Master 2 Machine Learning for Data Science

TP3
Analysis of genomic and ploidy alterations in breast tumors

Réalisé par:
Amira KOUIDER - 21904040 FI
Nadia RADOUANI – 21911973 FI

Enseignante: Mme Séverine AFFELDT

Année universitaire: 2020-2021

1. *Jeu de données:*

Nous voulons analyser les altérations génomiques des tumeurs du sein à partir du catalogue en ligne ***COSMIC***.

La base de données utilisée est ***CosmicCancer,*** elle contient 807 échantillons d'informations sur le niveau d'expression de 91 gènes. Ces gènes ont été sélectionnés sur la base d'études antérieures sur les mutations et/ou les altérations de l'expression dans le cancer du sein.

La base contient trois catégories de variables : les variables en majuscule représentent l'état d'expression d'un gène (Over/under/normal), les variables en minuscule représentent si un gène est muté (y/n), et finalement, la variable 'Ploidy' contient deux modalités : 1 représente la cellule diploïde et 2 représente la cellule triploïde, cette variable a été discrétisée, car, elle était à l'origine un mélange de 2 distributions gaussiennes.

2. *Nettoyage de la base :*

Afin d'appliquer l'algorithme hill-climbing, un nettoyage de la base de données s'avérait être nécessaire. Une première étape de ce nettoyage consistait à supprimer les valeurs manquantes. La base contient 8 NA présentes dans la variable 'Ploidy'. Du fait, après ce nettoyage, nous avons retiré 1% des échantillons de l'ensemble des données.

Une deuxième étape consistait à supprimer les variables constantes et une dernière étape pour convertir la variable 'Ploidy' en type 'factor'.

3. *Network reconstruction with the hill-climbing approach*

Le modèle hc :

```
nodes:                                  162
arcs:                                   199
  undirected arcs:                      0
  directed arcs:                        199
average markov blanket size:            2.96
average neighbourhood size:             2.46
average branching factor:               1.23

learning algorithm:                     Hill-Climbing
score:                                  BIC (disc.)
penalization coefficient:               3.34168
tests used in the learning procedure:   46368
optimized:                              TRUE
```

Figure 1 - aperçu du modèle hc

Igraph network

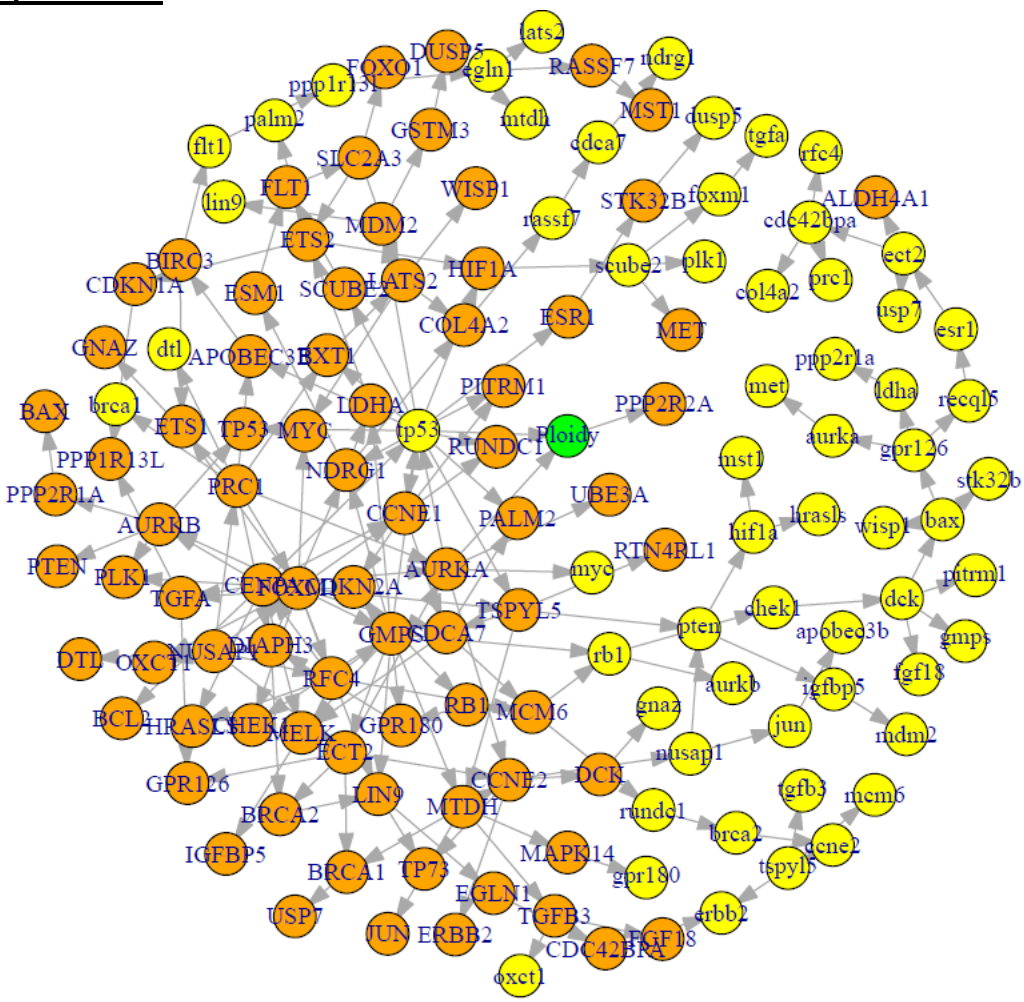


Figure 2- HC Graph

A partir du graphe, nous constatons que 'Ploidy' est affecté par 'tp53' et 'AURKA', tandis qu'il affecte 'PPP2R2A'.

Relations entre les gènes

```
> #Idendify the lower case nodes most related to upper case nodes
> lc_most_related_to_uc(graph_hc)[1:10]
  tp53 scube2    rb1  egln1   flt1 rassf7   gnaz   mtdh   ect2   lin9
    16      2      2      1      1      1      1      1      1      1
> #Identify the hubs
> sort(hub_score(graph_hc)$vector, decreasing = TRUE)[1:10]
     tp53     FOXM1      GMPS     CENPA     CCNE1      PRC1      RFC4    DIAPH3
1.0000000 0.9787804 0.6589622 0.4750144 0.2547987 0.2510718 0.1744653 0.1722574
    AURKA     CDCA7
0.1624072 0.1462756
> #Top 10 nodes and edges interms of betweenness centrality measure.
> sort(betweenness(graph_hc), decreasing = TRUE)[1:10]
    tp53     MCM6      dck      rb1    chek1      bax    CDCA7    NDRG1   DIAPH3
424.1667 332.0000 300.0000 299.0000 294.0000 256.0000 252.7333 240.0000 232.0000
  gpr126
221.0000
```

Figure 3 – Part 01 : Mutated genes most related to gene expression / Part 02 : Hubs of the HC graph /Part 03 : Top 10 nodes in terms of betweenness centrality measure.

Nous constatons que ‘tp53’ joue un rôle important, car, il est lié à ‘Ploidy’ et à 15 gènes d’expression. Il est aussi le premier nœud en termes de centralité d'intermédiarité (betweenness centrality), et le plus grand ‘hub’ dans le graphe.

	From	To
1	chek1	dck
2	rb1	chek1
3	CDCA7	MCM6
4	dck	bax
5	bax	gpr126
6	GPR180	DIAPH3
7	AURKA	tp53
8	MCM6	rb1
9	RB1	GPR180
10	DIAPH3	NDRG1

Tableau 1 – Top 10 edges in terms of betweenness centrality measure.

L’arête **rb1 -> chek1** se positionne dans le top 10 des arêtes en termes de betweenness centrality. Par ailleurs, nous remarquons que **rb1**, **chek1**, **dck**, **bax** et **gpr126** forment une de chaine.

4. *Network reconstruction with the PC approach*

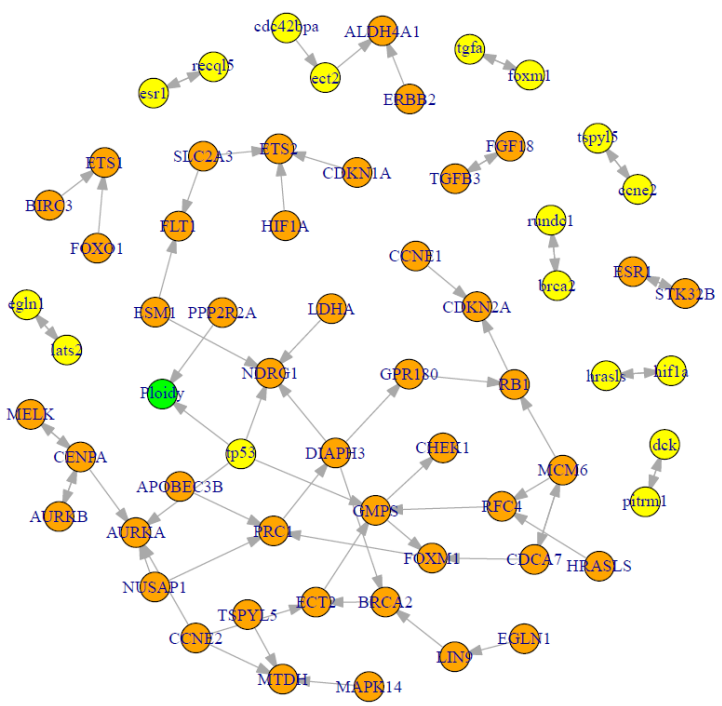
Le modèle pc :

```
Object of class 'pcAlgo', from Call:
pc(suffStat = suffStat, indepTest = disCItest, alpha = 0.01,
    labels = colnames(data_pc))
Number of undirected edges:  12
Number of directed edges:    45
Total number of edges:       57
```

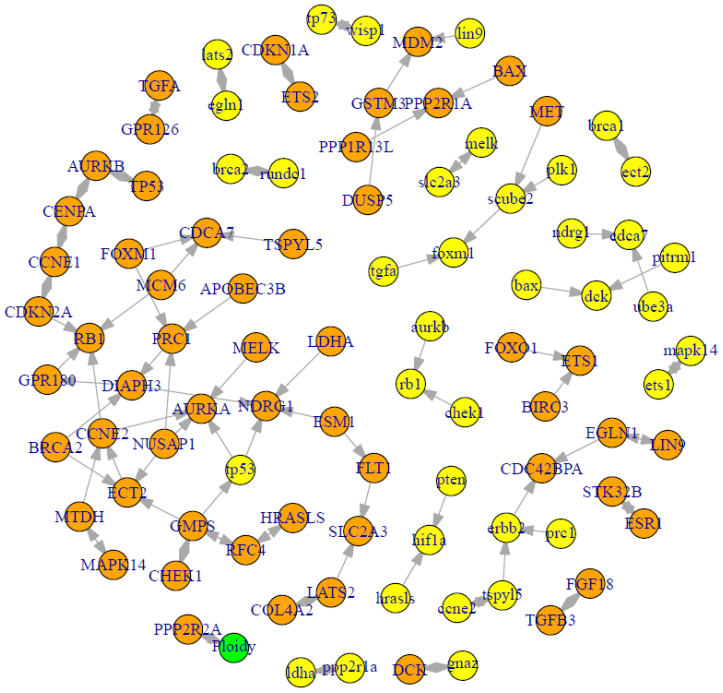
```
Object of class 'pcAlgo', from Call:
pc(suffStat = suffStat, indepTest = disCItest, alpha = 0.06,
    labels = colnames(data_pc))
Number of undirected edges:  24
Number of directed edges:    53
Total number of edges:       77
```

Figure 4- Aperçu du modèle PC avec deux seuils différents

Igraph network :



(a) Alpha=0.01



(b) Alpha=0.06

Figure 5- PC Graph

Le graphe PC est moins dense que celui de HC, il comporte un grand composant connecté et de nombreux petits composants connectés pour alpha=0.01 ou de nombreux petits et grands composants connectés pour alpha=0.06.

Nous remarquons, que pour alpha=0.01 ‘Ploidy’ est affecté par ‘tp53’ et ‘PPP2R2A’ et n’affecte aucun gène, tandis qu’avec un seuil alpha=0.06 il a une relation d’affectation mutuelle avec le gène ‘PPP2R2A’.

Comme nous l’avons vu avec HC, avec PC aussi, ‘tp53’ est plus en relation avec les gènes d’expression que les gènes mutés. Cependant, avec un seuil élevé, nous constatons que sur le graphe (b) nous avons perdu beaucoup d’information sur ‘tp53’ et son lien avec ‘Ploidy’ et plusieurs autres gènes d’expression.

Relations entre les gènes

```
> #Idendify the lower case nodes most related to upper case nodes
> lc_most_related_to_uc(graph_pc)[1:6]
  tp53   ect2   bbc3  egln1  tgfb3 igfbp5
     3      1      0      0      0      0
> #Identify the hubs
> sort(hub_score(graph_pc)$vector, decreasing = TRUE)[1:10]
     tp53      CCNE2      CENPA     NUSAP1     DIAPH3       ESM1       LDHA       ECT2
1.0000000 0.5441412 0.4843813 0.4327383 0.4193660 0.3513923 0.2911834 0.1968021
     RFC4   PPP2R2A
0.1968021 0.1644400
> #Top 10 nodes and edges interms of betweenness centrality measure.
> sort(betweenness(graph_pc), decreasing = TRUE)[1:10]
  PRC1 DIAPH3   GMPS  FOXM1   ECT2  BRCA2 GPR180     RB1   RFC4   LIN9
    93     88     83     76     52     46     28     17     15     11
>
```

(a) Alpha=0.01

```
> #Idendify the lower case nodes most related to upper case nodes
> lc_most_related_to_uc(graph_pc)[1:6]
  tp53   gnaz scube2   lin9  erbb2   bbc3
     3      2      1      1      1      0
> #Identify the hubs
> sort(hub_score(graph_pc)$vector, decreasing = TRUE)[1:10]
    NUSAP1      CCNE2       tp53       GMPS      MELK       MCM6      FOXM1      BRCA2
1.0000000 0.6954017 0.6378840 0.6142035 0.4296175 0.4191077 0.4136042 0.3800852
    CDKN2A     GPR180
0.3289849 0.2657842
> #Top 10 nodes and edges interms of betweenness centrality measure.
> sort(betweenness(graph_pc), decreasing = TRUE)[1:10]
  GMPS DIAPH3  CCNE2   ECT2   PRC1  CENPA   RFC4  CCNE1    tp53  AURKB
  22.0   13.5   12.5   12.5   11.0   10.0   10.0    9.0    8.0    7.0
```

(b) Alpha=0.06

Figure 6- Part 01 : Mutated genes most related to gene expression / Part 02 : Hubs of the PC graph /Part 03 : Top 10 nodes in terms of betweenness centrality measure.

'tp53' est le gène muté le plus en relation avec les gènes d'expression, que ce soit avec un grand ou un petit seuil. Il est également le plus grand 'hub' dans le graphe construit avec un seuil alpha=0.01, mais il perd son importance avec un seuil plus élevé (Figure 6 (b)), et le gène NUSAP1 devient le plus grand 'hub' pour un seuil alpha=0.06.

Nous remarquons aussi que pour alpha=0.01, PRC1 est le premier nœud dans le top 10 des nœuds en termes de centralité d'intermédiarité (betweenness centrality), tandis que GMPS est le premier dans ce classement pour un seuil alpha=0.06. (Figure 6)

	From	To
1	PRC1	DIAPH3
2	FOXM1	PRC1
3	GMPS	FOXM1
4	ECT2	GMPS
5	BRCA2	ECT2
6	DIAPH3	GPR180
7	DIAPH3	BRCA2
8	GPR180	RB1
9	RFC4	GMPS
10	LIN9	BRCA2

(b) Alpha=0.01

	From	To
1	RFC4	GMPS
2	ECT2	CCNE2
3	PRC1	DIAPH3
4	GMPS	tp53
5	GMPS	ECT2
6	DIAPH3	GPR180
7	CCNE2	RB1
8	HRASLS	RFC4
9	CENPA	CCNE1
10	CHEK1	GMPS

(a) Alpha=0.06

Tableau 2 – Top 10 edges in terms of betweenness centrality measure

Dans le top 10 des arêtes en termes de betweenness centrality, nous remarquons qu'il s'agit principalement d'arêtes construites par des gènes d'expression, notamment pour alpha=0.01, aucun gène muté ne figure dans ce classement, nous remarquons aussi pour ce seuil que les gènes **RFC4, GMPS, FOXM1, PRC1, DIAPH3, BRCA2** et **ECT2** forment une chaine. (Tableau 2 (a))

Pour alpha=0.06, les arêtes sont construites par des gènes d'expression à l'exception de tp53 un gène muté qui figure dans l'arête **GMPS -> tp53** qui prend la position 4 du top 10. (Tableau 2 (b))

5. ***Network reconstruction with the MIIC approach***

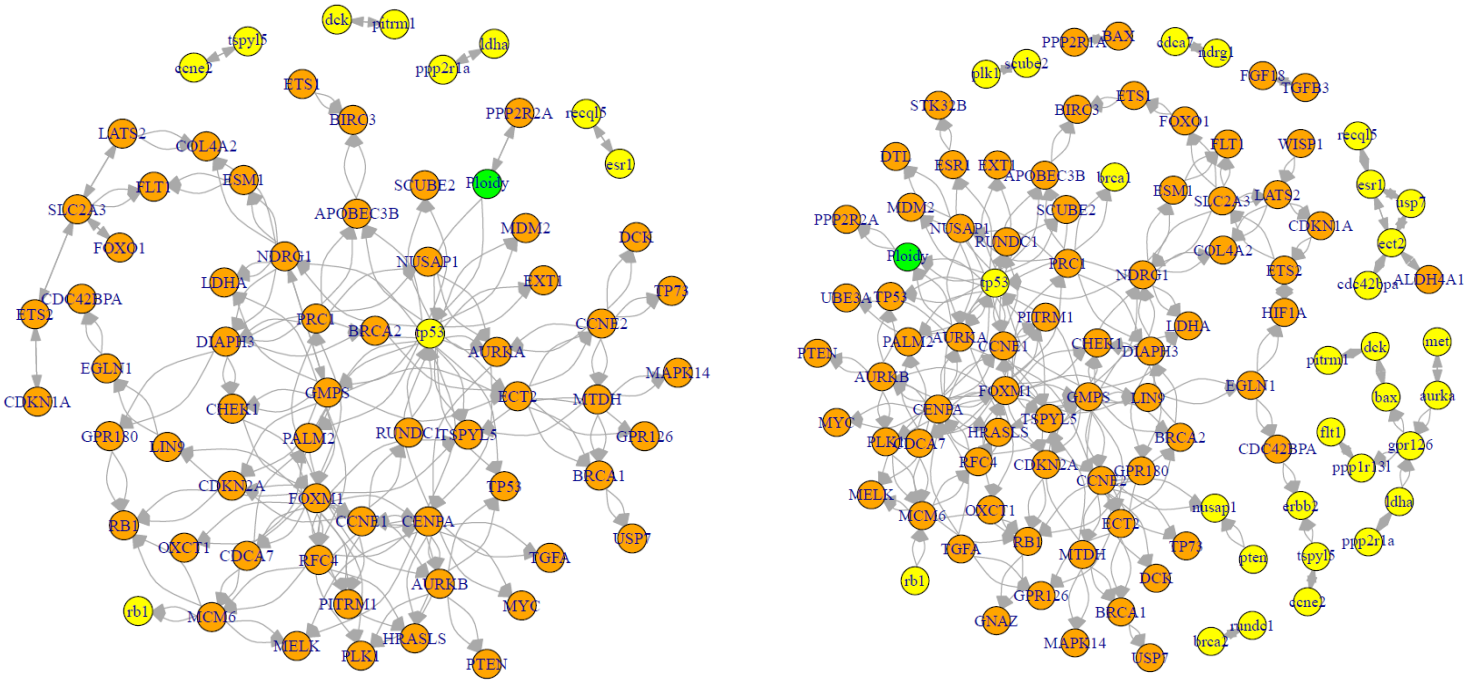
Le modèle miic :

L'algorithme miic est principalement configuré par les deux paramètres suivant :

confidenceShuffle : Le nombre de brassages de l'ensemble de données d'origine afin d'évaluer le rapport de confiance spécifique de tous les bords inférés.

confidenceThreshold : Le seuil utilisé pour filtrer les bords les moins probables suivant le pas du squelette. Voir Verny et al, PLoS Comp. Bio. 2017.

Igraph network:

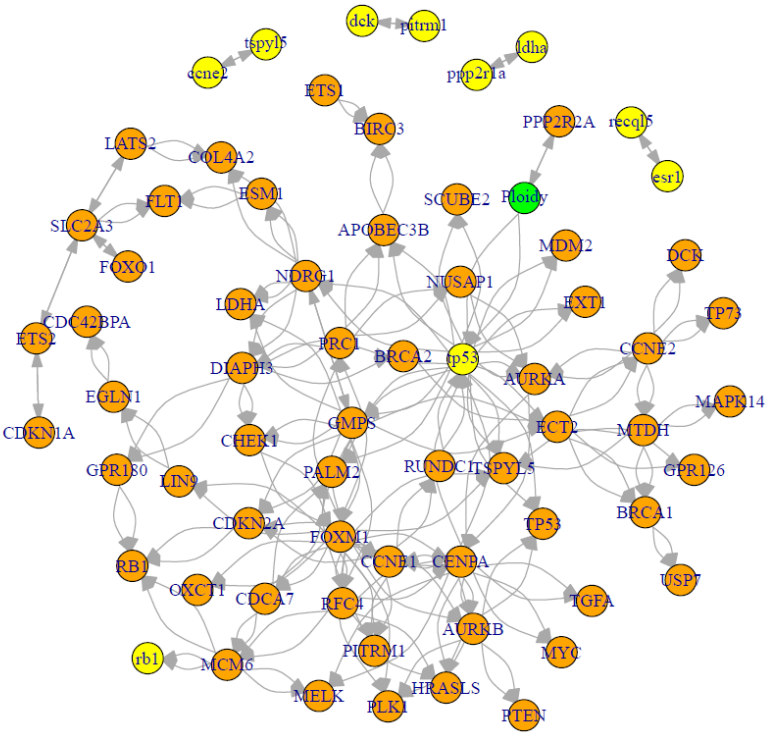


(a) n_shuffles = 100, conf_threshold = 0.001

(b) n_shuffles = 200, conf_threshold = 0.01

Figure 7- miic Graph

Le graphe construit avec miic est moins dense que celui construit avec hc, nous remarquons aussi qu'avec un seuil petit on a moins de liens, ceci apparait sur le graphe (a) (Figure 7) : avec un seuil 0.001, de nombreux gènes mutés se sont dissociés, ce qui fait qu'ils n'apparaissent pas dans le graphe. Le paramètre *n_shuffles* n'a pas d'impact sur le graphe mais plutôt sur la valeur de confiance pour chaque arête.



;

Figure 8 - miic graph (n_shuffles = 500, conf_threshold = 0.001)

Nous avons choisi les seuils n_shuffles = 500, conf_threshold = 0.001. D'après le graphe obtenu, Ploidy affecte tp53 et PPP2R2A et est affecté par PPP2R2A.

Relations entre les gènes

```
> #Idendify the lower case nodes most related to upper case nodes
> lc_most_related_to_uc(graph_miic)[1:6]
 tp53   rb1  bbc3 egln1 tgfb3  esm1
   24     2     0     0     0     0
> #Identify the hubs
> sort(hub_score(graph_miic)$vector, decreasing = TRUE)[1:10]
     FOXM1       tp53      CENPA       GMPS       MTDH      AURKB       RFC4
1.00000000 0.53171751 0.36634579 0.21806955 0.15238168 0.15090267 0.14995757
     CCNE1      AURKA      CCNE2
0.12000251 0.09464386 0.06346216
> #Top 10 nodes and edges interms of betweenness centrality measure.
> sort(betweenness(graph_miic), decreasing = TRUE)[1:10]
     tp53     CENPA      GMPS     FOXM1     AURKA      ECT2     CCNE2      NDRG1
486.33333 379.00000 310.66667 246.50000 227.66667 175.00000 158.00000  97.00000
   DIAPH3     CHEK1
 70.66667  53.33333
```

Figure 9- Mutated genes most related to gene expression / Part 02 : Hubs of the miic graph /Part 03 : Top 10 nodes in terms of betweenness centrality measure.

Nous constatons que 'tp53' joue un rôle important, car, il est lié à 'Ploidy' et à 23 gènes d'expression. Il est aussi le premier nœud en termes de centralité d'intermédiarité (betweenness centrality), par contre le gène FOXM1 est le plus grand 'hub' dans le graphe, suivi par tp53.

	From	To
1	CENPA	tp53
2	CENPA	tp53
3	tp53	GMPS
4	tp53	GMPS
5	AURKA	CENPA
6	AURKA	CENPA
7	ECT2	CCNE2
8	ECT2	CCNE2
9	GMPS	FOXM1
10	GMPS	FOXM1

Tableau 3 – Top 10 edges in terms of betweenness centrality measure

Nous remarquons que tp53 apparait dans les 4 premières arêtes en termes de betweenness centrality, confirmant par ceci les résultats obtenus par les graphes sur l'importance de ce gène.

Par ailleurs, nous remarquons aussi dans ce classement qu'il s'agit principalement d'arêtes construites par des gènes d'expression.

6. Conclusion

Chaque approche donne un graphe particulier, mais dans les trois, le gène tp53 est le plus important : le plus en relation avec les gènes d'expression, et aussi le plus grand hub. De plus, les approches MIIC et PC ont des paramètres qui permettent le contrôle des arêtes dans le but de créer un graphe plus ou moins dense et mieux interprétable.

Cependant, le manque de connaissances dans le domaine de la biologie a rendu difficile l'évaluation de la qualité des algorithmes.