Université Paris Cité

Rapport
Traitement Numérique des Données

Yeast

Élèves :
Sepanta Farzollahi
Hagop Hannachian
Jean-Baptiste Hochet

Enseignante :
Nicoleta Rogovschi

8 mai 2024

Table des matières

1 Introduction

1.1 Contexte

La prédiction de la localisation cellulaire des protéines est une tâche essentielle en bioinformatique et en biologie moléculaire. Comprendre où une protéine est localisée dans une cellule est crucial pour déchiffrer ses fonctions biologiques et ses interactions au sein des organismes vivants. Cette information peut fournir des insights précieux pour la recherche sur les maladies, le développement de médicaments et la biologie fondamentale.

Le jeu de données utilisé dans ce projet provient du domaine de la bioinformatique et comprend des mesures expérimentales et des prédictions informatiques pour prédire la localisation subcellulaire des protéines chez la levure. Ces données sont extrêmement utiles pour développer des modèles de prédiction et des algorithmes d'apprentissage automatique afin de classifier les protéines en fonction de leur localisation.

L'objectif de ce projet est d'appliquer des méthodes enseignées dans le cours pour analyser ce jeu de données, de choisir et d'évaluer des approches adaptées, et enfin d'interpréter les résultats obtenus. Cette analyse permettra de mieux comprendre les caractéristiques des protéines qui déterminent leur localisation cellulaire, en utilisant des techniques statistiques et informatiques avancées.

1.2 Description du jeu de données

Le jeu de données **Yeast** utilisé pour ce projet concerne la prédiction de la localisation cellulaire des protéines, comprenant des informations sur différents attributs des protéines et leur localisation présumée. Voici un aperçu des principales caractéristiques du jeu de données :

- **Nombre d'instances** : Le jeu de données contient un total de 1484 instances, représentant des observations individuelles pour les protéines.
- **Nombre de variables** : Il comporte 10 variables, dont 8 sont de type continu décrivant des caractéristiques telles que les scores de reconnaissance des séquences de signal, la présence de motifs spécifiques, etc., et 2 sont des variables catégorielles.
- **Type de données** : Les variables continues fournissent des mesures numériques ou des scores pour chaque caractéristique des protéines, tandis que les variables catégorielles représentent des informations telles que l'ID et la localisation cellulaire prévue de la protéine.
- **Attribut cible** : La variable cible est `localization_site`, qui représente la localisation cellulaire prévue de la protéine. Il s'agit d'une variable catégorielle indiquant les différents sites de localisation, tels que `MIT` (mitochondrie), `NUC` (noyau), `CYT` (cytoplasme), etc.
- **Sources** : Le jeu de données est basé sur des études antérieures sur la prédiction de la localisation des protéines, telles que les références citées dans le jeu de données.
- **Prétraitement** : Aucune valeur manquante n'est signalée dans le jeu de données, ce qui simplifie le processus de prétraitement initial.

2 Analyse Exploratoire des Données

2.1 Description des variables

Les variables de **Yeast** fournissent des informations sur différentes caractéristiques des protéines, chacune étant essentielle pour la prédiction de leur localisation cellulaire. Voici une description détaillée des variables :

- `Sequence_Name` (ID) : Variable **catégorielle** représentant le numéro d'accession pour la base de données SWISS-PROT.
- `mcg` : Variable **continue** représentant le score obtenu par la méthode de McGeoch pour la reconnaissance de séquences de signal.
- `gvh` : Variable **continue** représentant le score obtenu par la méthode de von Heijne pour la reconnaissance de séquences de signal.
- `alm` : Variable **continue** représentant le score obtenu par le programme de prédiction de régions transmembranaires ALOM.
- `mit` : Variable **continue** représentant le score obtenu par l'analyse discriminante de la composition en acides aminés de la région N-terminale des protéines mitochondriales et non mitochondriales.
- `erl` : Variable **continue** indiquant la présence du sous-chaînon HDEL (censé agir comme un signal de rétention dans le réticulum endoplasmique). Il s'agit d'un attribut binaire.
- `pox` : Variable **continue** représentant le signal de ciblage peroxisomique dans la partie C-terminale des protéines.
- `vac` : Variable **continue** représentant le score obtenu par l'analyse discriminante de la composition en acides aminés des protéines vacuolaires et extracellulaires.
- `nuc` : Variable **continue** représentant le score obtenu par l'analyse discriminante des signaux de localisation nucléaire des protéines nucléaires et non nucléaires.
- `localization_site` (cible) : Variable **catégorielle** représentant la localisation cellulaire prévue de la protéine, avec 10 valeurs possibles : `CYT`, `NUC`, `MIT`, `ME3`, `ME2`, `ME1`, `EXC`, `VAC`, `POX`, `ERL`.

2.2 Visualisation des données

Les variables numériques (continues) extraites du jeu de données fournissent des mesures et des scores importants pour la prédiction de la localisation cellulaire des protéines. Chaque variable représente un aspect spécifique des protéines, allant des scores de reconnaissance de séquences de signal à la présence de motifs caractéristiques. Voici un aperçu des principales statistiques des variables numériques :

```
Sequence_Name           mcg              gvh              alm
Length:1484        Min.   :0.1100   Min.   :0.1300   Min.   :0.21
Class :character   1st Qu.:0.4100   1st Qu.:0.4200   1st Qu.:0.46
Mode  :character   Median :0.4900   Median :0.4900   Median :0.51
                   Mean   :0.5001   Mean   :0.4999   Mean   :0.50
                   3rd Qu.:0.5800   3rd Qu.:0.5700   3rd Qu.:0.55
                   Max.   :1.0000   Max.   :1.0000   Max.   :1.00
     mit              erl              pox              vac
Min.   :0.0000   Min.   :0.5000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.1700   1st Qu.:0.5000   1st Qu.:0.0000   1st Qu.:0.4800
Median :0.2200   Median :0.5000   Median :0.0000   Median :0.5100
Mean   :0.2612   Mean   :0.5047   Mean   :0.0075   Mean   :0.4999
3rd Qu.:0.3200   3rd Qu.:0.5000   3rd Qu.:0.0000   3rd Qu.:0.5300
Max.   :1.0000   Max.   :1.0000   Max.   :0.8300   Max.   :0.7300
     nuc         localization_site
Min.   :0.0000   Length:1484
1st Qu.:0.2200   Class :character
Median :0.2200   Mode  :character
Mean   :0.2762
3rd Qu.:0.3000
Max.   :1.0000
```

FIGURE 1 – Aperçu général des plages de valeurs et des tendances dans les variables

En examinant les caractéristiques numériques à travers des graphiques de boîtes à moustaches (boxplots), il est possible d'observer une grande hétérogénéité dans les valeurs pour la plupart des variables, à l'exception d'`erl` et `pox`. Cette hétérogénéité se manifeste par la présence de nombreuses valeurs aberrantes (outliers), qui sont situées au-delà des seuils minimum et maximum généralement observés pour ces variables.

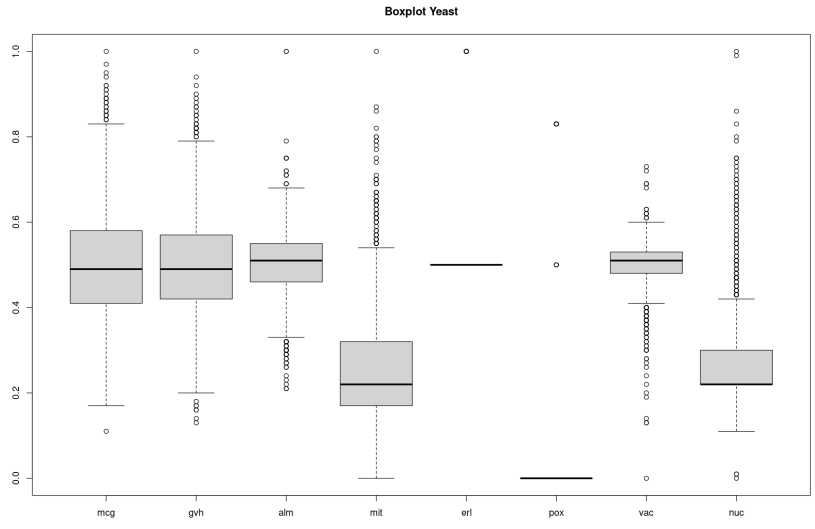

FIGURE 2 – Boîtes à moustaches (boxplots) des variables

Par exemple, pour la variable `mit`, la médiane est plus proche du premier quartile que du troisième quartile, ce qui suggère une distribution asymétrique à droite (right-skewed). Cela indique que la majorité des valeurs sont concentrées du côté gauche de la distribution, avec une étendue plus importante vers la droite.

En revanche, les variables `erl` et `pox` se distinguent par leur stabilité, affichant une faible variabilité et peu de valeurs aberrantes. La variable `erl` est quasi constante, avec un seul outlier, tandis que `pox` présente deux valeurs aberrantes.

Les observations sur `erl` et `pox` soulignent la nécessité de leur évaluation plus approfondie pour déterminer leur utilité dans la modélisation et l'analyse prédictive.

Pour analyser les relations linéaires entre les variables numériques de l'ensemble de données, un corrplot a été utilisé. Cette visualisation met en évidence les associations entre les caractéristiques des protéines étudiées, révélant les corrélations positives et négatives entre différentes mesures. Ces résultats aident à mieux comprendre les interactions entre les attributs des protéines avant d'approfondir l'analyse.

Les valeurs de corrélation indiquent la force et la direction des associations linéaires entre ces variables numériques. Une corrélation proche de 1 ou -1 indique une forte relation linéaire, tandis qu'une corrélation proche de 0 indique une faible relation linéaire.

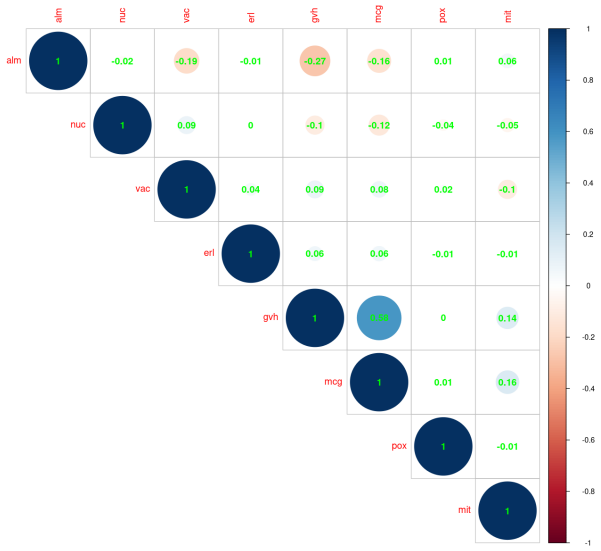

FIGURE 3 – Corrplot des variables

- `mcg` et `gvh` : Une corrélation **positive** forte de **0.58**, ce qui indique une relation significative entre elles. Lorsque la valeur de `mcg` augmente, la valeur de `gvh` tend également à augmenter.

- `alm` et `gvh` : Une corrélation **négative** modérée de **-0.27** suggère une relation inverse entre `alm` et `gvh`. Ainsi, lorsque la valeur de `alm` augmente, la valeur de `gvh` a tendance à diminuer.

- `mit` et `mcg` : Une corrélation **positive** faible de **0.16**, ce qui indique une relation moins prononcée entre ces deux variables.

- `erl` et `mcg` : Une corrélation **positive** très faible de **0.06**, ce qui suggère une relation presque inexistante entre ces deux variables.

- `pox` et `mcg` : Une corrélation **positive** extrêmement faible de **0.005**, indiquant une quasi-absence de relation linéaire entre ces deux variables.
- `vac` et `mcg` : Une corrélation **positive** très faible de **0.075** suggère une relation minime entre ces deux variables.
- `nuc` et `mcg` : Une corrélation **négative** faible de **-0.12**, ce qui suggère une relation inverse relativement faible entre ces deux variables.

La distribution des variables du jeu de données a été explorée en calculant leur skewness, permettant de comprendre les niveaux d'asymétrie présents dans leurs distributions. La skewness mesure la symétrie d'une distribution : une skewness positive indique une asymétrie à droite, tandis qu'une skewness négative indique une asymétrie à gauche.

En examinant les résultats, divers niveaux d'asymétrie sont observés :

- `mcg` : Légère asymétrie à droite (skewness ≈ 0.60).
- `gvh` : Légère asymétrie à droite (skewness ≈ 0.42).
- `alm` : Légère asymétrie à gauche (skewness ≈ -0.22).
- `mit` : Forte asymétrie à droite (skewness ≈ 1.44).
- `erl` : Très forte asymétrie à droite (skewness ≈ 10.14).
- `pox` : Très forte asymétrie à droite (skewness ≈ 10.26).
- `vac` : Forte asymétrie à gauche (skewness ≈ -1.79).
- `nuc` : Forte asymétrie à droite (skewness ≈ 2.41).

Ces valeurs de skewness fournissent des informations sur la forme des distributions des variables. Les histogrammes correspondants illustrent également la réparttion des valeurs. Par exemple, comme évoqué précédemment, pour les variables `erl` et `pox`, la quasi-totalité des valeurs sont identiques, à l'exception de quelques valeurs aberrantes.

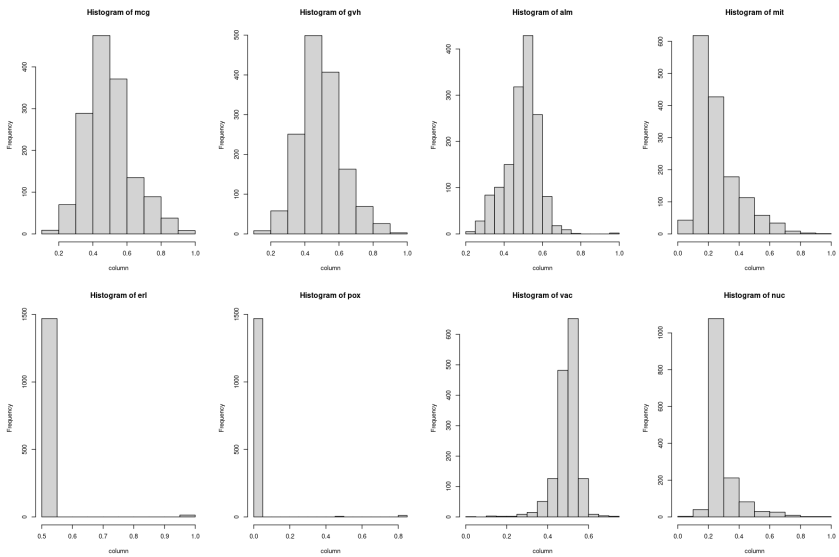

FIGURE 4 – Histogrammes des variables

Les diagrammes de dispersion (scatterplots) ont été utilisés pour explorer les relations entre les paires de variables (scatterplot matrix). Ces graphiques permettent de visualiser les associations potentielles entre les caractéristiques numériques du jeu de données.

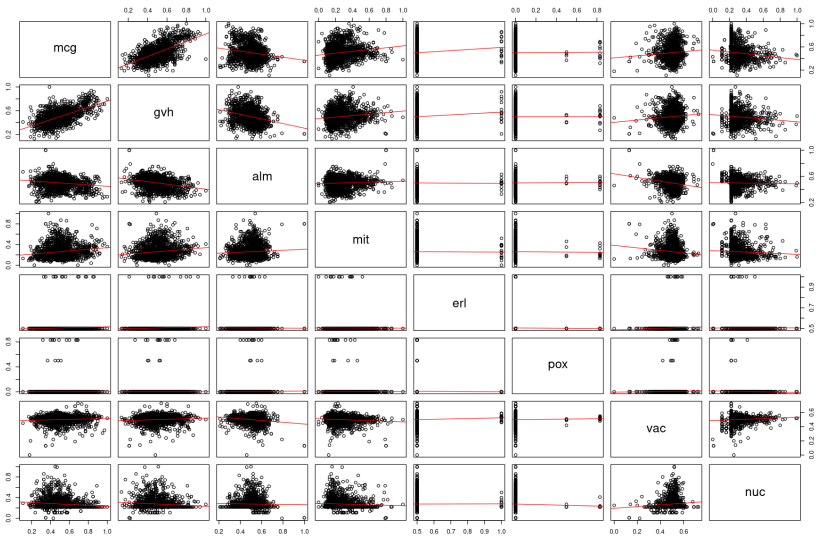

FIGURE 5 – Matrices de scatterplots des variables

Dans l'analyse des scatterplots, une relation positive entre `mcg` et `gvh` est observée, où les valeurs semblent augmenter conjointement, suggérant une corrélation positive entre ces deux variables. Cette tendance est illustrée par une courbe ascendante qui se forme à mesure que les valeurs de `mcg` augmentent par rapport à `gvh`.

Pour les autres paires de variables, comme `alm` et `gvh`, `mit` et `mcg`, ainsi que `erl` et `mcg`, les scatterplots ne révèlent pas de motifs évidents ou de relations claires. Les points semblent dispersés aléatoirement sans former de schémas distincts, ce qui indique une faible corrélation linéaire entre ces variables.

3 Analyse en Composantes Principales (ACP)

L'Analyse en Composantes Principales (ACP) a été appliquée pour explorer la structure sous-jacente des données et identifier les principales dimensions qui capturent la variance dans le jeu de données. Cette analyse multivariée permet de réduire la dimensionnalité des données tout en conservant autant d'informations que possible.

3.1 Description de l'ACP

L'ACP a généré six dimensions principales (Dim.1 à Dim.6), correspondant au nombre de variables dans le jeu de données. Chaque dimension principale explique une part de la variance totale dans les données. Par exemple, la première composante principale (Dim.1) explique **30%** de la variance totale, tandis que la deuxième composante principale (Dim.2) explique **21%** de la variance totale.

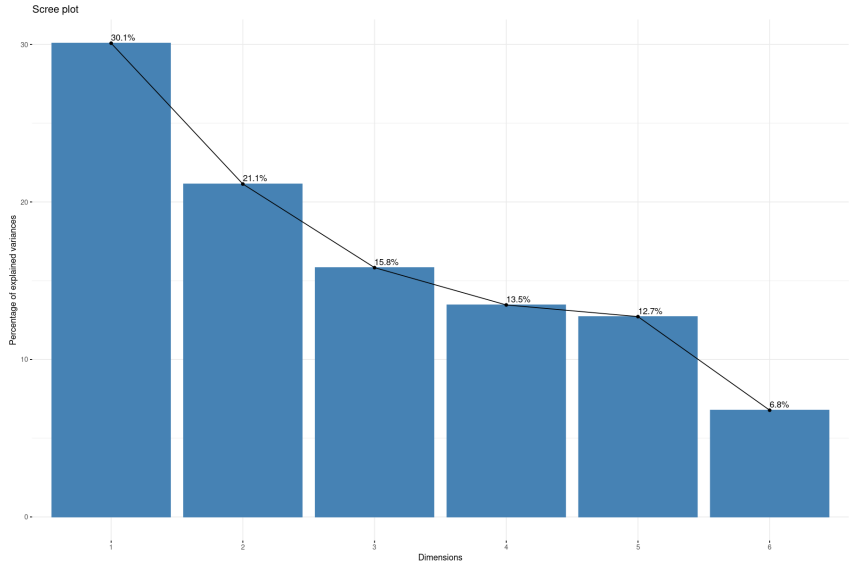

FIGURE 6 – Valeurs propres des variables

La proportion cumulative des deux premières dimensions principales (Dim.1 et Dim.2) explique près de **51%** de la variance totale, ce qui suggère que plus de la moitié de la variation dans les données est capturée par ces deux dimensions.

Des variables supplémentaires telles que `erl` et `pox` ont été incluses dans l'analyse, bien qu'elles n'aient pas été utilisées pour générer les composantes principales. Leur projection dans l'espace des dimensions principales offre des insights sur leur association avec les variables principales du jeu de données.

De même, des catégories supplémentaires ont été projetées dans l'espace des dimensions principales pour évaluer leur contribution à la variance totale et leur relation avec les dimensions principales.

3.2 Interprétation des résultats

Les résultats complets de l'analyse, y compris le résumé des statistiques des composantes principales, sont disponibles dans le fichier PDF du notebook pour une consultation détaillée.

En observant les coordonnées des individus dans l'espace des dimensions principales, il apparaît que les positions relatives des individus révèlent une dispersion significative. Certains individus sont fortement associés à une dimension particulière, ce qui indique des caractéristiques distinctives par rapport aux variables analysées. Par exemple, certains individus peuvent être positivement corrélés avec Dim.1 et négativement corrélés avec Dim.2, ce qui suggère des profils spécifiques parmi les variables étudiées.

Cette dispersion met en évidence la diversité des profils au sein de l'échantillon, reflétant les différentes combinaisons de valeurs pour les variables de l'ensemble de données. Certains individus peuvent être plus proches de l'origine des axes, indiquant une variabilité plus faible dans les caractéristiques qu'ils représentent, tandis que d'autres sont plus éloignés, suggérant une plus grande variabilité.

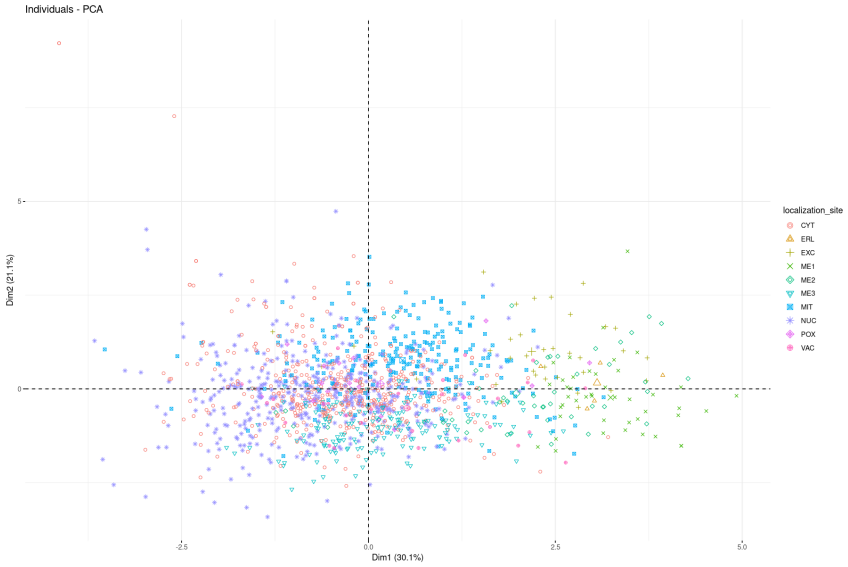

FIGURE 7 – ACP des individus

En examinant la position des individus dans l'espace des dimensions principales, on peut également identifier des clusters ou des regroupements d'individus ayant des profils similaires.

De même, les variables sont projetées dans l'espace des dimensions principales, ce qui permet d'évaluer leur contribution à chaque dimension. Par exemple, une forte corrélation positive entre une variable et une dimension principale indique une contribution significative de cette variable à la variance capturée par la dimension correspondante.

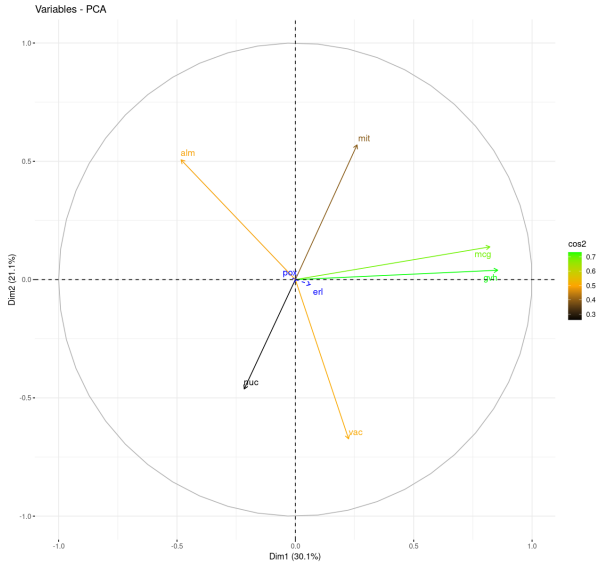

FIGURE 8 – ACP des variables

Dans l'interprétation de l'ACP des variables, il est remarquable que `mcg` et `gvh` présentent une forte corrélation positive, ce qui se traduit par leur magnitude élevée dans le biplot. De plus, une anti-corrélation est observée entre `mit` et `nuc`, ainsi qu'entre `alm` et

vac. Les variables négativement corrélées sont positionnées du côté gauche de l'origine du biplot, comme alm ou nuc.

Concernant la contribution des variables à la formation des axes, mcg et gvh contribuent au premier axe du côté positif, tandis que alm participe au premier axe du côté négatif.

Pour le deuxième axe, vac participe à sa formation du côté négatif, et mit est davantage impliqué dans la création de cet axe du côté positif par rapport à alm. Ces observations fournissent des informations précieuses sur la manière dont les variables sont positionnées et contribuent à la structure de l'espace des dimensions principales dans l'analyse en composantes principales.

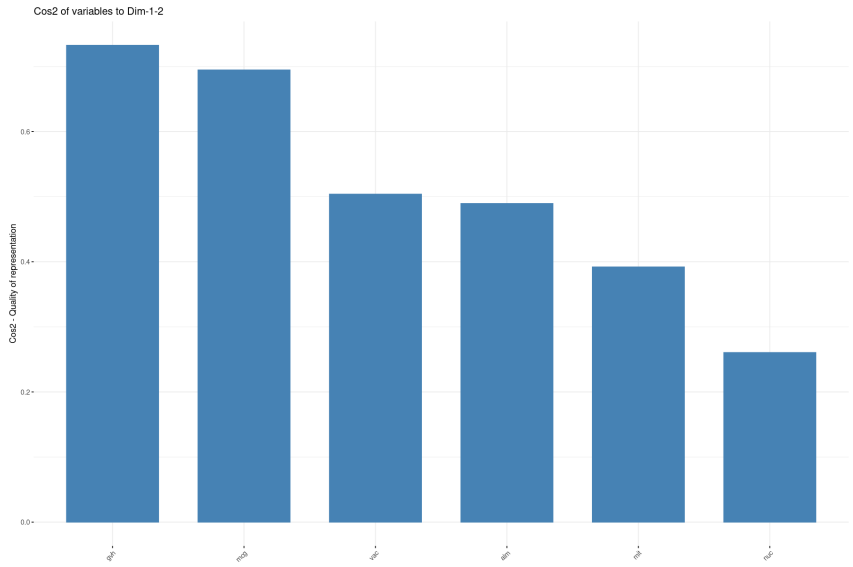

FIGURE 9 – Contributions des variables aux deux premiers axes principaux

gvh et mcg se distinguent par les plus forts $\cos^2$ parmi les variables, indiquant leur contribution significative à la variance expliquée par les premières composantes de l'ACP. Ces variables expliquent une part importante de la variance capturée par les premières dimensions principales. Les variables vac, alm, mit et nuc contribuent également à la variance expliquée par les premières composantes de l'ACP, bien que leur contribution soit moindre par rapport à gvh et mcg.

4 Clustering des Données

4.1 Méthode de Ward

Le dendrogramme, réalisé avec la méthode de Ward, offre une représentation visuelle de la structure hiérarchique des données. Cette méthode agglomérative vise à former des clusters homogènes en minimisant la variance intra-cluster.

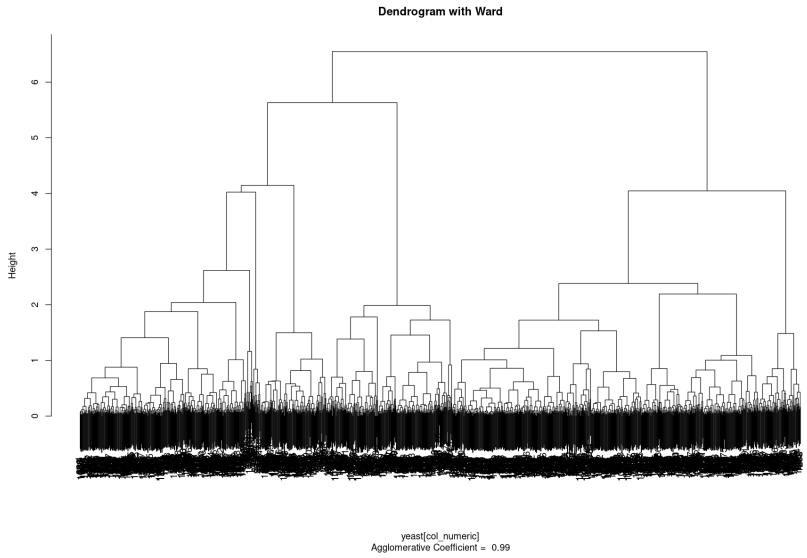

FIGURE 10 – Dendrogramme avec la méthode de Ward

Le coefficient d'agglomération élevé de 0.99 indique des regroupements solides et bien définis dans le dendrogramme, ce qui renforce la validité des clusters identifiés. Cette analyse est cruciale pour comprendre la distribution des données et guider le processus de clustering.

4.2 k-means

Pour l'analyse de clustering, une méthode non hiérarchique a été utilisée pour regrouper les observations en clusters de manière à minimiser la variance intra-cluster et maximiser la variance inter-cluster. L'algorithme k-means initialise les centroïdes de chaque cluster de façon aléatoire, puis attribue chaque observation au cluster le plus proche en fonction de la distance euclidienne. Les centroïdes sont ensuite mis à jour en calculant la moyenne des observations de chaque cluster, et ce processus itératif se répète jusqu'à atteindre la convergence.

Cette méthode est efficace pour identifier des groupes homogènes dans un espace multidimensionnel en fonction de leur similarité. Pour déterminer le nombre optimal de clusters, différents critères tels que la compacité intra-cluster et la séparation inter-cluster ont été évalués à l'aide d'indices de validation de cluster.

Le jeu de données **Yeast** contient **10** classes distinctes représentant différentes localisations cellulaires (la variable `localization_site`) : `CYT` (cytosolique ou cytosquelettique), `NUC` (nucléaire), `MIT` (mitochondriale), `ME3` (protéine membranaire sans signal N-terminal), `ME2` (protéine membranaire avec signal non clivé), `ME1` (protéine membranaire avec signal clivé), `EXC` (extracellulaire), `VAC` (vacuolaire), `POX` (peroxisomale) et `ERL` (lumen du réticulum endoplasmique).

Pour déterminer le nombre optimal de clusters lors de l'analyse de clustering, plusieurs indices de validation ont été comparés : **Calinski-Harabasz**, **Davies-Bouldin**, **Silhouette** et **Dunn**. Chaque indice évalue la compacité intra-cluster et la séparation

inter-cluster pour différents nombres de clusters, visant ainsi à minimiser ou maximiser ces critères selon le contexte spécifique de chaque mesure.

- **Calinski-Harabasz** : Cet indice mesure la dispersion entre les clusters par rapport à la dispersion au sein des clusters. Un score plus élevé indique des clusters compacts et bien séparés, ce qui est souhaitable pour obtenir des groupes distincts. **Deux** clusters maximisaient cet indice ont été identifiés, ce qui suggère une forte séparation entre les deux groupes.

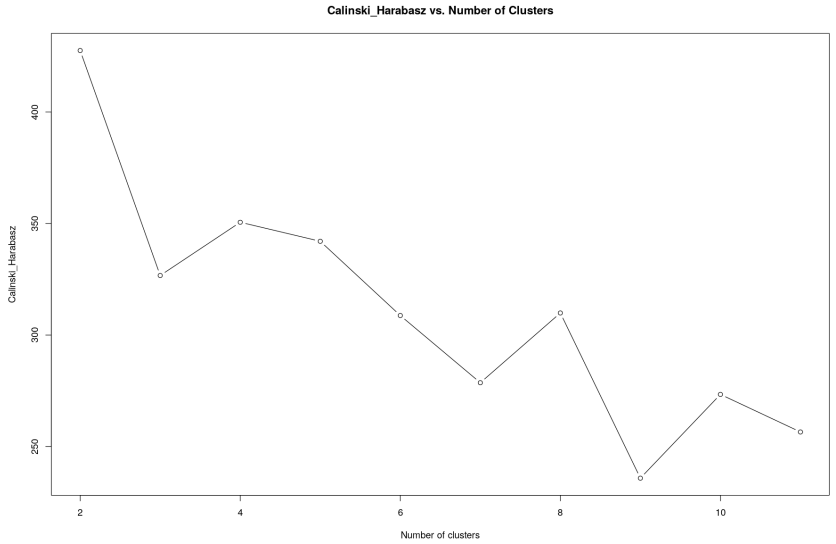

FIGURE 11 – Nombre de clusters en fonction de l'indice Calinski-Harabasz

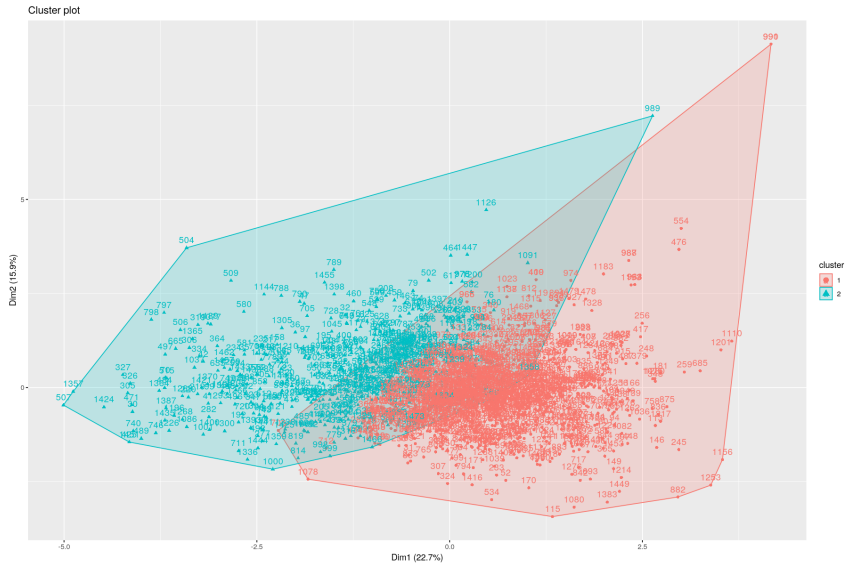

FIGURE 12 – Cluster plot pour l'indice Calinski-Harabasz maximal

- **Davies-Bouldin** : Cet indice évalue la similarité moyenne entre chaque cluster et son cluster le plus proche. Un score plus faible indique des clusters plus compacts et mieux séparés, ce qui est également un critère de bonne performance pour le

clustering. Selon cet indice, **dix** clusters étaient optimaux, correspondant au nombre initial de classes dans le jeu de données.

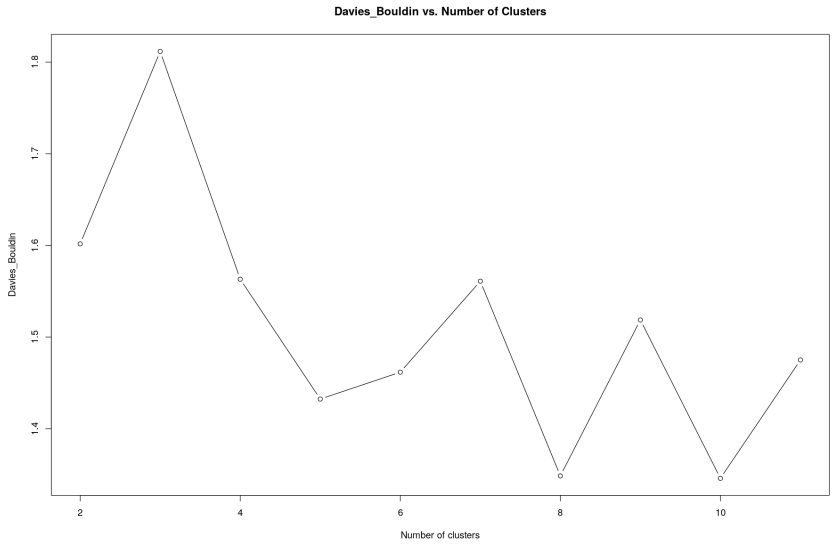

FIGURE 13 – Nombre de clusters en fonction de l'indice Davies-Bouldin

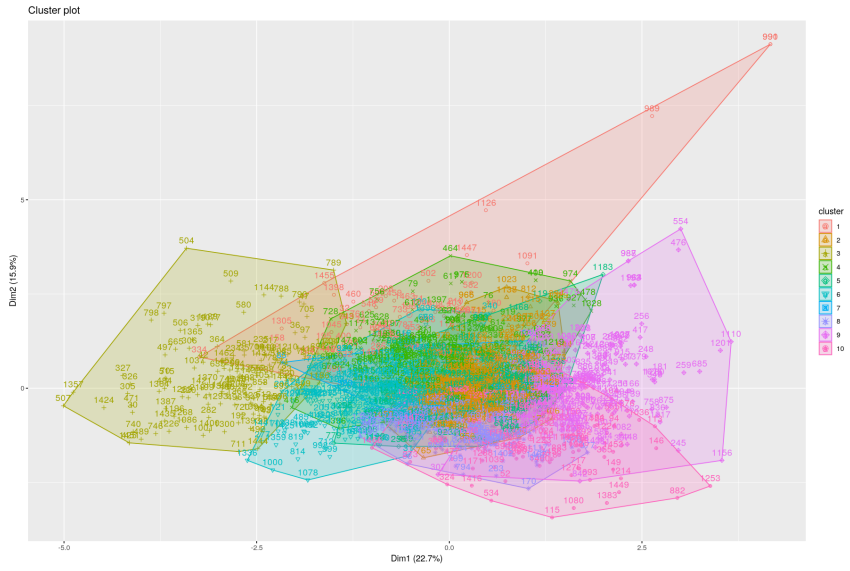

FIGURE 14 – Cluster plot en fonction de l'indice Davies-Bouldin minimal

- **Silhouette** : Cet indice mesure à quel point chaque point de données est proche de son propre cluster par rapport aux clusters voisins. Un score plus élevé indique des clusters mieux définis et séparés, ce qui reflète une bonne structure de clustering. Les résultats ont montré que **deux** clusters maximisaient cet indice, mettant en évidence une bonne séparation entre les groupes.

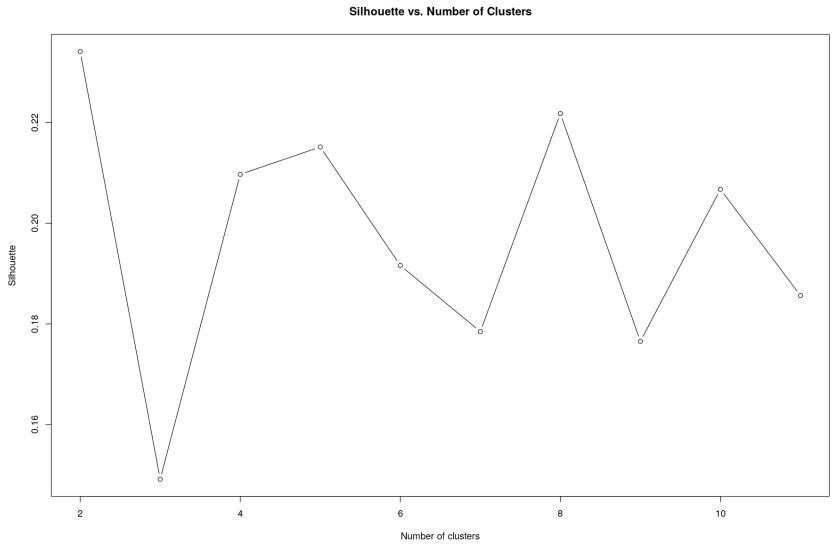

FIGURE 15 – Nombre de clusters en fonction de l'indice Silhouette

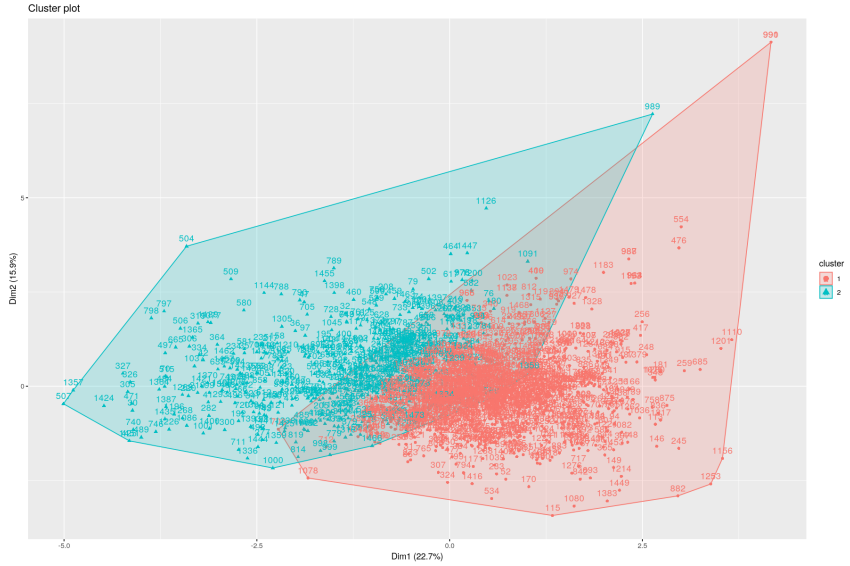

FIGURE 16 – Cluster plot en fonction de l'indice Silhouette maximal

- **Dunn** : Cet indice évalue la séparation entre les clusters. Un score plus élevé indique des clusters mieux définis et plus distincts, ce qui est bénéfique pour l'interprétation des clusters. Les résultats ont indiqué que **cinq** clusters optimisaient cet indice, ce qui suggère une séparation efficace des données en cinq groupes distincts.

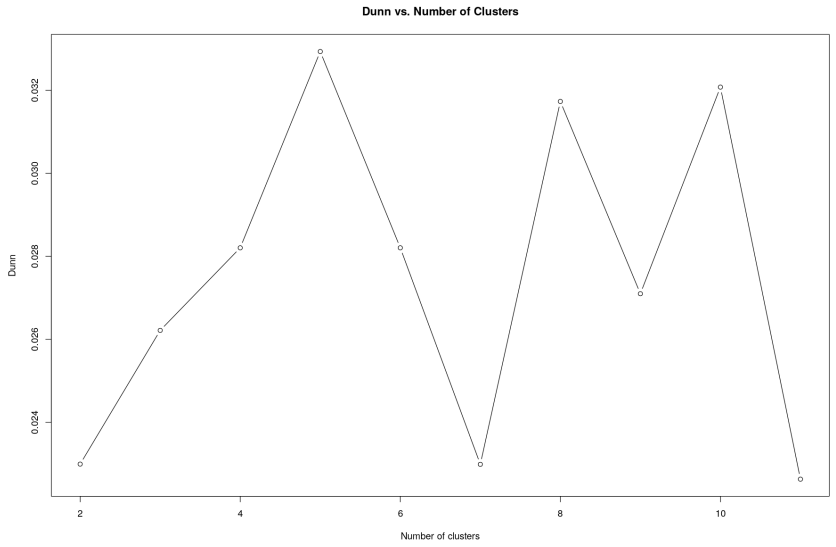

FIGURE 17 – Nombre de clusters en fonction de l'indice Dunn

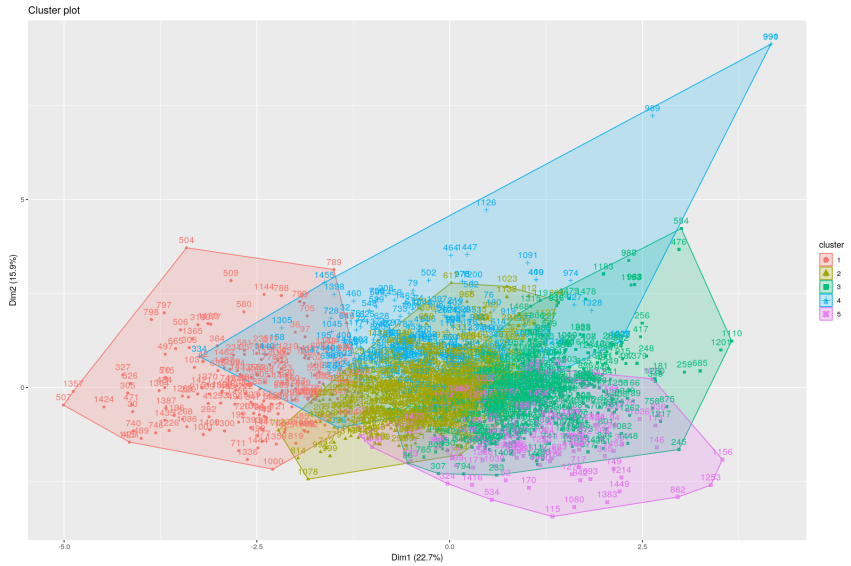

FIGURE 18 – Cluster plot en fonction de l'indice Dunn maximal

Les valeurs de Normalized Mutual Information (NMI) ont été également calculées pour chaque indice de validation de clustering, ce qui permet d'évaluer la similarité entre les clusters obtenus et les classes réelles du jeu de données. Le NMI mesure la qualité des clusters par rapport aux classes réelles, avec des valeurs plus élevées indiquant une meilleure correspondance entre les clusters et les classes.

- NMI pour **Calinski-Harabasz** ≈ 0.1
- NMI pour **Davies-Bouldin** ≈ 0.22
- NMI pour **Silhouette** ≈ 0.1
- NMI pour **Dunn** ≈ 0.19

Parmi ces indices, **Davies-Bouldin** a obtenu le meilleur NMI avec une valeur d'environ 0.22, indiquant une correspondance plus forte entre les clusters identifiés et les classes réelles du jeu de données.

Après l'analyse des résultats de clustering avec l'indice **Davies-Bouldin** qui a identifié **dix** clusters optimaux et obtenu un bon score de NMI d'environ 0.22, il est apparu que **Davies-Bouldin** était le meilleur choix parmi les indices évalués. Ce résultat suggère que les clusters générés par l'algorithme k-means, en utilisant **Davies-Bouldin** comme critère d'évaluation, correspondent mieux aux classes réelles du jeu de données que les autres indices. De plus, une **pureté** d'environ **0.35** a été trouvée pour les clusters identifiés, indiquant ainsi une bonne séparation et compacité des clusters selon cet indice de pureté.

5 Conclusion

Diverses techniques d'analyse de données ont été utilisées pour explorer les caractéristiques des observations. Une analyse descriptive des variables a mis en évidence des corrélations et des asymétries dans les distributions. Ensuite, l'Analyse en Composantes Principales (ACP) a été appliquée pour réduire la dimensionnalité et identifier les variables les plus influentes.

Par la suite, une analyse de clustering avec k-means a été réalisée pour regrouper les observations en clusters homogènes. En évaluant l'indice Davies-Bouldin, un nombre optimal de clusters a été identifié correspondant au nombre initial de classes. Cette approche a permis de mieux comprendre la structure des données et de les regrouper en fonction de leurs similarités.

Enfin, la qualité des clusters générés a été évaluée à l'aide de mesures comme l'indice de pureté, indiquant une bonne séparation et compacité des clusters. Ce projet fournit des informations clés sur la structure des données, facilitant ainsi la compréhension des relations entre les variables et des caractéristiques des observations.

6 Ressources

https://archive.ics.uci.edu/dataset/110/yeast

https://fr.wikipedia.org/wiki/Indice_de_Calinski-Harabasz

https://fr.wikipedia.org/wiki/Indice_de_Davies-Bouldin

https://fr.wikipedia.org/wiki/Silhouette_(clustering)

https://fr.wikipedia.org/wiki/Indice_de_Dunn