Université de Bordeaux

Data Mining : Recherche de protéines similaires.

Auteurs :
Franck Soubes
Guillamaury Debras
Tristan Frances
Mercia Ngoma Komb

Superviseurs :
Pascal Desbarats

12 janvier 2018

Table des matières

1 Introduction

Les protéines[1] sont impliquées dans le fonctionnement de toutes les cellules d'un organisme. Avec l'amélioration des techniques de séquençage et d'études protéomiques, nous avons de plus en plus de données a notre disposition pour l'étude de ces macromolécules essentielles a la vie. L'une des problématiques principales existantes est de savoir comment comparer ces protéines entre elles, et comment les organiser en différents groupes. Devrait t-on les grouper par similarité de séquence, de structure 2D, 3D, ou encore en fonction de certains critères structurels ou fonctionnels comme l'hydrophobicité ou le point isoélectrique ?

Dans le cadre de l'UE DEA, il nous a été demandé de réaliser un programme capable de trier des protéines parmi un jeu de donnée conséquent (B*ig Data*) selon le respect des techniques vu en cours et travaux dirigés. Le *data mining*[2] ou fouille de données est une technique qui permet d'extraire différentes informations ou motifs à partir d'une grande quantité de données selon plusieurs méthodes.

Ces techniques sont utilisées dans tous les types de domaines professionnels passant par la gestion de relation client ou l'optimisation site web, assurance et marketing. La fouille de données permet d'analyser et de trouver des d'autres informations intéressantes comme des pattern. La bioinformatique n'y échappe pas, permettant d'obtenir différents motifs biologiques à partir de différentes bases de données (protéines,gènes)

Le principe du *datamining* repose sur le fait qu'il n'y a pas besoin d'avoir d'hypothèse de départ. Cette méthode se base sur Le pré-traitement des données qui se déroule en plusieurs étapes, celles-ci sont :

La récupération afin d'extraire les données à partir de différentes banques de données.

La purification pour détecter des doublons ou erreurs éventuellement présents dans le jeu de données constitué et d'y remédier.

La transformation des données afin de convertir tout les formats présents en un seul et unique, pour une utilisation optimale.

L'étape suivante est la création de l'entrepôt de données (*Data-Wharehouse*) soit le chargement de toutes les données pour chaque organisme.

Une fois le traitement de ces étapes réalisé, on peut s'occuper de la partie création des groupes de données (*clusters*) contenant des protéines similaires suivant les critères que nous avons choisi.

1.1 Sélection des données

Afin de répondre à la question du sujet, nous avons en premier lieu déterminer les critères permettant de clusteriser[3] les protéines. Par la suite, nous avons sélectionné une liste des différents organismes sur lesquels notre classe pouvait travailler et ainsi récupérer les données qui les intéresse. Les données ont été recueillies sur un site web : Uniprot. et ont subit une curation via swiss-prot. Uniprot répertorie les protéines et leurs caractéristiques, et permet à n'importe quel utilisateur d'enregistrer les informations relatives à ces protéines selon plusieurs formats. Le format de fichier de sortie pour chacune de ses protéines est le format JSON. Ce choix a été fait puisqu'il permet de recueillir toutes les caractéristiques de la protéine et de les réutiliser très facilement par le langage Python.

L'ensemble des données récoltées formeront ainsi notre *data-warehouse*, celui-ci est formé de protéines contenues dans 14 organismes différents.

1.2 Tri des données

Les données sur lesquelles nous avons travaillées ont été choisies selon différents critères que nous avons nous mêmes choisis, en outre : la longueur de séquence, le pHi, l'arômaticité, le poids moléculaire, l'hydrophobicité et le nombre de structures secondaires comme le nombre d'hélices Alpha, de feuillets Béta et de coudes. Les données comme arômaticité, phi, poids moléculaire et hydrophobicité ne sont pas disponibles nativement dans le *data-warehouse*, de ce fait nous avons via l'implémentation d'un script python générer notre propre fichier JSON à l'aide de de modules déjà existants (Seq, IUPAC et ProteinAnalysis).

2 Création du *data-warehouse* personnel

2.1 Choix et récupération des données

Afin de répondre à la problématique, nous avons décidé de nous focaliser sur les protéines humaines ayant été publiées, et plus particulièrement à celles disponibles sur la base de données *uniprot*. Nous récupérons donc les données sous la forme d'un fichier JSON nous permettant de lire et de traiter facilement les données.

2.2 Préparation des données & représentation des données

La préparation des données est une étape indispensable car les données récupérées peuvent être incomplètes ou biaisées.
De plus, certaines protéines n'avaient pas de structure secondaire, nous avons décidé de ne pas les prendre et de calculer la fraction des acides aminées qui tendent à devenir des hélices, coudes ou feuillets conformément au module *ProtParam*.

Afin de préparer nos données nous avons écrit un script python qui permet de *parser* notre fichier JSON originel, de récupérer les infos que nous voulons (taille de la séquence, identifiant), calculer les informations dont nous avons besoin (aromaticité, phi, hydrophobicité..) et de les retourner sous un nouveau JSON. Pour la représentativité des données deux tour sur notre même jeu de données est effectué. Lors du premier tour nous mesurons la moyenne des différentes paramètres dans le *data-warehouse* via le module *numpy*. Le deuxième tour sert à créer notre nouveau JSON (*data base*) en ajoutant pour les données manquantes la valeur moyenne obtenu lors du premier tour de *parsing* pour les différents critères.

Nous obtenons ainsi un fichier JSON contenant uniquement les données que nous avons sélectionnées. Il contient 20237 protéines.

A partir de ce fichier et pour faciliter la création des *clusters*, nous avons décidé de créer une table.

La table structure comprend toutes les données récupérées et calculées.

Cette table est en réalité un dictionnaire de dictionnaire dans lesquels la clé est le numéro d'accession de la protéine. La valeur associée à la clé est un autre dictionnaire dans lequel les clés sont les noms des paramètres.

3 Critères choisis

Nous avons choisi 9 critères.

3.1 Structure primaire

3.1.1 Taille de la chaîne

Les protéines sont constituées d'acides aminés reliés entre eux par des liaisons peptidiques. Cette taille peut grandement varier selon les protéines. Après réalisation d'un histogramme de la taille des séquences, nous avons confirmer ce paramètre comme critère de clusterisation.

3.1.2 Poids moléculaire

Le poids moléculaire d'une protéine est la somme des poids des éléments la constituant, le poids peut être un critère intéressant pour une étape de clusterisation, de plus l'histogramme réalisé sur ce paramètre montre une bonne répartition des données.

3.2 Structure secondaire

Les structures secondaires définissent également les protéines (hélices alpha, feuillets beta et coude) qui ont une influence sur son repliement.

3.3 Structure tertiaire

La structure tertiaire fait appel à des interactions hydrophobes, des liaisons ioniques, des liaisons de Van der Walls et des liaisons covalentes (comme le pont disulfure) dans le cas d'une protéine de œ. Elle fait référence à l'organisation dans l'espace de ses structures secondaires.

3.3.1 Hydrophobicité

L'hydrophobicité d'une protéine est déterminée à partir du nombre d'acides aminés hydrophobes qu'elle possède. Selon l'index d'hydrophobicité (GRAVY : *Grand average of hydropathicity index, Kyte and Doolittle*) qui indique la solubilité d'une protéine. Le caractère hydrophobe ou hydrophile d'une protéine influence généralement sa localisation au niveau cellulaire ou encore son repliement, c'est pour cela que nous avons juger utile d'inclure ce critère dans notre clusterisation.

3.3.2 pHi

Le pHi ou point isoéléctrique d'une protéine, c'est à dire le pH auquel cette protéine a une charge nulle et donc limite son déplacement dans un champ électrique physiologique. Une protéine peut ainsi être chargée positivement ou négativement. Une protéine dans un milieu égal à son pHi a en général

— Une solubilité qui est minimale,

— Une mobilité qui lui permet de migrer plus vers un milieu qu'un autre,

3.3.3 Aromaticité

L'aromaticité est une propriété des structures moléculaires avec un composé cyclique qui est particulièrement stable. L'aromaticité est définie par la règle de Hückel. Le calcul se fait via la fréquence entre la phénylalanine la tryptophane et la tyrosine.

3.3.4 Cystéines

Deux cystéines peuvent générer des ponts disulfures créant des liaisons inter-chaînes. Ayant une influence sur la structure tridimensionnelle de la protéine, il est intéressant de garder ce paramètre.

4 Réalisation et Analyse des *Clusters*

4.1 Stratégie de clusterisation

Nous avons choisi de réaliser un *clustering* hiérarchique de type DIANA(*Divisive Analysis*)[4], c'est une approche *"top-down"*, l'algorithme est l'inverse de celui d'AGNES, il commence avec un seul *cluster* contenant tous les objets. A chaque itération le *cluster* le plus hétérogène est divisé en deux comme représenté sur la figure 1. La méthode utilisée est celle de Ward, car nous travaillons sur des données quantitatives non binaires. Cette méthode est basée sur la somme des carrés produisant des groupes qui minimisent la dispersion intra-groupe à chaque fusion.

Une fois la matrice de distance générée via la distance euclidienne, on génère nos *clusters* paramètre après paramètre avec la méthode de Ward.

Afin de déterminer l'ordre des paramètres à réaliser nous nous sommes appuyer sur une Analyse en Composantes Principales ou (ACP) sur les 20000 protéines et 9 paramètres représentés sur la figure 2. Les 2 premières dimensions expliquent à 50% la répartition des données, l'hydrophobicité, hélice alpha et aromaticité représentent la première dimension, ceci est logique du fait que les acides aminés aromatiques sont hydrophobes, ceci est souvent le cas des régions transmembranaires qui s'organisent en hélice alpha. La deuxième dimension représente la taille de la séquence et cystéine, ces 2 paramètres réunis ont du sens dans la mesure où plus la taille est élevée plus le nombre de cystéine peut statistiquement augmenté. Pour ce qui est du poids, feuillet bêta, coude et phi, ces 4 derniers paramètres ne semblent pas être influencés par ces 2 dimensions. Notre démarche d'ordre de clusterisation est la suivante : nous décidons d'utiliser les paramètres d'ordre primaire en premier (poids moléculaire suivi de la taille de séquence) puis les paramètres de structures secondaires (feuillet,coude, hélice) et enfin les paramètres liés aux structures tertiaires (aromaticité, hydrophobicité, cystéine et phi). L'ACP est possible car nous avons un grand nombres d'objets (20000protéines) bien supérieur aux 9 critères (mode direct) pour faire notre *clusterisation*. De plus une étape de normalisation des données est effectué au préalable via la méthode *scale* pour homogénéiser les données qui n'ont pas les mêmes unités de mesures (exemple Da pour le poids moléculaire contre nombre d'acides aminés pour la taille de la séquence).

4.2 Résultat d'analyse

Sur la figure 3 nous avons le résultat du dendrogramme pour l'ordre des paramètres décrit au préalable avec un résultat de 490 *clusters*. En ordonnée nous retrouvons la distance euclidienne qui sépare chaque découpage des *clusters*, et en abscisse le nombre de protéines contenues dans chaque *cluster*. Nous obtenons ainsi pour chaque suite de paramètres le nombre de *cluster* suivant :

- Longueur de la séquence : 2 *clusters*
- Poids Moléculaire : 4 *clusters*
- Feuillet : 8 *clusters*
- Coude : 16 *clusters*
- Hélice : 32 *clusters*
- aromaticité : 64 *clusters*
- hydrophobicité : 128 *clusters*

— pHi : 256 *clusters*
— cystéine : 490 *clusters*

L'algorithme peut parfois rencontrer des *clusters* vides, ceci explique le fait que nous n'avons pas 256*2 = 512 *clusters* pour le dernier paramètre. Après de multiples essais sur l'ordre de combinaison des paramètres, il s'avère que cet ordre est celui qui permet d'obtenir le plus de *clusters*.
Nous pourrions obtenir beaucoup plus de *clusters* en augmentant le nombre de paramètres, nous avons essayé avec l'indice d'instabilité qui pour une valeur supérieur à 40 est considéré comme instable et en dessous est considéré stable. Cependant ce paramètre ne peut pas être utilisé étant donné que via notre clusterisation, l'utilisation de la bibliothèque *scipy* ne nous permet pas de couper à cette valeur étant donné qu'il se base sur une matrice de distance euclidienne et non sur une valeur prédéfini.
Nous retrouvons à la fin de nos analyses des *clusters* allant de 1 protéine jusqu'à 80 protéines, ceci reste dans la logique de la méthode DIANA avec un jeu de données de 20000 protéines avec 9 paramètres.

Afin de visualiser la différence intra et inter-*cluster* nous avons décider de comparer les écart types et moyennes entre les protéines des différents *clusters*. Ceci est calculé via l'implémentation d'un script python qui prends en entrée le résultat texte obtenu après clusterisation et calcule la moyenne et écart type de chaque paramètre de chaque *cluster* et retourne les résultats des calculs sous format texte (res_intra.txt).

Après l'obtention de ces résultats, nous avons d'abord comparer les différentes moyennes et écarts-types, les unités n'étant pas les mêmes, et la présence d'une seule protéine dans certains *clusters* font que les valeurs sont à prendre en considération de manière relative. Cependant nous obtenons pour certains écarts-types de paramètres tels que les hélices, coudes, feuillets, et aromaticité des valeurs proches de 0.01.

Pour les paramètres taille de séquence et poids moléculaire la répartition des données étant vaste, leurs écarts-types associés augmentent également variant de 20 à 200 pour la taille et de plusieurs millions pour le poids moléculaire.

Lorsque l'on compare les différents paramètres pour les *clusters* 1 et 2 on retrouve des moyennes très similaires de même pour les *clusters* 464 et 465 avec pour exemple une taille de 81 et 166 acides aminés, une aromaticité de 0.21 et 0.16 pour un écart type de 0.05 et enfin, un point isoéléctrique de 8.27 et 10.51 pour les deux premiers *clusters* (figure annexe 4). En ce qui

concerne les *clusters* 489 et 490 (figure annexe 5) leurs tailles respectives sont de 1151 et 1164, l'aromaticité est de 0.084 et 0.084 et le point isoéléctrique de 5.581 et 7.512. et des écarts de moyennes conséquent entres les groupes de *clusters* 1,2 et 464, 465. On peut clairement observer une similarité de proche en proche et une dissimilarité lorsque les *clusters* sont éloignés entres eux. Cette observation conforte notre analyse concernant le bon *clustering* effectué par notre méthode.

Pour aller plus loin, on a regardé si les clusters pouvaient contenir via les paramètres choisis des fonctions similaires. Pour cela, on a regardé les protéines appartenant au cluster 1, et constater que ces 2 protéines sont impliquées dans la formation de la kératine : Q3LI58 et Q3LI59 qui correspondent respectivement aux protéines KRTAP21-2 et KRTAP21-1 (voir figure annexe 6). A la suite de cela on a voulu voir si les protéines des protéines des clusters proches à ce premier cluster avaient également des fonctions pouvant être similaires. Il s'est avéré que la majorité des protéines impliquées dans ce troisième cluster étaient aussi impliqué dans la formation de la kératine. Cette observation appuie le fait que notre démarche de clusterisation possède certaines qualités.

5 Conclusion

L'établissement des règles, la réflexion sur papier et la conception au moyen d'un langage informatique sont les trois étapes nécessaires pour aboutir à un logiciel fonctionnel. Le projet Datamining a eu un début de développement rapide mais l'écriture du programme afin de clusteriser le jeu de données contenant plus de 20 000 protéines a pris la majeur partie du temps, La vérification des similarités intra et inter *cluster* a été délicate du au fait que nous avions 9 paramètres avec des unités différentes. Par ailleurs, plusieurs méthodes sont disponibles cependant les résultats peuvent varier selon l'utilisation de ces différentes méthodes, ici nous avons opter pour la technique DIANA par la méthode de Ward avec la distance euclidienne et avons ainsi obtenu 490 *clusters*. Le code a été pensé et écrit en suivant les bonnes pratiques de programmation.

En ce qui concerne les modifications qui peuvent être apportés à notre code, de nombreuses implémentations sont possibles. Pour exemple nous pouvons citer l'ajout d'autres paramètres afin d'arriver à un nombre de *cluster*

plus grand pour une meilleur différenciation, l'utilisation de la méthode Agnès est aussi envisageable afin de comparer nos résultats de *clusters* et ainsi vérifier que nous avons bien les mêmes protéines dans certains *clusters* .
Ce projet nous a permis de réaliser une approche approfondie de la fouille de données qui risquent d'être de plus en plus présente dans tous les domaines que ce soit.
Enfin nous remercions Monsieur Pascal Desbarats pour ses connaissances apportées sur l'analyse et l'exploration de données.

6 Annexes

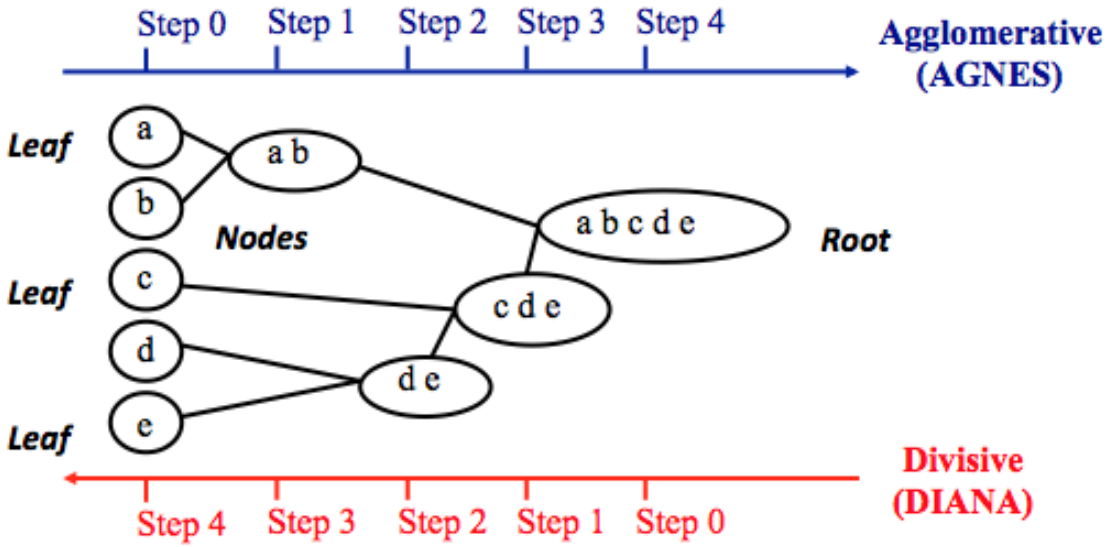

FIGURE 1 – Représentation graphique de la différenciation entre méthode AGNES et DIANA
source : http ://uc-r.github.io/hc_ clustering

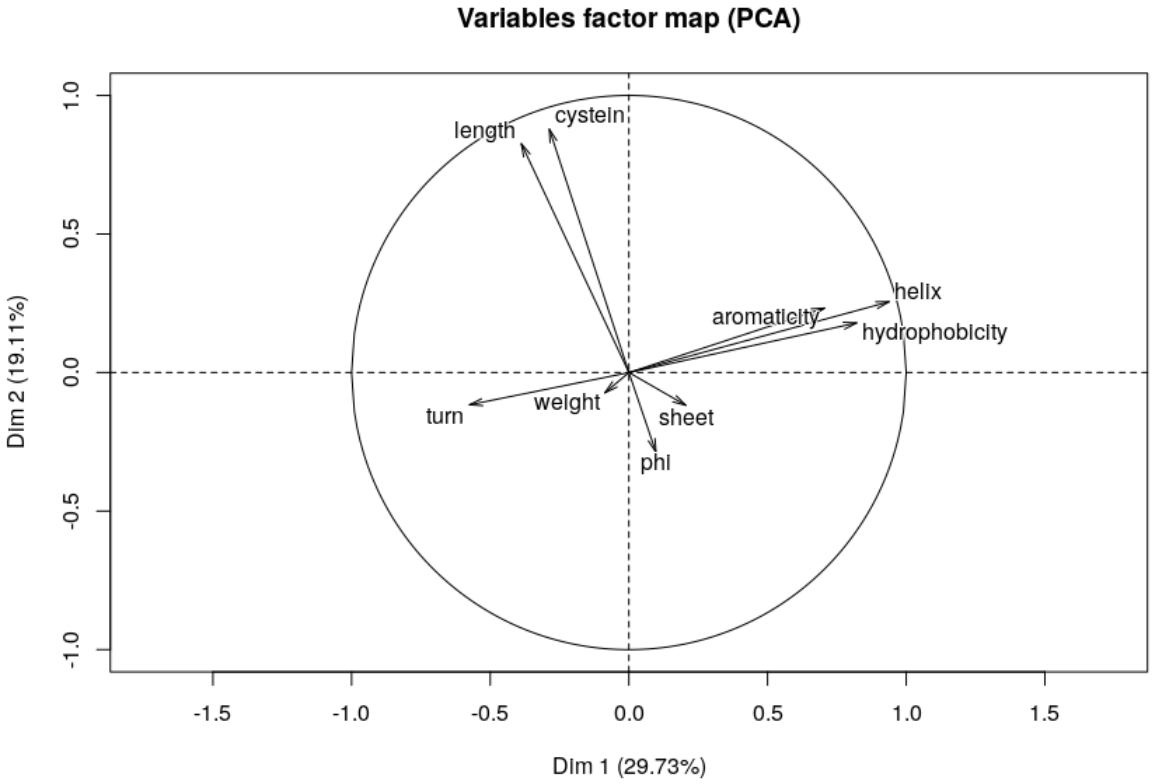

FIGURE 2 – Représentation graphique de la Analyse par Correspondance Principale pour le jeu de donnée *realtable*.csv

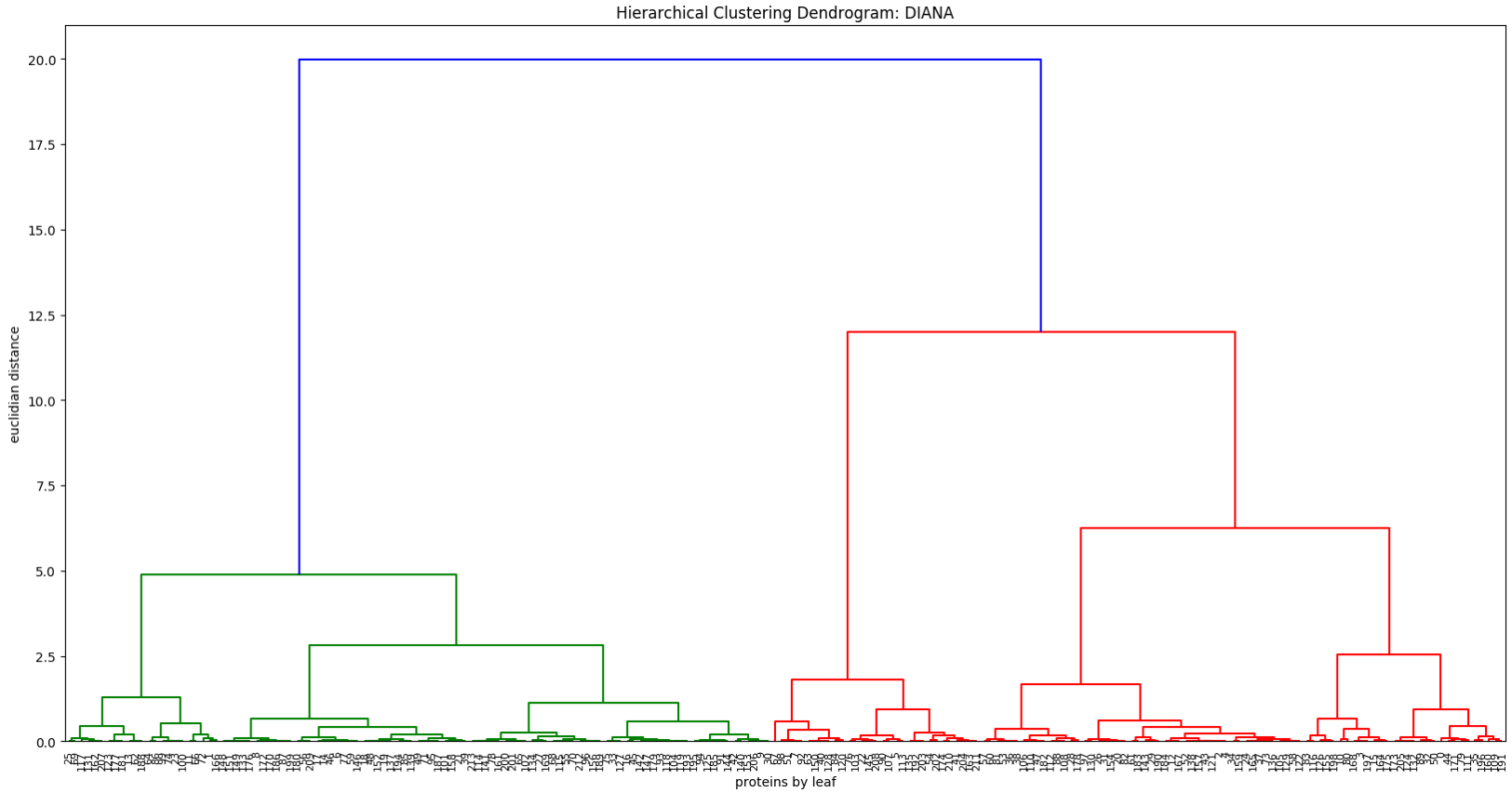

FIGURE 3 – Dendrogramme obtenu après avoir effectué la clusterisation avec la méthode DIANA

```
clusterisation of Human proteins
cluster numero:1
moyenne taille :81.0 ecart-type taille :2.0
moyenne poids :5057141.95 ecart-type poids :39792.55
moyenne point isoéléctrique :8.27941894531 ecart-type point isoéléctrique :0.124694824219
moyenne helix :0.216028671649  ecart-type helix :0.000838798230898
moyenne turn : 0.494128412384  ecart-type turn : 0.0122007015403
moyenne feuillet : 0.0183773066951  ecart-type feuillet : 0.00571907884703
moyenne hydrophobicité :-0.230631386305  ecart-type  hydrophobicité : 0.0344288546591
moyenne cysteines :16.5  ecart-type cysteines : 1.5
moyenne aromaticité :0.210004575263  ecart-type aromaticité : 0.00518529815464
cluster numero:2
moyenne taille :166.0 ecart-type taille :0.0
moyenne poids :416610449.1 ecart-type poids :0.0
moyenne point isoéléctrique :10.5162963867 ecart-type point isoéléctrique :0.0
moyenne helix :0.319277108434  ecart-type helix :0.0
moyenne turn : 0.39156626506  ecart-type turn : 0.0
moyenne feuillet : 0.198795180723  ecart-type feuillet : 0.0
moyenne hydrophobicité :0.159036144578  ecart-type  hydrophobicité : 0.0
moyenne cysteines :4.0  ecart-type cysteines : 0.0
moyenne aromaticité :0.168674698795  ecart-type aromaticité : 0.0
cluster numero:3
moyenne taille :69.25 ecart-type taille :17.246376431
moyenne poids :299808053.666 ecart-type poids :190290904.133
moyenne point isoéléctrique :8.27751159668 ecart-type point isoéléctrique :0.650559802684
moyenne helix :0.325293083572  ecart-type helix :0.0286225371219
moyenne turn : 0.453214488972  ecart-type turn : 0.0637290680613
moyenne feuillet : 0.0691499768572  ecart-type feuillet : 0.0353199900172
moyenne hydrophobicité :-0.308481687168  ecart-type  hydrophobicité : 0.12344435253
moyenne cysteines :7.0  ecart-type cysteines : 1.65831239518
moyenne aromaticité :0.270292604163  ecart-type aromaticité : 0.043318368738
```

FIGURE 4 – Extrait du fichier texte contenant les valeurs des écart-types et des moyennes de chaque paramètres pour les *clusters* 1 et 2

```
4621 moyenne aromaticite :0.07650673400b7  ecart-type aromaticite : 0.002011784511718
4622 cluster numero:463
4623 moyenne taille :2974.0 ecart-type taille :1031.09116959
4624 moyenne poids :1156964927.05 ecart-type poids :228897705.83
4625 moyenne point isoéléctrique :5.87994384766 ecart-type point isoéléctrique :0.452209530322
4626 moyenne helix :0.303673439971  ecart-type helix :0.0138476869683
4627 moyenne turn : 0.238536310965  ecart-type turn : 0.0146141434752
4628 moyenne feuillet : 0.255781387352  ecart-type feuillet : 0.0148777906621
4629 moyenne hydrophobicité :-0.284138161219  ecart-type  hydrophobicité : 0.060470354709
4630 moyenne cysteines :78.375  ecart-type cysteines : 21.4414172806
4631 moyenne aromaticité :0.0822366115821  ecart-type aromaticité : 0.00427253070307
4632 cluster numero:464
4633 moyenne taille :1151.84761905 ecart-type taille :364.812065928
4634 moyenne poids :1076798879.12 ecart-type poids :237115107.624
4635 moyenne point isoéléctrique :5.58140055339 ecart-type point isoéléctrique :0.416913268479
4636 moyenne helix :0.283724489043  ecart-type helix :0.0200120666245
4637 moyenne turn : 0.24767923211  ecart-type turn : 0.0209210904293
4638 moyenne feuillet : 0.262348084668  ecart-type feuillet : 0.0168754275585
4639 moyenne hydrophobicité :-0.442615583785  ecart-type  hydrophobicité : 0.146696075768
4640 moyenne cysteines :22.7428571429  ecart-type cysteines : 10.0323558183
4641 moyenne aromaticité :0.0840169883248  ecart-type aromaticité : 0.00903584913296
4642 cluster numero:465
4643 moyenne taille :1164.26605505 ecart-type taille :353.696019442
4644 moyenne poids :1030042316.69 ecart-type poids :223652000.968
4645 moyenne point isoéléctrique :7.51268243352 ecart-type point isoéléctrique :0.957129518301
4646 moyenne helix :0.291064937372  ecart-type helix :0.021863206181
4647 moyenne turn : 0.244192317959  ecart-type turn : 0.0193727023202
4648 moyenne feuillet : 0.256557033032  ecart-type feuillet : 0.0127032905998
4649 moyenne hydrophobicité :-0.409324154993  ecart-type  hydrophobicité : 0.153649120422
4650 moyenne cysteines :21.7981651376  ecart-type cysteines : 9.58379731345
4651 moyenne aromaticité :0.0843521773715  ecart-type aromaticité : 0.00892958564085
```

FIGURE 5 – Extrait du fichier texte contenant les valeurs des écart-types et des moyennes de chaque paramètres *clusters* 464 et 465

```
 1 Q3LI58, Q3LI59
 2 O96M19
 3 Q3LI66, Q3LHN0, Q3LI61, Q3LI63, Q3LI64, Q3LI68, Q3SYF9, Q8IUB9
 4 Q96E09, P31942, Q8IZP0, Q99217, A0A1B0GTS1, A0A1B0GWH4, Q496A3, Q6ZTU2, P50548, Q9H1C7, Q92734,
   O95429, P98082, P20073, Q9UL17, Q969T9, Q96IK1, Q13094, Q9BXJ4, Q13542, Q92997
 5 Q6ZS52, A6NJB7, Q5PR19, Q14814, O00401, P35908, Q96AH0, Q9Y4X4, Q8N365, Q9Y6J3, P26367, F2Z3F1,
   Q8NEP4, Q13151, Q9Y2V2, P41162, P0C862, O75909, Q9NZV6, Q12857, Q9Y5A9, Q9P2K5, Q12906, P98179,
   Q9BXJ5, P04156, Q15434, Q32P51, P51991, P31276
 6 Q569K4, Q8TDC3, A6NFR6, O75593, P01860, Q07352, Q8NEA6, Q9UJM3, Q9UQC2, P14866, Q9Y215, Q9NZJ0,
   P11161, Q6Q6R5, Q9Y2X9, Q5JXC2, P26651, Q9NYJ8, P20393
 7 Q13480, P02671, O00409, Q9UGP4, Q9NUC0, Q9H9S0, Q8NBF1, Q2WGN9, Q5QP82, Q96SF7, Q9UPW0, Q4G112,
   Q7Z5Q1
 8 Q9H1U4
 9 Q52LG2, Q3LI83, P59991, Q6PEX3
10 Q6ZSA8, Q8N7U7, Q5VSD8, Q86Z23, Q9H6D8, O15255, P02747, Q01167, Q9HAK2, P10323, P02745, Q8WUH6,
   Q8IYJ0, O95081, Q8N7I0, Q0VFX4, P02746, P52594, P49840, A6NDX4, Q9Y4X0, Q6ZN03, Q8N0V1, Q9HB31,
   Q8NFH5
11 A0A075B6I0, A2A2V5, P01721, A0A075B6J9, Q6ICB0, Q9BXJ2, P80748, P06731, P01700, Q9H305, Q8N814,
   O75973, Q96MT4, Q19T08, Q8N1V8, Q5SWW7, B1ATL7, Q9BXJ0, Q9UI25, Q96LM9, Q06416, P01706, A0A0A0MT36,
   A0A0B4J1Y8, Q6ZN04, P01701, Q8WWR9
12 Q9BUV0
13 P38159
14 Q15415, A6NDE4, P0C7P1
15 Q16629, Q8WXF0
16 Q8NDC0, Q9BRQ0, Q5T035, P0C841, P56693, P81877, Q92934, P14653, Q9HC44, Q9Y3Y4, Q9BWW4
17 P17483, Q8ND56, P55316, O60481, Q8IWX8, O43365, P14651, Q8N9P0, Q6PB30
18 Q8IUG1, Q07627
19 P59990, P60329
20 A8MUX0
21 P60368, P60014, P60369, P60413, P60412, P60331, P60411, P60410
```

FIGURE 6 – Extrait du fichier texte contenant les *clusters* avec les identifiants des protéines qu'ils contiennent

Références

[1] Gregory A Petsko, Dagmar Ringe, and Mme Dominique Charmot. *Structure et fonction des protéines*. De Boeck Supérieur, 2008.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning : Data mining, inference, and prediction. *Biometrics*, 2002.

[3] Anil K Jain. Data clustering : 50 years beyond k-means. *Pattern recognition letters*, 31(8) :651–666, 2010.

[4] Ashish Kumar Patnaik, Prasanta Kumar Bhuyan, and KV Krishna Rao. Divisive analysis (diana) of hierarchical clustering and gps data for level of service criteria of urban streets. *Alexandria Engineering Journal*, 55(1) :407–418, 2016.