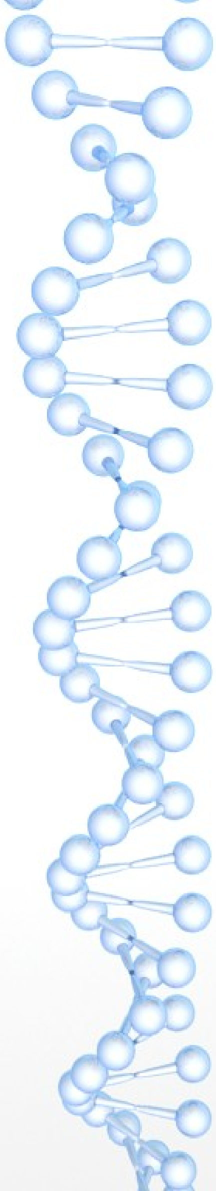
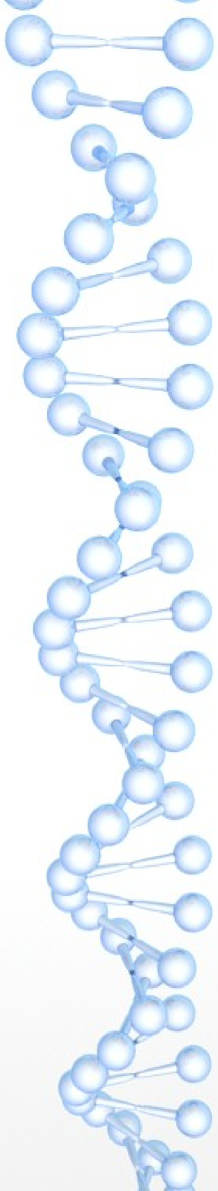


Short report progress

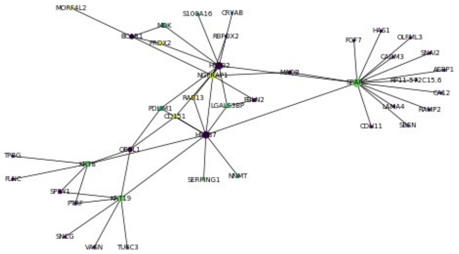
How to find genes network into single cell RNA data

By Romuald MARIN

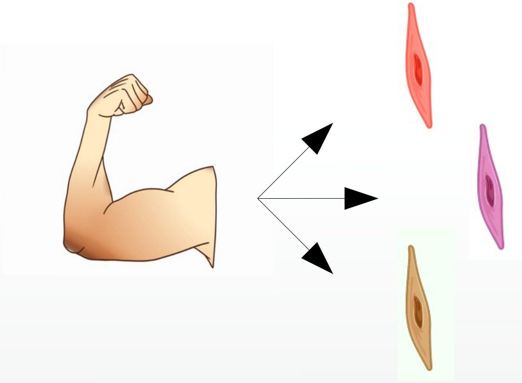


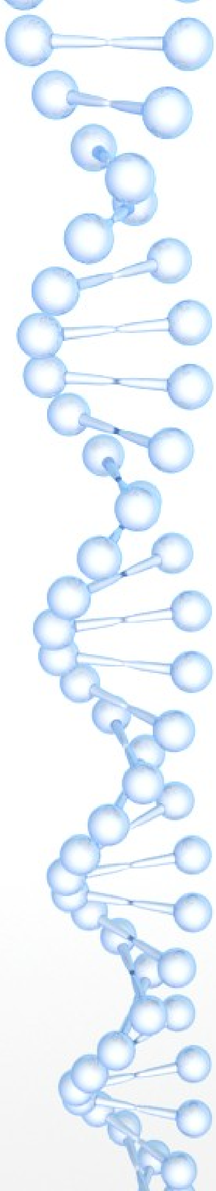
Why do we want to find genes network?

- Genes network :
  - Groupe of genes which interact with each other
  - Same biological function
  - Transcription factor gene



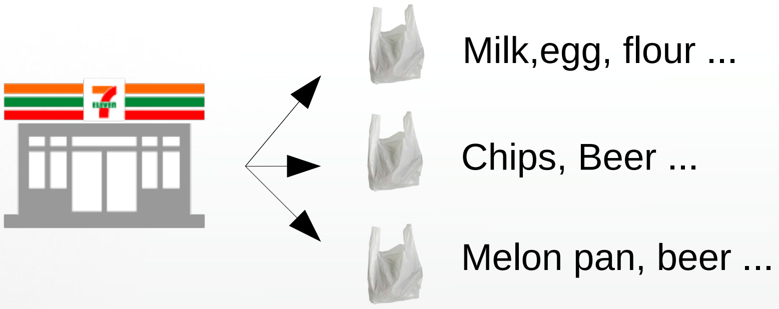
- Single cell RNA data :
  - Cellular level study
  - Big variation
  - Non-expressed gene





My approach?

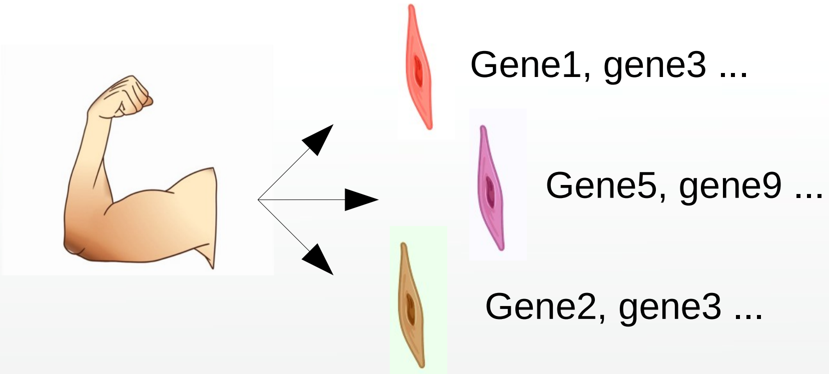
- Association rules learning :
  - Discovering interesting relations between variables in large databases
  - Discovering genes relation into count matrix expression
- Apriori Algorithm
  - Find objects bought together
  - Find relations between differents items



Milk,egg, flour ...

Chips, Beer ...

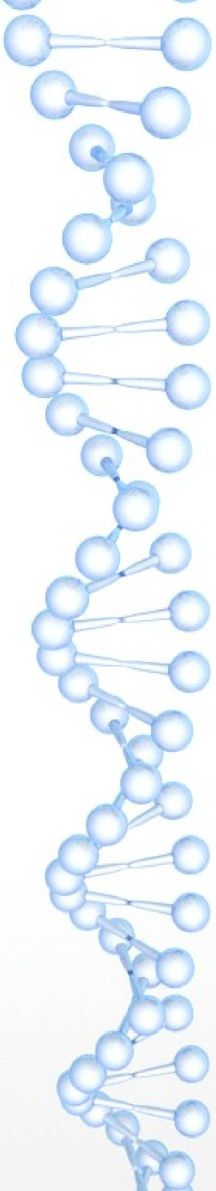
Melon pan, beer ...



Gene1, gene3 ...

Gene5, gene9 ...

Gene2, gene3 ...



How does it work ? (1)

	A1BG	ACTB	A1CF	...	FAT1
Cell 1	0	364	1014	...	40
Cell 2	0	909	0	...	0
Cell 3	0	501	590	...	2046
Cell 4	0	107	0	...	0

Count Matrix with expressed genes

→

	ACTB	A1CF	...	L3HYPDH
Cell 1	True	True	...	False
Cell 2	True	False	...	False
Cell 3	True	True	...	True
Cell 4	True	False	...	False

Boolean Matrix with expressed genes

Normalize and bool :

- Remove no expressed gene
- Remove too small count

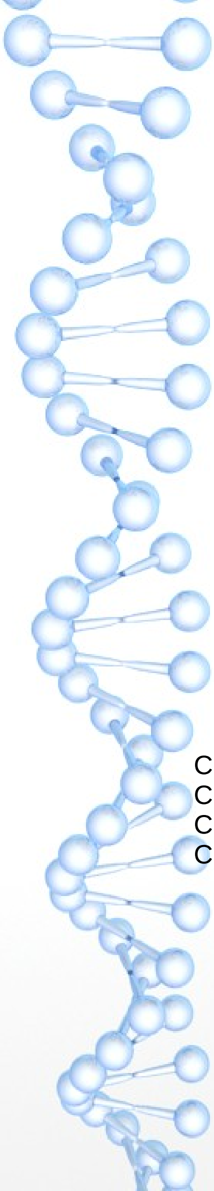
Support : Fraction of cells who expressed genes (or a group of genes)

Confidence : Fraction of gene expressed if an other gene (or a group of genes) is expressed

Lift : Measure of correlation between 2 genes or groups of genes

Support(A1CF) = 2/4 = 1

Confidence(ACTB → A1CF) =
support(ACTB,A1CF)
/support(ACTB)
= 0.5



How does it work ? (2)

1st step : Define threeshold

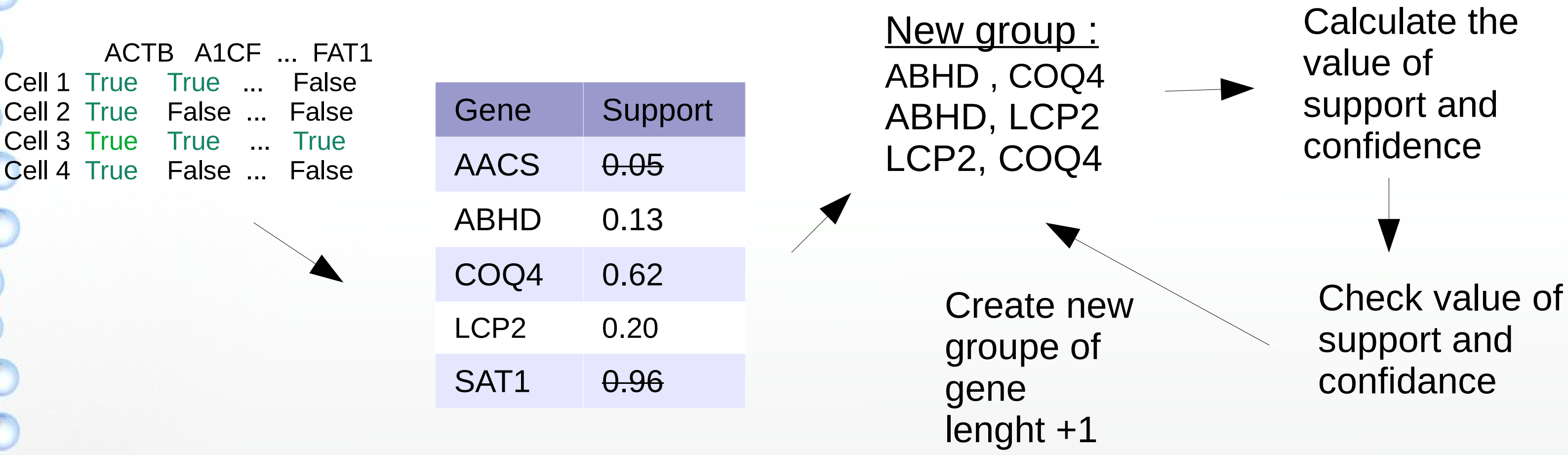
2nd step : Find gene who pass different threshold

3rd step : Create group of gene of lenght L+1 and check values

4th step : Repeat until lengh threeshold

Threeshold

- Min Support = 0,1
- Max Support = 0,7
- Min Confidence = 0.9
- Lenght = 3



	ACTB	A1CF	...	FAT1
Cell 1	True	True	...	False
Cell 2	True	False	...	False
Cell 3	True	True	...	True
Cell 4	True	False	...	False

Gene	Support
AACS	~~0.05~~
ABHD	0.13
COQ4	0.62
LCP2	0.20
SAT1	~~0.96~~

New group :

ABHD , COQ4

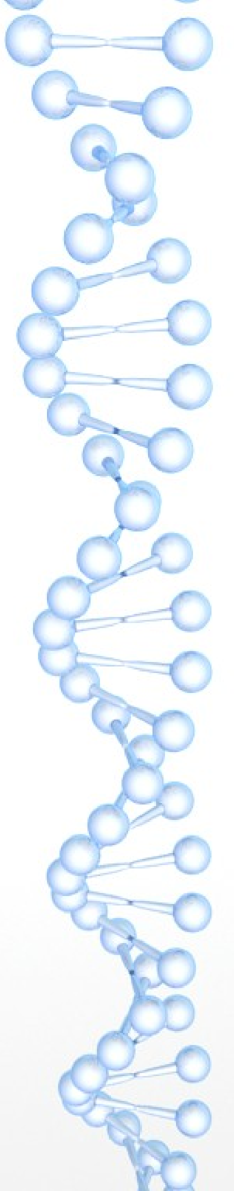
ABHD, LCP2

LCP2, COQ4

Calculate the value of support and confidence

Check value of support and confidance

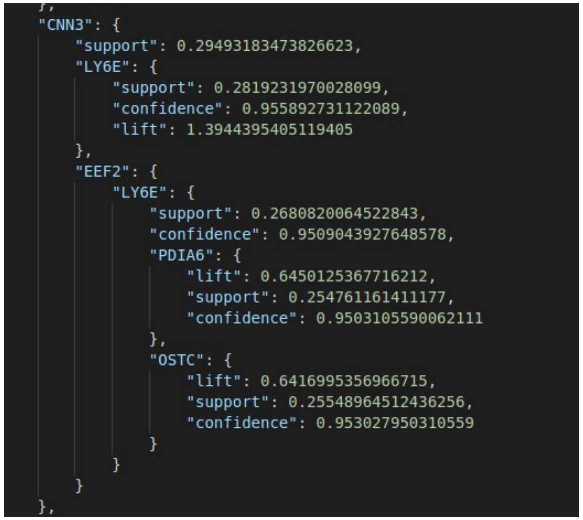
Create new groupe of gene lenght +1



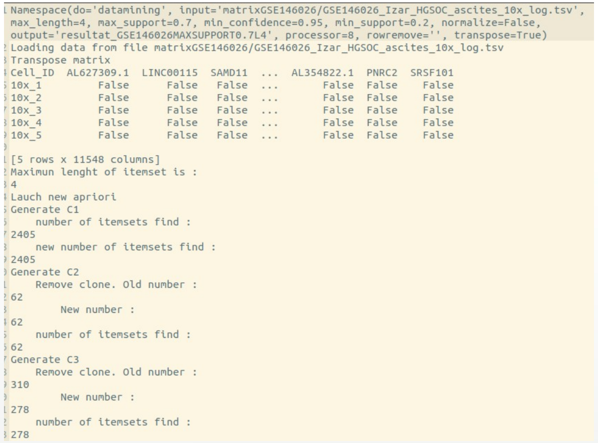
Output

Differents files are created :

- results json format
- List in txt format
- Log file
- Network representation

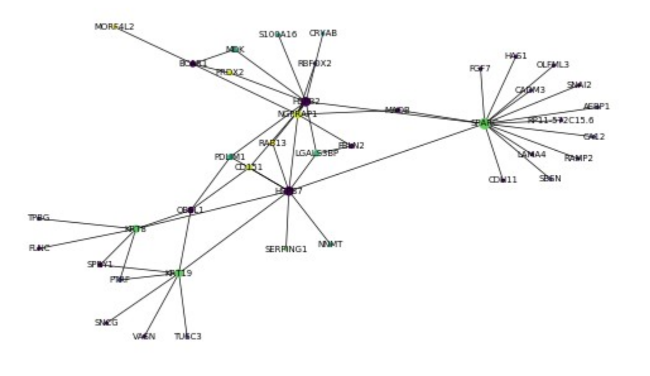


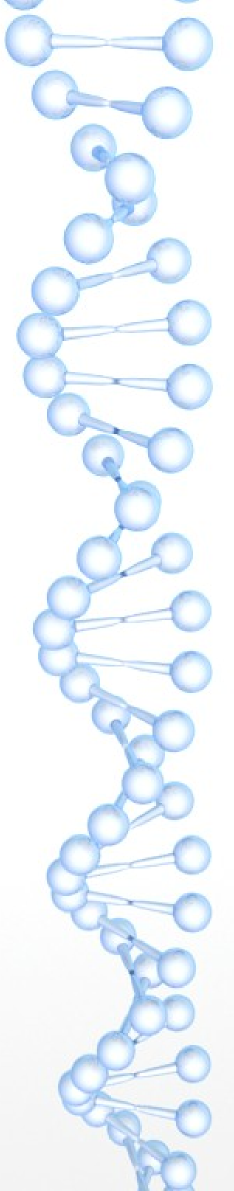
```
    },
    "CNN3": {
        "support": 0.29493183473826623,
        "LY6E": {
            "support": 0.2819231970028099,
            "confidence": 0.955892731122089,
            "lift": 1.3944395405119405
        },
        "EEF2": {
            "LY6E": {
                "support": 0.2680820064522843,
                "confidence": 0.9509043927648578,
                "PDIA6": {
                    "lift": 0.6450125367716212,
                    "support": 0.254761161411177,
                    "confidence": 0.9503105590062111
                },
                "OSTC": {
                    "lift": 0.6416995356966715,
                    "support": 0.25548964512436256,
                    "confidence": 0.953027950310559
                }
            }
        }
    },
```



```
Namespace(do='datamining', input='matrixGSE146026/GSE146026_Izar_HGSOC_ascites_10x_log.tsv',
max_length=4, max_support=0.7, min_confidence=0.95, min_support=0.2, normalize=False,
output='resultat_GSE146026MAXSUPPORT0.7L4', processor=8, rowremove='', transpose=True)
Loading data from file matrixGSE146026/GSE146026_Izar_HGSOC_ascites_10x_log.tsv
Transpose matrix
Cell_ID  AL627309.1  LINC00115  SAMD11  ...  AL354822.1  PNRC2  SRSF101
10x_1         False      False   False  ...       False  False    False
10x_2         False      False   False  ...       False  False    False
10x_3         False      False   False  ...       False  False    False
10x_4         False      False   False  ...       False  False    False
10x_5         False      False   False  ...       False  False    False

[5 rows x 11548 columns]
Maximun lenght of itemset is :
4
Lauch new apriori
Generate C1
    number of itemsets find :
2405
    new number of itemsets find :
2405
Generate C2
    Remove clone. Old number :
62
        New number :
62
    number of itemsets find :
62
Generate C3
    Remove clone. Old number :
310
        New number :
278
    number of itemsets find :
278
```





Importance of different thresholds

Support threshold:

- Maximum support allow to remove genes who expressed in to many cell (default=0.8)
- Minimum support allow to discovered low frequent network (default=0.3)

Confidence and lift

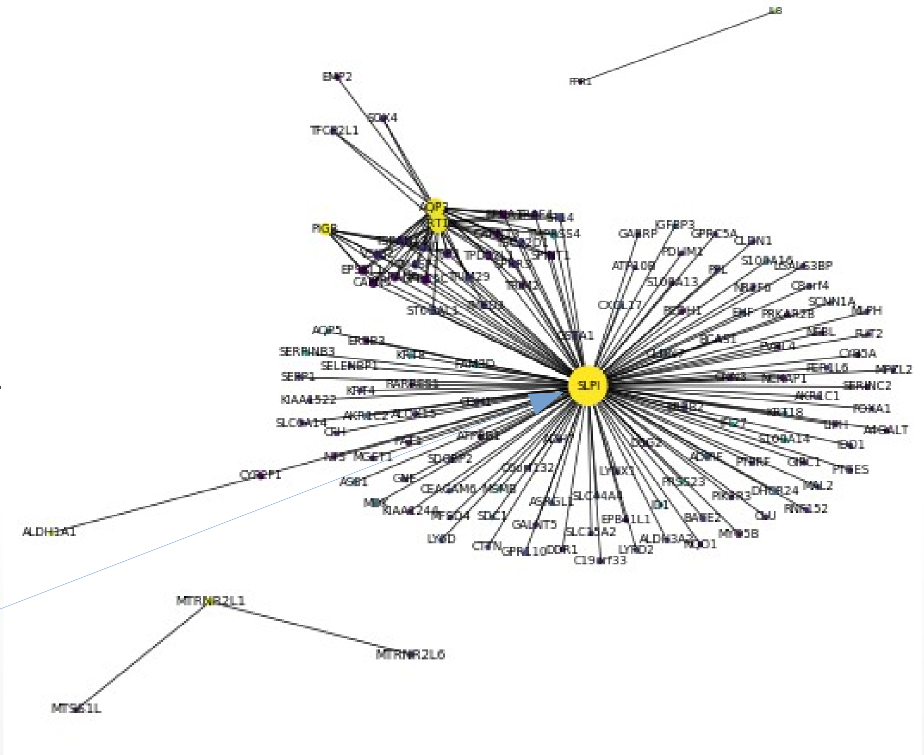
- allow to keep only gene network interesting (default confidence=0.9 & lift > 1)

Lenght

- Reduce compute time (default=4)

Results

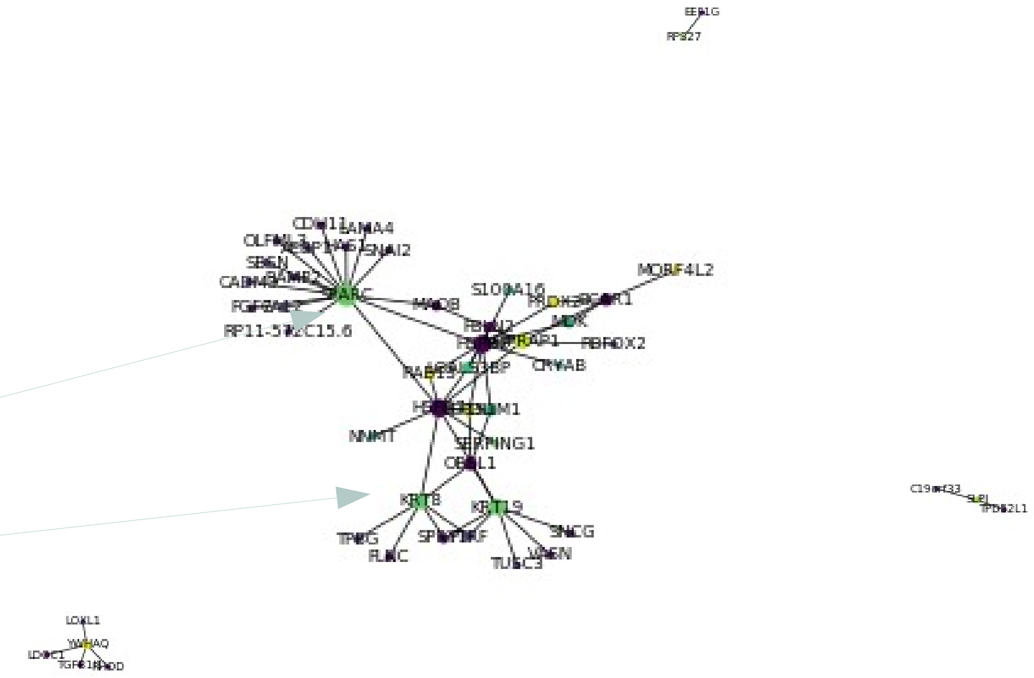
- Nasal epithelial cells from mouse (GSE148829)
- Threshold :
  - max_length=3 , min_confidence=0.95
  - min_support=0.2 , max_support=0.8
- Results :
  - 1394 genes
  - 440 couple of 2 genes
  - 808 groups of 3 genes
- Exemple of relation :
  - Interleukine 8 & Formyl Peptide Receptor 1
    - host defense and inflammation
  - Aquaporin 3 & Keratin 19 & Aquaporin TIP3-1
    - Inflamation response
  - SPL1 : Transcription factor

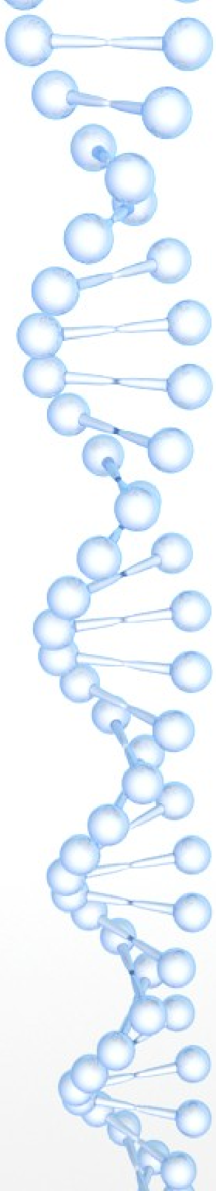


Source : uniprot

Results

- Ovarian cancer ascites from human (GSE146026)
- Threshold :
  - max_length=4 , min_confidence=0.95
  - max_support=0.5, min_support=0.1
- Results :
  - 4260 genes
  - 201 couple of genes
  - 1843 goups of 3 genes
  - 18457 groups of 4 genes
- Exemple of relation :
  - Carbonic Anhydrase 12 & SPARC
    - Carcinomas and Induces Apoptosis in Ovarian Cancer Cells
  - Keratin 19 & Caveolae Associated Protein 1 & Keratin 8
    - Structure of caveole





Conclusion

ADVANTAGE

- Easy to use
- Very fast compare to other method
- Print graph and different results files
- Allow to study different frequence with threshold

DISADVANTAGE

- Only up regulated genes network are find
- More accurate methods are available