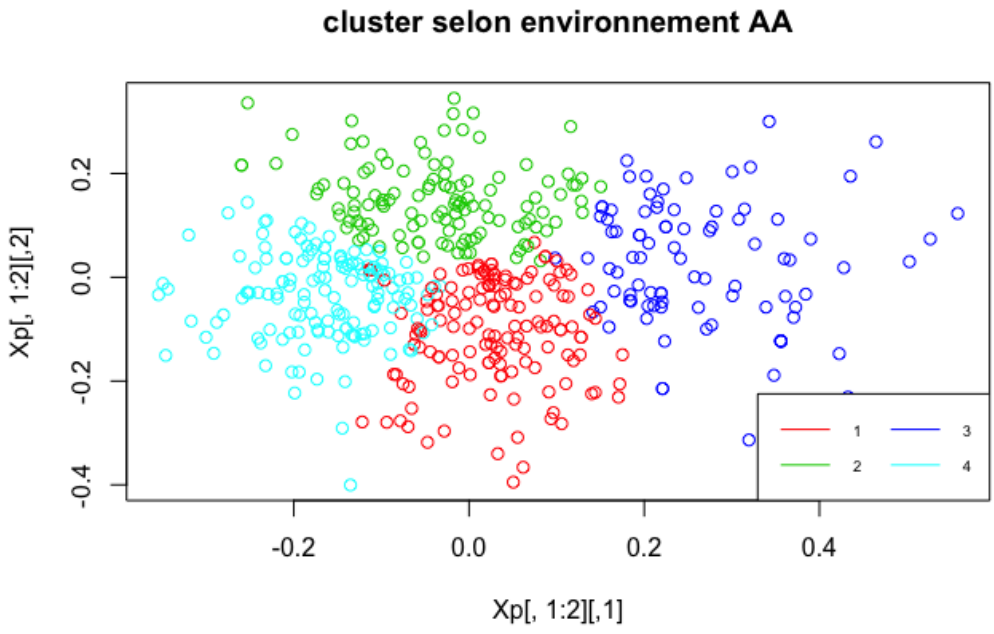

Figure 1 : Représentation des données avec cmdscale

3) Test anova et post-hoc

Dans tous les tests suivants, on choisit un risque $\alpha = 5\%$.

On cherche ensuite à identifier quels sont descripteurs différenciés entre les différents clusters calculés. On effectue un test anova pour chaque descripteur en fonction des clusters en posant pour chaque test les hypothèses suivantes :
H0 : Toutes les moyennes sont égales
H1 : Au moins l'une des moyennes est différente

On admet que les données suivent une loi normale.
On teste l'homoscédasticité avec un test de bartlett (H0 : l' homoscédasticité est respectée). Si la p-value est inférieure à 0.05, les conditions d'applications ne sont donc pas respectées et on effectue un test de kruskal wallis non paramétrique au lieu d'un test anova.

Les tests anova sont soulignés dans le tableau suivant.
On note 0 les p-value inférieures à 2.2×10^{-16}.

A	R	N	D	C	E	Q	G	H	I	L
6.6×10^{-6}	1.3×10^{-16}	0.04	1.0×10^{-10}	0.006	0	0	0	2.1×10^{-4}	3.1×10^{-4}	0.003

K	M	F	P	S	T	W	Y	V
1.7×10^{-7}	5.0×10^{-4}	0.1	0.002	0	0.883	0.677	0.835	0

aromatic	polar	aliphatic	charged	negative	positive	hydrophobic	small	tiny
2.4×10^{-4}	0	0	0	0	0	0	0	0

Les descripteurs pour lequel le test est significatif au risque $\alpha = 5\%$ sont indiqués en rouge.

Pour chaque descripteur dont le test est significatif, on effectue des tests post-hoc de comparaison de moyennes 2 à 2. Pour visualiser les résultats, on représente dans un tableau les moyennes par ordre croissant en indiquant le cluster correspondant associé à sa moyenne.

Moyenne croissante

1) A	2 0.074	3 0.080	1 0.101	4 0.128
R	4 0.028	1 0.039	2 0.049	3 0.070
N	2 0.031	4 0.033	1 0.042	3 0.043
D	4 0.039	2 0.041	1 0.046	3 0.085

5) C	2 0.019	1 0.020	3 0.021	4 0.028
E	4 0.025	1 0.037	2 0.042	3 0.069
Q	4 0.026	1 0.032	2 0.040	3 0.060
G	3 0.052	2 0.053	4 0.088	1 0.105
H	4 0.012	2 0.023	1 0.025	3 0.025
10) I	1 0.049	3 0.056	4 0.078	2 0.094
L	3 0.075	1 0.081	4 0.113	2 0.132
K	4 0.030	1 0.035	2 0.049	3 0.057
M	3 0.017	2 0.029	4 0.030	1 0.031
P	2 0.031	4 0.037	3 0.045	1 0.049
15) S	2 0.035	4 0.042	3 0.052	1 0.079
V	3 0.053	1 0.078	2 0.110	4 0.123
Aromatic	4 0.089	3 0.107	1 0.115	2 0.121
Polar	4 0.355	2 0.417	1 0.461	3 0.585
Aliphatic	3 0.184	1 0.207	4 0.320	2 0.337
20) Charged	4 0.133	1 0.182	2 0.205	3 0.307
Negative	4 0.063	1 0.082	2 0.084	3 0.155
Positive	4 0.070	1 0.099	2 0.121	3 0.152
Hydrophobic	3 0.575	1 0.676	2 0.731	4 0.771
Small	2 0.443	3 0.489	1 0.581	4 0.582
25) Tiny	2 0.181	3 0.206	4 0.285	1 0.306

Résultats des tests post-hoc entre les 4 clusters déterminés par k-means

= différence significative permettant d'établir des groupes au risque alpha 5%

On peut noter :

- Groupe 3 le plus chargé, groupe 4 le moins chargé
- Groupe 4 le + hydrophobique, groupe 3 le – hyddrophobique
- Groupe 4/1 plus petits résidus (small + tiny)
- Groupe 2/3 plus grands résidus (avec différences 2 < 3 pour freq. small)

4) Test du chi2 : la répartition des modifications est-elle la même selon les clusters déterminés par la méthode k-means

Test du chi2 d'homogénéité des populations au risque alpha = 5%.

H0 : la distribution des modifications est la même selon les clusters
H1 : la distribution des modifications n'est pas la même selon les clusters

NB : C.A. → nombre d'individus > 5
On regroupe NRG, NrG, nRG, nrG.

```
[1] "chi2 : modif~cluster"
> (Xsq <- chisq.test(M))  # Prints test summary

        Pearson's Chi-squared test

data:  M
X-squared = 8.3204, df = 9, p-value = 0.5022
```

p-value > 0.05, on ne rejette pas H0 → pas de différence dans la répartition des modifications même s'il y a des différences dans les propriétés des acides aminés de l'environnement des cystéines.

5) Test du chi2 : la répartition des plis de rossmann est-elle la même selon les clusters déterminés par la méthode k-means

Test du chi2 d'homogénéité des populations au risque alpha = 5%.

H0 : la distribution des plis de rossmann est la même selon les clusters
H1 : la distribution des plis de rossmann n'est pas la même selon les clusters

```
[1] "chi2 : Rossman~cluster"
> (Xsq <- chisq.test(M))  # Prints test summary

        Pearson's Chi-squared test

data:  M
X-squared = 43.271, df = 3, p-value = 2.155e-09
```

p-value<0.05, rejet de H0

Observation des résidus

group	modif 0	1
1	1.9745434	-2.6141249
2	-0.4331559	0.5734610
3	1.7154404	-2.2710949
4	-2.9479139	3.9027833

Plis de Rossmann en proportion plus importantes dans les groupes 2 et 4, moins présents dans les groupes 1 et 3.

Représentation des cystéines
→ Environnement différent en fonction des cystéines, + ou – chargés, + ou – hydrophobes, mais même types de modifications
→ (en cours) différences sur les autres descripteurs par rapport aux groupes déterminés avec environnement + compte-rendu sur étude précédente k-means selon tous les descripteurs normalisés
→ Plis de rossmann répartis différemment selon environnement : lié à fonction ou lien avec cystéine modifiée ?
→ + intéressant de représenter charges et hydrophobicité plutôt que surface ? Peut-être éléments observables visuellement