

Méthodologie pour l'analyse des données SAGE : Discovery Small 74*822

1. Planifier des buts clairs et précis.

Etablir des buts fixes et établir avec l'aide du biologiste les informations intéressantes.

2. Prétraitement des données.

Dans les données du type SAGE il est indispensable faire deux types de traitement. Dans un premier temps nous devons identifier les « tags », présents de façon unique dans une gène, afin d'éviter ambivalence dans l'étape d'interprétation. Dans un deuxième temps nous avons fait scénarios :

- Données brutes,
- Données normalisées par nombre de tags par expérience.

3. Classification et Validation des expériences

Nous avons appliqué à nos scénarios possibles des données des algorithmes de classification afin de trouver des outliers au sein des 74 expériences. Nous avons utilisé des algorithmes de classification non supervisée tels que : clustering hiérarchiques (UPGMA, agnes, clara, diana), fanny, k-means, et des algorithmes de classification supervisée, tels que les arbre des décision : C.5.0 et CRT et les réseaux de neurones. Nous avons également testé plusieurs mesures de distances : euclidienne, manhattan, produit vectoriel, covariance, Pearson et Spearman. Le but de cette partie était de valider nos données et de trouver des classifications pertinentes par tissus et par cellules cancérigènes et non cancérigènes avec une confiance élevée (Plus de 90%).

4. Sélection et Classification des gènes

- Dans un premier temps nous avons appliqué des tests tels que les statistiques F, T et de Fischer afin de trouver les gènes qui se sont exprimés de façon différentielle.
- Ensuite nous avons classifié et validé nos gènes en utilisant les mêmes algorithmes que dans l'étape précédente et en classifiant cette fois ci les gènes.
- Identification des clusters forts, c'est-à-dire ceux qui apparaissent de façon constante dans plusieurs méthodes de classification, ainsi comme des associations entre les gènes. Nous comprenons comme « Cluster Fort » à l'ensemble de gènes qui apparaissent dans le même cluster quotidiennement avec une confiance élevée, en faisant varier la méthode de classification et la distance.

5. Résultats et Interprétation des données

- Sélectionner à la main les clusters « forts ». Valider et interpréter les clusters « forts » à partir des programmes qui font des annotations, classifications et graphiques sur des bases de données génomiques, tels que THEA, GO Tools, Microarray Tools etc.
- Validation par les biologistes des résultats obtenus.
- Définition de nouvelles informations « intéressantes » du point de vue des biologistes

Premier Scénario à partir des données brutes

Classification et validation des Expériences :

Dans un premier temps, nous avons utilisé les données brutes par tissu et nous avons appliqué les algorithmes de classification supervisés et non supervisé pour identifier les tissus des classes Cancer et Non Cancer pour chaque tissu. Dans le tableau suivant nous pouvons observer la répartition par tissu pour chacune des classes Cancer Bulk, Cancer Cell Line, Normal Bulk et Normal Cell Line.

	Répartition par tissu selon l'état et le type de cellule				
	Cancer Bulk	Cancer Cell Line	Normal Bulk	Normal Cell Line	Total
Brain	8	7	5	1	21
Breast	6	3	2	0	11
Colon	2	4	2	0	8
Kidney	0	2	0	0	2
Ovary	3	4	0	2	9
Pancreas	0	3	2	2	7
Prostate	3	6	2	0	11
Peritoneum	0	0	1	0	1
Skin	1	0	0	1	2
Vessel	0	0	0	2	2
Total	23	29	14	8	74

Table 1 : Répartition Par tissu-état-technique.

Nous avons choisi les 6 premiers tissus, car il nous permettront de faire notre classification par rapport à l'état : cancer ou non cancer, et ainsi trouvé les possibles outliers au seins des expériences. (Voir Table 1).

Nous présenterons les résultats des classifications pour chaque organe de la façon suivant, les règles extraites à partir des algorithmes supervisés : C.5.0 et CRT et le dendrogramme le plus pertinent pour classifier les tissus cancéreux et non cancéreux en indiquant l'algorithme de classification et la mesure de distance utilisée. Finalement nous indiquons les outliers pour chaque type de tissu, c'est-à-dire les conditions expérimentales (tissus) présentant des « anomalies ».

PANCREAS

Méthode de C.5.0 :

a. Règle :

Si $ENSG00000173402 > 2 \rightarrow$ Cancer
Sinon Normal.

Effectivité de la règle : 100 %

Méthode CRT :

b. Règle :

Si $ENSG00000198561 > 2.5 \rightarrow$ Cancer
Sinon Normal

Effectivité de la règle : 100 %

Résultats de Clustering

Algorithmes Appliquées :

- Agnes
- Diana
- UPGMA
- K-means

Mesures de distance entre clusters :

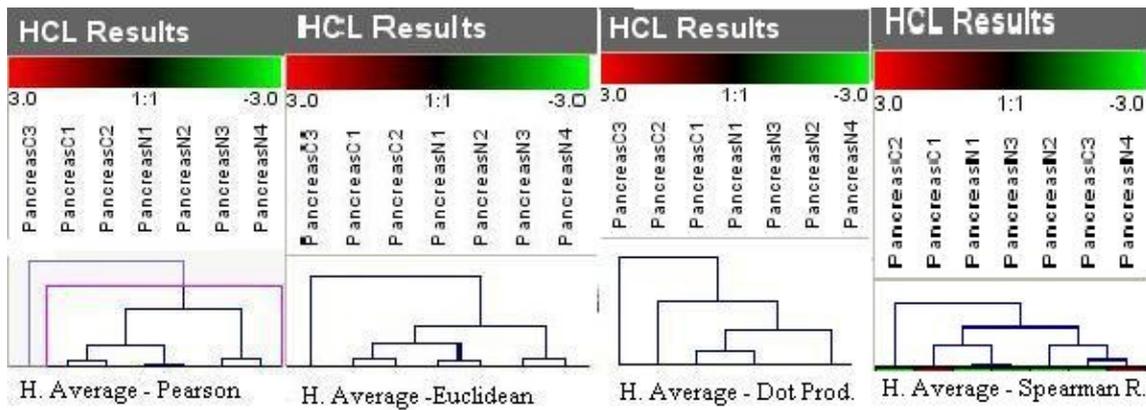
- Single linkage
- Average linkage
- Complete linkage

Mesures de distances entre conditions :

- Euclidienne
- Manhattan
- Pearson
- Spearman
- Produit Vectorielle

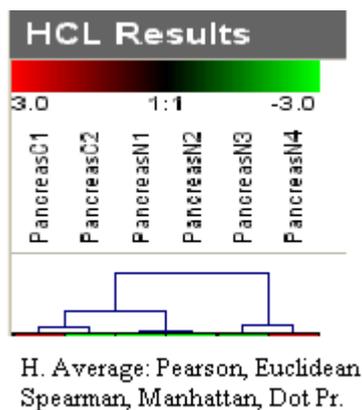
L'analyse en composantes principales nous a permis de déterminer que les 3 premières composantes expliquent 97.2% de la variance totale (1^{er} Composant 69.579%, 2^{eme} Composant 21.296%, 3^{eme} Composant 6.334 %).

D'après les résultats résumés dans la figure ci-dessous nous avons trouvé que l'expérience Pancreas C3 se comporte de façon isolé aux autres.



Nous avons réalisé la même expérience en enlevant l'outlier : Pancreas C3 et ainsi vérifié que la distribution des composantes principales reste sans changements : 1^{er} Composant 73.23%, 2^{ème} Composant 19.525%, 3^{ème} Composant 4.439 %. C'est-à-dire que les trois premières composantes expliquent 97.2% de la variance totale.

En appliquant les divers algorithmes de clustering nous avons trouvé de façon constante la même répartition en clusters quelque soient les mesures de distances utilisées, ce qui nous indique la robustesse de nos expériences.



Nous pouvons observer la proximité entre les quatre premières expériences, lesquelles sont des expériences de type « cell line » à la différence de PancreasN3 et PancreasN4 qui sont de type Bulk line.

BRAIN

Méthode de C.5.0 :

a. Règle :

Si **ENS00000198763** > **75** → Cancer

Sinon Normal.

Effectivité de la règle : 94 % soft pour le cas de la condition BrainN4.

b. Règle :

Si **ENS00000198888** ≤ **97** → Cancer

Sinon Normal.

Effectivité de la règle : 94 % soft pour le cas de la condition BrainN4.

Méthode CRT :

a. Règle :

Si **ENS00000197029** > **8.5** → Cancer

Sinon Si **ENS00000099964** < **6.5** → Cancer

Sinon Normal

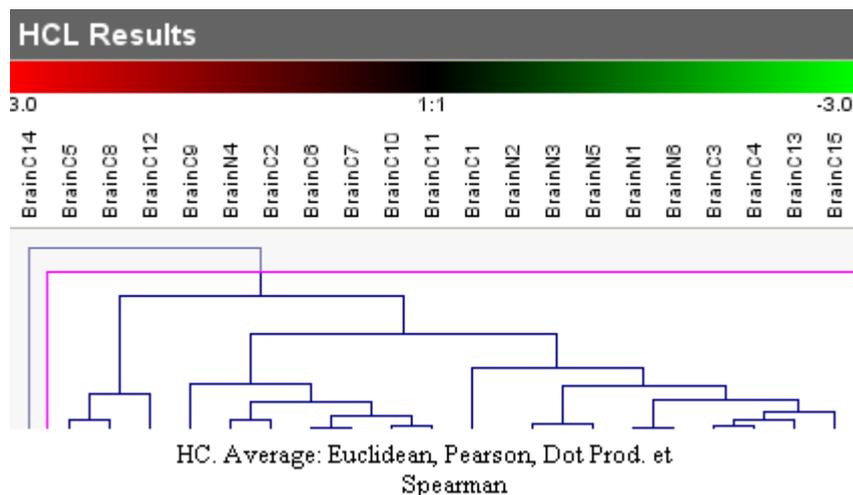
Effectivité de la règle : 94 % soft pour le cas de la condition BrainN4.

Tant les algorithmes C.5, CRT nous indiquent que la condition Brain N4 est une condition a enlever.

Résultats de Clustering :

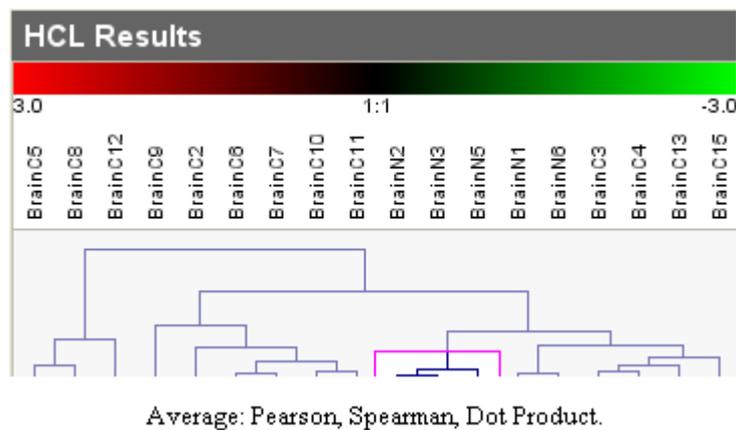
L'analyse en Composantes Principales nous a permis de déterminer que les 3 premières composantes expliquent 86.6% de la variance total (1^{er} Composant 57.12%, 2^{eme} Composant 20.981%, 3^{eme} Composant 8.571 %).

D'après les résultats résumés dans la figure ci-dessous nous avons trouvé que l'expérience BrainN4 nous ajoute du bruit, ainsi comme les conditions cancérigènes BrainC14 (qui est un outlier) et la condition BrainC1 qui est un outlier aussi (même s'il est juste un mineur degré).



Nous avons réalisé la même expérience en enlevant N4, C14 et C1, et ainsi vérifié que la des composantes principales reste sans changements : 1^{er} Composant 67.97%, 2^{eme} Composant 15.93%, 3^{eme} Composant 6.435 %. C'est-à-dire que le trois premières composants explique 90.3% de la variance totale.

En appliquant les diverses algorithmes de clustering nous avons trouvé de façon constant la même répartition en clusters quelque soient les mesures de distances utilisées, ceux qui nous indique la robustesse de nos expériences



Nous pouvons observer dans la figure antérieure la division en quatre clusters, le groupe des Brain Normales N1-N6, le groupe de Brain Bulk : C13, C15, C3, C4, le groupe des brain cancer cell line (7 en totale) et les deux cancer bulk : C2 et C5, sont mis ensemble.

Breast

Méthode de C.5.0 :

a. Règle :

Si **ENSG00000130175=0** → Cancer

Sinon Normal.

Effectivité de la règle : 100 %

Méthode CRT :

b. Règle :

Si **ENSG0000010278<10.5** → Cancer

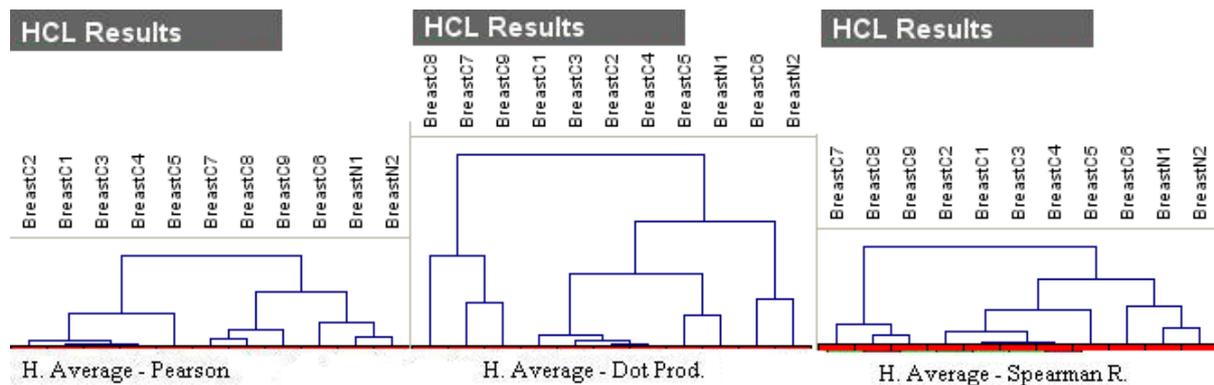
Sinon Normal

Effectivité de la règle : 100 %

Résultats de clustering

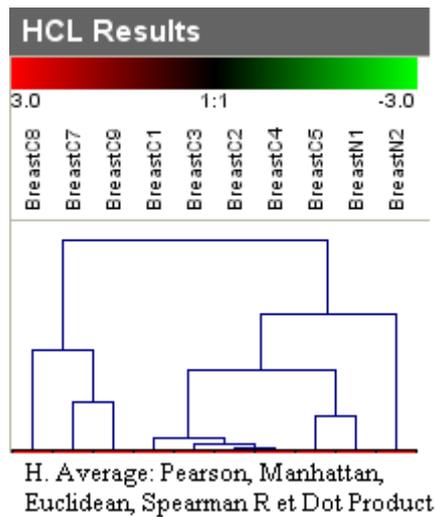
L'analyse en Composantes Principales nous a permis de déterminer que les 3 premières composantes expliquent 90.58% de la variance total (1^{er} Composant 67.27%, 2^{eme} Composant 14.286%, 3^{eme} Composant 9.09 %, 4^{me} Composant 6.59 %).

D'après les résultats résumés dans la figure ci-dessous nous avons trouvé que l'expérience BreastC6 nous ajoute du bruit.



Nous avons réalisé la même expérience en enlevant Breast C6, et ainsi vérifié que l'analyse des composantes principales a été amélioré significativement: 1^{er} Composant 69.88%, 2^{eme} Composant 14.398., 3^{eme} Composant 9.911 %. C'est-à-dire que le trois premières composantes expliquent 94.2% de la variance total.

En appliquant les diverses algorithmes de clustering nous avons trouvé de façon constant la même répartition en clusters quelque soient les mesures de distances utilisées, ceux qui nous indique la robustesse de nos expériences



Nous pouvons observer dans la figure antérieure la distribution des expériences : Les expériences C7, C8 et C9 sont de type cancer cell line alors ils sont classés ensemble, en tant que les expériences C1-C5, ainsi comme les expériences N1 et N2 sont de type bulk et ils sont classés séparément. Ceux-ci nous indique comme pour les Pancréas que nous avons deux classes, par maladie : Cancer et Non Cancer ainsi comme par type d'expérience : Bulk ou Cell line

Prostate

Méthode de C.5.0 :

a. Règle :

Si **ENSG00000140740** ≤ 5 → Cancer

Sinon Normal.

Effectivité de la règle : 100 %

Méthode CRT :

b. Règle :

Si **ENSG00000110442** < 2.5 → Cancer

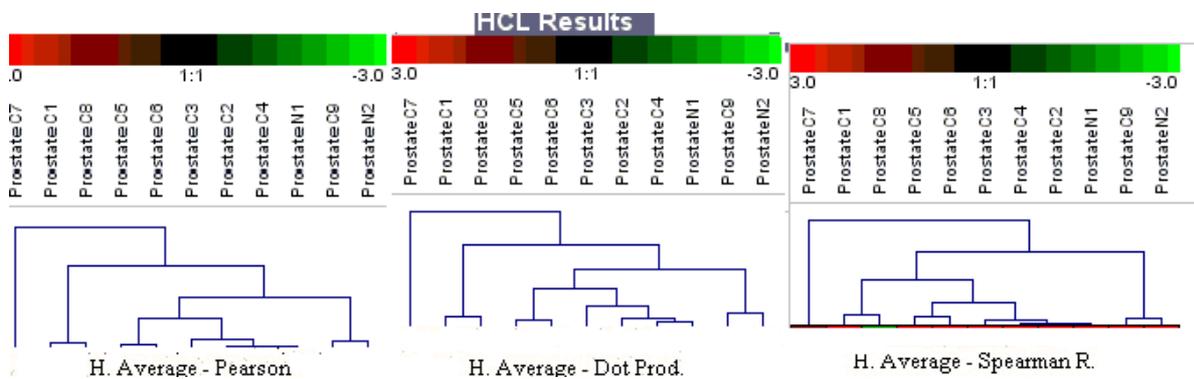
Sinon Normal

Effectivité de la règle : 100 %

Résultats de clustering

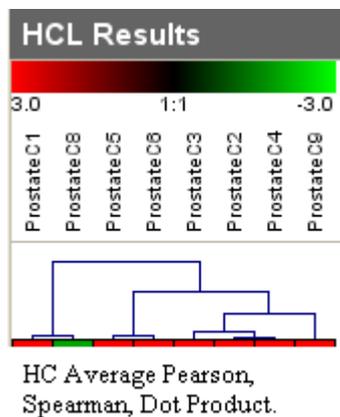
L'analyse en Composantes Principales nous a permis de déterminer que les 3 premières composantes expliquent 82.96% de la variance total (1^{er} Composant 66.14%, 2^{eme} Composant 9.30%, 3^{eme} Composant 7.50 %, 4^{me} Composant 4.79 %). En termes de clusters, nous avons besoin au plus de quatre clusters afin de prendre la plus parte de variance dans les données. Comme nous avons seulement quatrepossibles classes : Cancer, Normal, Bulk et Line, ça nous indique que il y en beaucoup de bruit dans ces conditions.

D'après les résultats résumés dans la figure ci-dessous nous avons trouvé que l'expérience C7 est un outlier. Dans le cas de N1 il se comporte comme C2 et C4 alors on a décidé de l'éliminer également, car il est une expérience qui ne serve pas du tout à décrire l'état Normal. Le cas de N2 en étant si proche à ses voisins cancérigènes (C9, C2 et C4) nous avons décidé d'enlever aussi.



Nous avons réalisé la même expérience en enlevant ProstateC7, ProstateN1 et ProstateN2, et nous avons regardé que la distribution des composantes principales a été amélioré significativement: 1^{er} Composant 77.10%, 2^{eme} Composant 11.875, 3^{eme} Composant 4.9 % et la 4^{eme} Composant 4.17 % C'est-à-dire que le trois premières composants explique 93.88% de la variance total.

En appliquant les diverses algorithmes de clustering nous avons trouvé de façon constant la même répartition en clusters quelque soient les mesures de distances utilisées, ceux qui nous indique la robustesse de nos expériences



En ce qui concerne à la classification de bulk et normal, nous pouvons regarder dans la figure ci-dessus, que les deux expériences « bulk » : C4 et C9 apparaissent proches et tous les restantes ou expériences cell line apparaissent plus au moins près les un des autres.

Colon + Peritonéum

Méthode de C.5.0 :

a. Règle :

Si **ENSG00000010278** <8 → Cancer

Sinon Normal.

Effectivité de la règle : 100 %

Méthode CRT :

b. Règle :

Si **ENSG00000175061** >=2 → Cancer

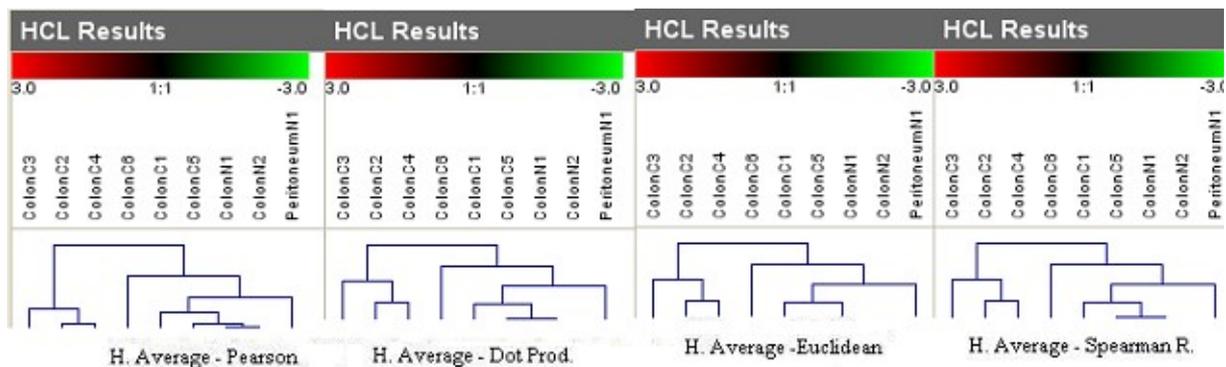
Sinon Normal

Effectivité de la règle : 100 %

Résultats de clustering

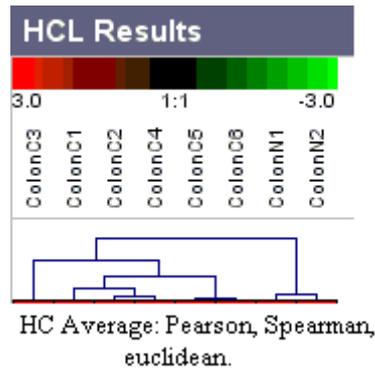
L'analyse en composantes principales nous a permis de déterminer que les 3 premières composantes expliquent 93.467% de la variance totale (1^{er} Composant 79.09%, 2^{eme} Composant 9.363%, 3^{eme} Composant 5.007 %, 4^{me} Composant 2.417 %).

D'après les résultats résumés dans la figure ci-dessous nous avons trouvé que l'expérience Peritonéum, même s'il se trouve dans le colon, cette tissu a des caractéristiques particuliers, qui lui signaler comme outlier dans notre ensemble pour le Colon.



Nous avons réalisé la même expérience en enlevant Peritonéum N1, et ainsi vérifié que la distribution des composantes principales a été amélioré suffisamment: 1^{er} Composant 80.47%, 2^{eme} Composant 9.99, 3^{eme} Composant 4.872 %. C'est-à-dire que le trois premières composants explique 95.328% de la variance total.

En appliquant les algorithmes de clustering nous avons trouvé de façon constant la même répartition des clusters a travers de tous les méthodes et des distances, ceux qui nous indique la robustesse de nos expériences (a exception des légères différences dans la distance produit vectorielle).



Nous pouvons observer dans la figure ci-dessus la distribution des expériences : Les expériences C1-C4 sont de type cell, alors ils sont classés ensemble, en tant que les expériences C5, C6 ainsi comme les expériences N1 et N2 sont de type bulk et ils sont classés séparément. Ceux-ci nous indique que nous avons deux classes, par maladie : Cancer et Non Cancer ainsi comme par type d'expérience : Bulk ou Cell line.

Ovaire

Méthode CRT :

a. Règle :

Si ENSG00000196563 < 17.5 → Cancer

Sinon Normal

Effectivité de la règle : 100 %

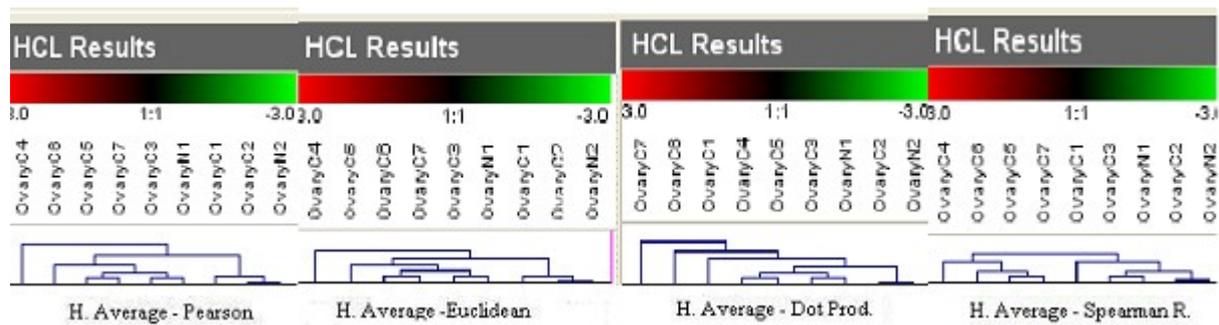
Méthode de C.5.0 :

Non trouvé

Resultats de Clustering

L'analyse en composantes principales nous a permis de déterminer que les 3 premières composantes expliquent seulement 88.596% de la variance total (1^{er} Composant 66.11%, 2^{eme} Composant 12.53%, 3^{eme} Composant 9.956 %, 4^{me} Composant 3.92 %). Comme nous avons seulement quatre possibles classes : Cancer, Normal, Bulk et Line, ca nous indique que il y en beaucoup de bruit dans ces conditions.

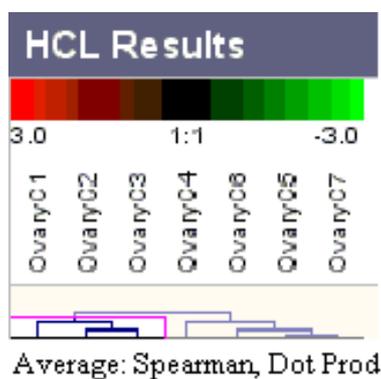
D'après les résultats résumés dans la figure ci-dessous nous avons trouvé que l'expérience que l'expérience N1 est jumelée de C3 et N2 de C2, au fur est au mesure de tous les algorithmes et des distances. En plus, ils se comportent comme les expériences cancérigènes, ainsi nous l'enlèverons, car il ajoute du bruit à l'expérience normale.



Nous avons réalisé la même expérience en enlevant OvaireN1 et Ovaire N2, et ainsi vérifié que la distribution des composantes principales a été amélioré significativement: 1^{er} Composant 65.756%, 2^{eme} Composant 13.707, 3^{eme} Composant 10.24 % et la 4^{eme} Composant 4.61 % C'est-à-dire que le trois premières composants explique 89.7% de la variance total.

Nous avons observé que en enlevant les expériences « normales » nous avons gagné en variance capté et nous avons enlevé le bruit.

En appliquant les algorithmes de clustering nous avons trouvé de façon constant la même répartition des clusters a travers de tous les méthodes et des distances, ceux qui nous indique la robustesse de nos expériences (a exception des légères différences dans la distance produit vectorielle).

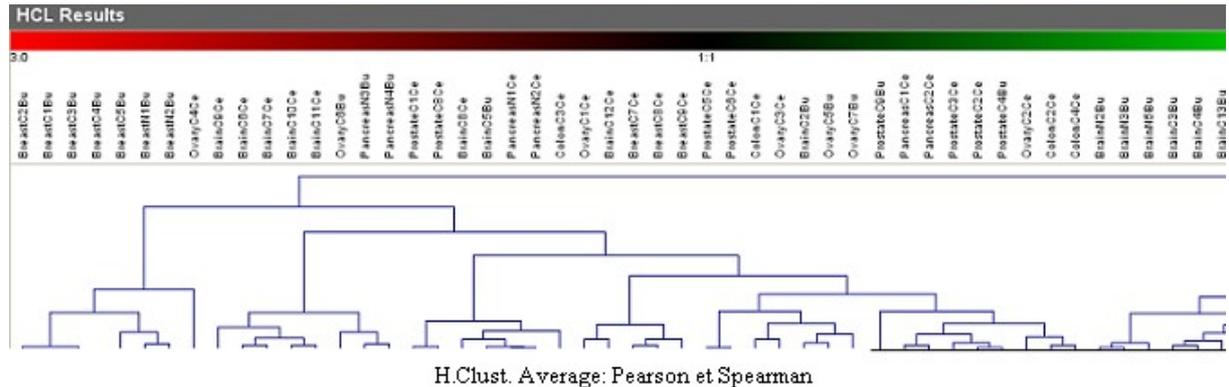


En appliquant les algorithmes de clustering nous avons trouvé un division cohérent des expériences cancérigènes : les expériences bulk C5-C7 groupes ensemble, ainsi comme les expériences du type cell line C1-C4.

Jeux des données sans les outliers des conditions expérimentales

Nous avons enlevé les 10 conditions ou outliers suivantes : PancreasC3, BrainN4, BrainC14, BrainC1, BreastC6, ProstateC7, ProstateN1, ProstateN2, OvaireN1, OvaireN2, par les raisons que nous avons indiqué dans les paragraphes précédentes. Finalement nous avons resté avec 64 conditions, auxquelles il existe 7 conditions ou tissus très peu représentés : Skin (2), Kidney(2), Vessel (2), et Peritonneum (1).

Une fois que nous avons enlevé tous les conditions expérimentales outliers, nous avons testé avec nos algorithmes de clustering la répartition de l'ensemble total avec des différentes distances



Nous pouvons regarder la correcte distribution des tissus en quatre classes par tissu : cancérigène bulk ou cancérigène cell, ou bien normal bulk ou normal cell. Nous pouvons regarder que les conditions se classifient par groupes de Cancer et Normal et de Bulk et Cell. Le seul organe qui se classifie de façon plutôt différentielle est l'ovaire, peut être à cause qu'il est très peu représenté. Dans ce graphe nous avons enlevé les tissus très peu représenté pour montrer la correct répartition des tissus représentatives : Brain, Breast, Prostate, Pancreas, Colon et Ovary.

En regardant l'analyse en composants principaux nous avons 6 composants représentatives, ce qui nous confirme notre division en six tissus. Ainsi, les 6 premières composants contiennent 92% de la variance total du jeux des données.

Nous avons appliqués les mêmes analyses aux jeux des données normalisées par Tags, cela veut dire avec des normalisations typiques pour des données SAGE :

$$nx_{ij} = X_{ij} * \text{Max}_j \{ X_{ij} \} / \text{Totag}_j$$

x_{ij} est un élément i, j de la matrice des jeux des données.

$\text{Max}_j \{ X_{ij} \}$ est le valeur maximal entre les éléments du colonne j .

Totag_j Nombre Total de Tags par Colonne j .

Nous avons appliqué la normalisation aux jeux des données original et nous avons trouvé les mêmes outliers que dans celle-ci (le document il est en cours de rédaction).

Il reste faire le pas 4 et 5 que nous avons vu au début de cette rapport.