

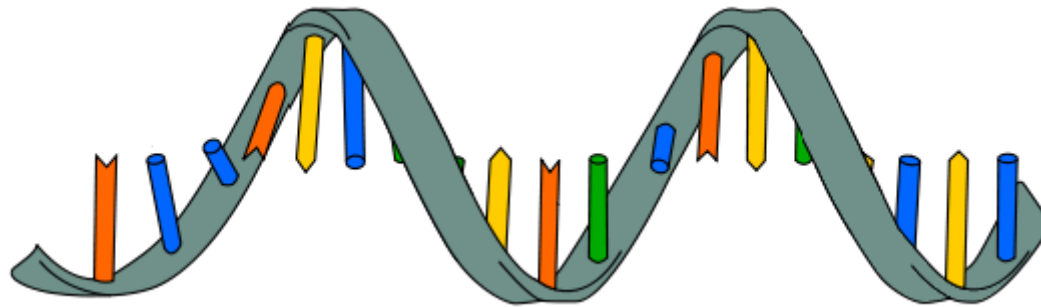
The background features abstract, overlapping green geometric shapes, primarily triangles and polygons, in various shades of green, creating a modern and dynamic visual effect.

BBC

Test miARN pour diagnostiquer le cancer de la prostate à partir de l'urine

Rappel - But du projet

- Comprendre le contexte (prostate et micro-ARN)
- Comprendre les données micro-arrays reçues
- Développer un modèle statistique de diagnostic avec les données reçues



Démarche d'analyse



- ▶ Analyse des fichiers *matrix* et *soft*
- ▶ Extraction et formatage des données
- ▶ Etude de la corrélation entre Sains et PCa
- ▶ Sélection utilisant la p-value de chaque corrélation
 - ▶ Essais avec le t-test -> inutile dans notre cas
- ▶ Entraînement des modèles -> résultats moyens
- ▶ Amélioration du set en testant les micro-ARN 2 par 2
 - ▶ Les modèles donnent des résultats bien mieux

Analyse des fichiers

- ▶ Fichier *matrix*
 - ▶ 60 patients
 - ▶ Une dizaine de patients sains
 - ▶ Niveau d'expression de plus de 20'000 micro-ARN
 - ▶ Données principales étudiées
- ▶ Fichier *soft*
 - ▶ Intensité observée sur chaque micro-ARN
 - ▶ P-value des valeurs présentes
 - ▶ Utilisé à des fins de vérification

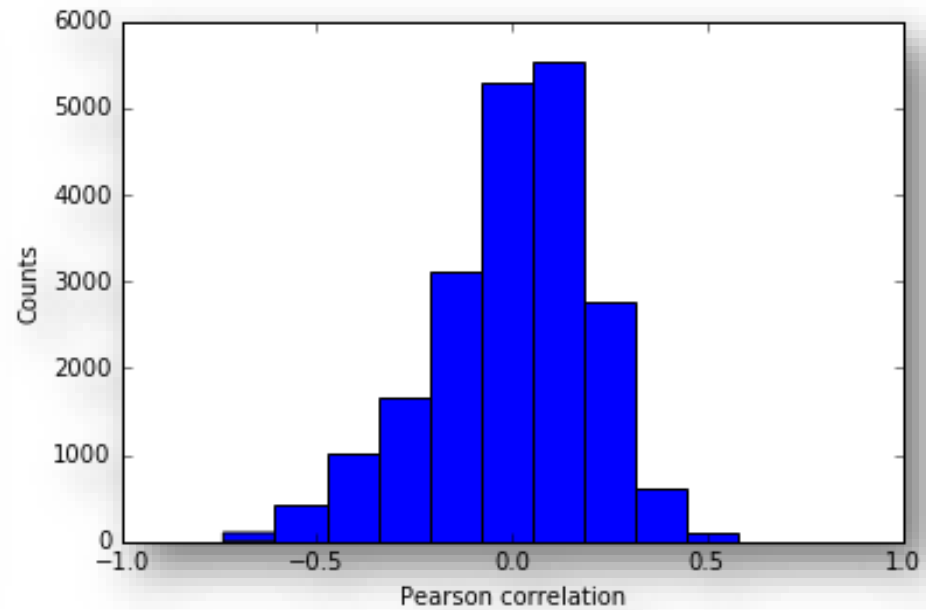
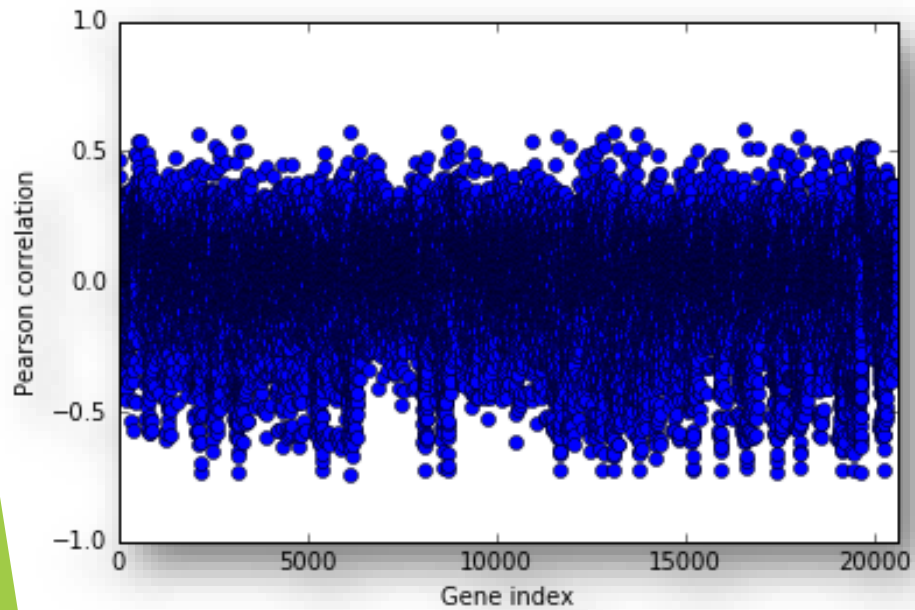


Extraction et formatage des données

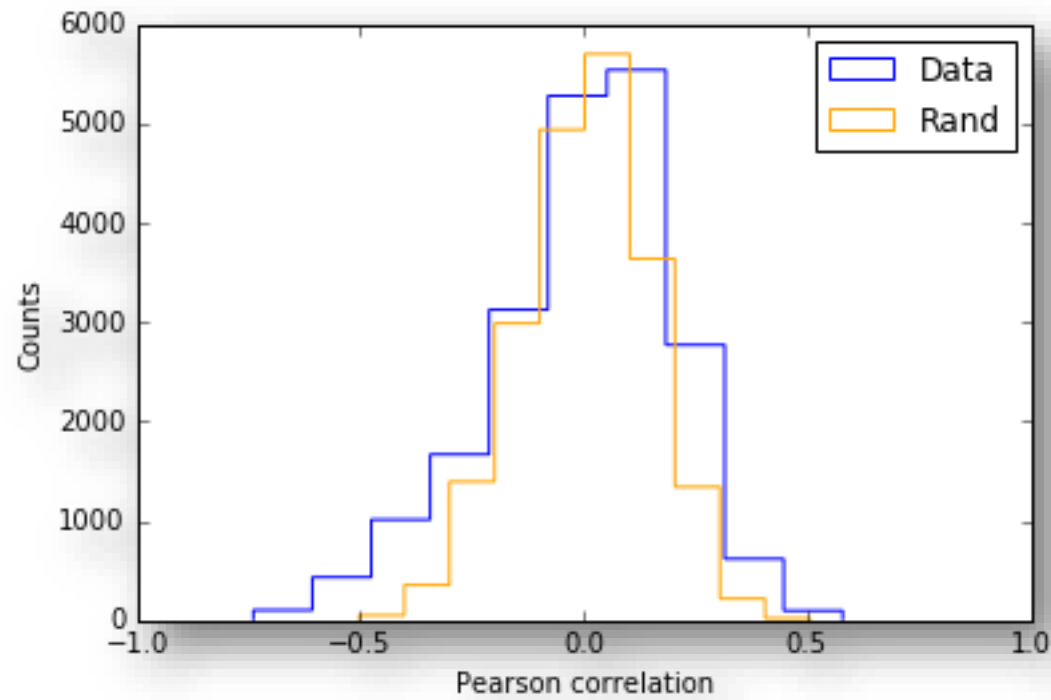
- ▶ Classification des types de patients (malade ou non)
- ▶ Séparer les données en test-set et train-set

Corrélation

- Calcul de la corrélation entre les données et un profil désiré

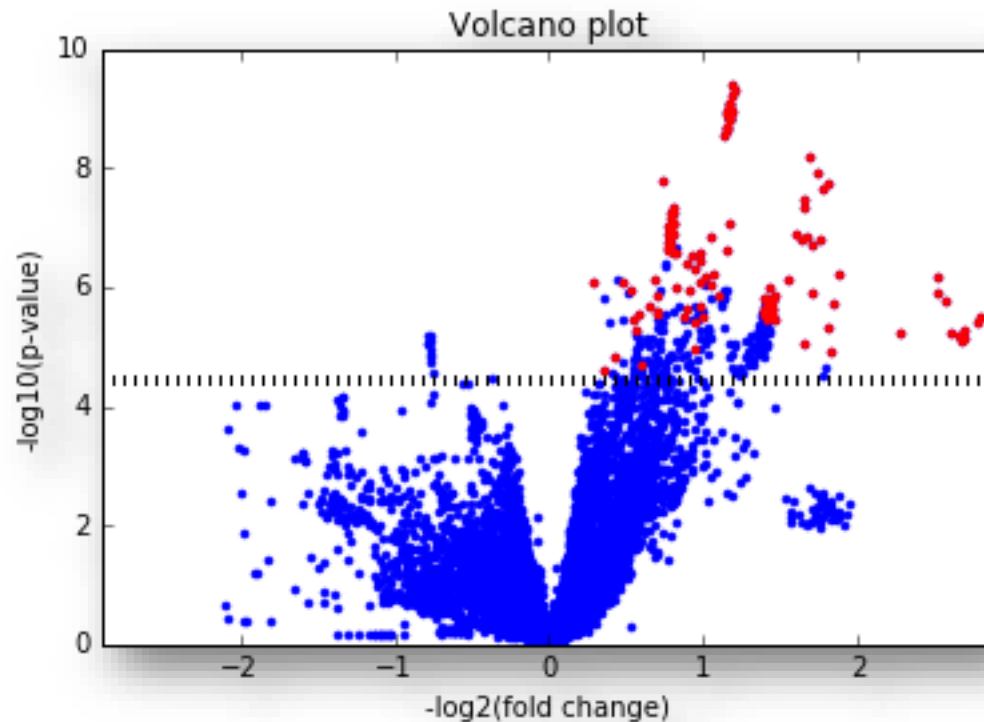


Comparaison avec un set aléatoire



Sélection des micro-ARN

- Nous choisissons ceux dont la corrélation est la plus significative
 - P-value la plus petite -> moins de chance d'être dû au hasard



Entrainement des modèles

► Tests avec plusieurs classificateur

- kNN 0.71%
- SVM 0.71%
- Random Forest 0.71%

Résultats très moyens, peut mieux faire !?

Amélioration du data set

- ▶ On calcule avec kNN les scores de réussite avec chaque paire des 60 micro-ARN précédemment sélectionnés
- ▶ On prend les micro-ARN qui ont la somme des scores la plus élevée

Résultats

- ▶ On obtient une nette amélioration tout en réduisant le nombre de micro-ARN nécessaires à 20
 - ▶ kNN: 85%
 - ▶ SVM: 71%
 - ▶ Random Forest: 85%

Seul SVM ne s'améliore pas, les autres donnent des résultats stables à 85%.

Vérification des résultats

- ▶ Dummy estimator
 - ▶ Avec des micro-ARN sélectionnés au hasard, nous obtenons environ 73% de résultats corrects
 - ▶ Si on prend le maximum de score réalisé par le dummy estimator on obtient des résultats environ entre 20% et 100% -> **c'est là qu'on se sent mal**
- ▶ Fichier soft
 - ▶ Contient des p-values différentes
 - ▶ En sélectionnant les micro-ARN d'après ces valeurs, on obtient des valeurs inférieures à nos précédent tests

Conclusion...

- ▶ Nous ne sommes pas sûr d'avoir trouvée quoi que ce soit de significatif étant donné qu'un dummy estimator donne des 100%
- ▶ Soit c'est un véritable hasard sans réelle corrélation, soit nos calculs ne sont pas totalement corrects



... ou pas

- ▶ En voyant ces résultats on s'est dit qu'on pouvait bien changer quelque chose
 - ▶ On ne met que des malades dans le test-set
- ▶ Conséquences
 - ▶ Les calculs utilisent le double de micro-ARN (126 au lieu de 60)
 - ▶ Tous les classificateurs donnent 100% de résultats juste avant même d'avoir comparé les micro-ARN par paires
 - ▶ Après réduction du set donnée à 20 micro-ARN seul kNN ne donne que 87% de juste
- ▶ Le dummy estimator donne toujours des très bon résultats (95% de résultats justes)

Conclusion

- ▶ Il est difficile de travailler avec ces données
 - ▶ Beaucoup de dépendance à la construction du test-set
 - ▶ On obtient de meilleurs résultats en ne mettant que des malades pour les tests
 - ▶ Les données aléatoires donnent d'excellents résultats
- ▶ Les bases mathématiques sont nouvelles pour nous
 - ▶ On se perds vite dans les différentes notions
 - ▶ Difficile de savoir quoi faire pour améliorer les résultats

Questions ?

