Utilisation des méthodes à noyaux en vue de l'analyse des données biologiques.

Xavier Grand : Formation continue Master BCD

17 Juillet 2020

Tuteurs scientifiques : Sébastien Déjean,
Jérôme Mariette,
Tuteur pédagogique : Thérèse Commes

MASTER SCIENCES ET NUMERIQUE POUR LA SANTÉ

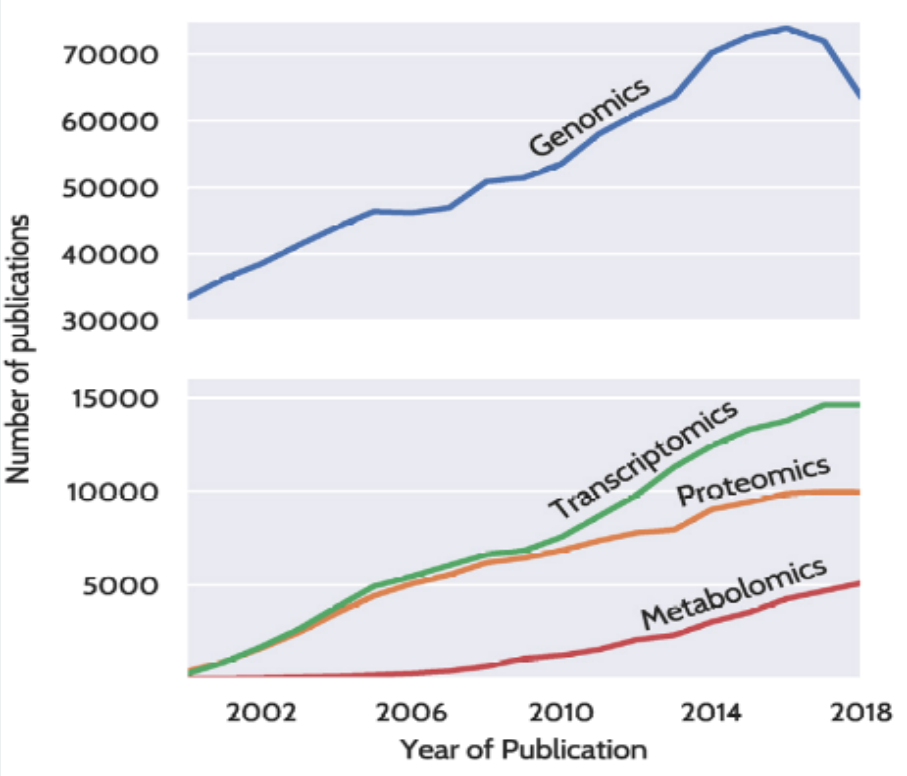
FACULTÉ DES SCIENCES MONTPELLIER

INRAE

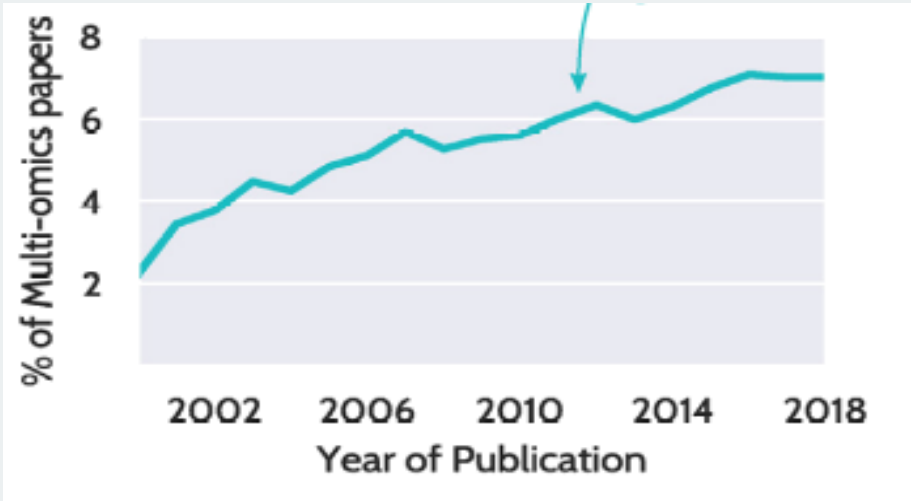
INSTITUT de MATHEMATIQUES de TOULOUSE

INTRODUCTION

Evolution de la publication de données omiques entre 2000 et 2018



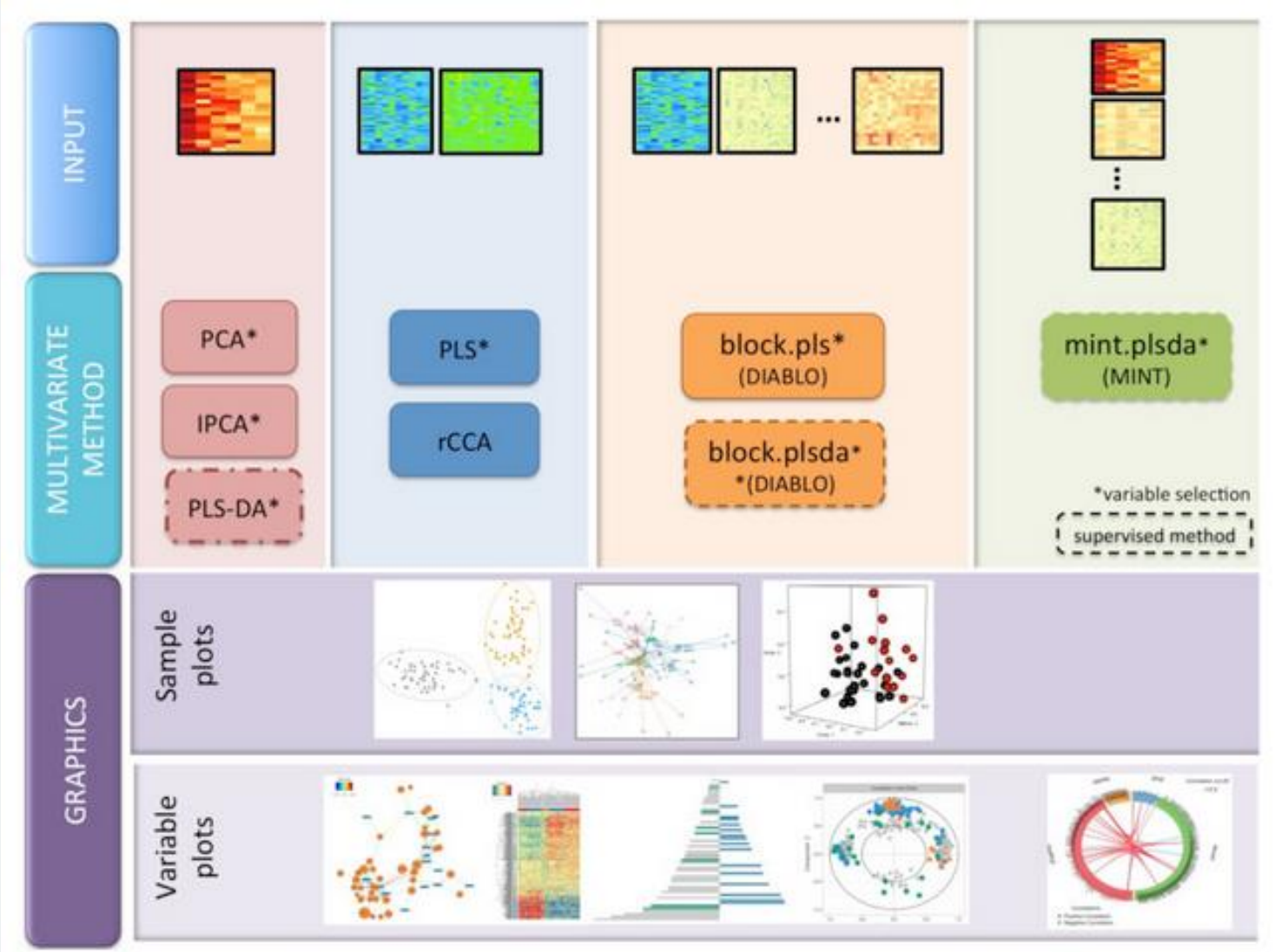
Etudes multi-omiques incluant des données de transcriptomiques, de protéomique et de métabolomique.



Noor *et al.*, 2019, Biological Insights through Omics Data Integration.

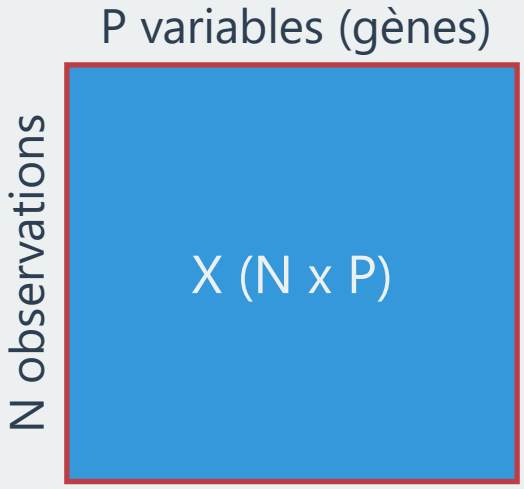
MASTER SCIENCES
ET NUMERIQUE POUR LA SANTÉ

Analyses multi-omiques : "mixOmics"



Analyses multi-omiques : "mixOmics"

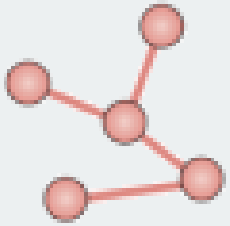
mixOmics offre un cadre d'analyse complet pour des données numériques…
Sous forme de tableau X avec N observations sur P variables.



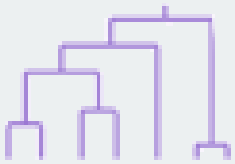
Données biologiques sous d'autres formes

Matrices de génotypages

Réseaux



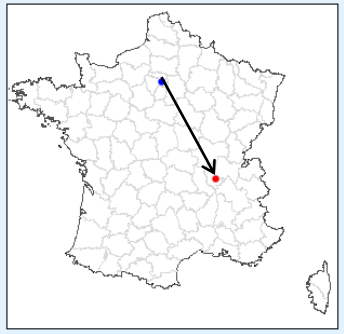
Arbres phylogénétiques



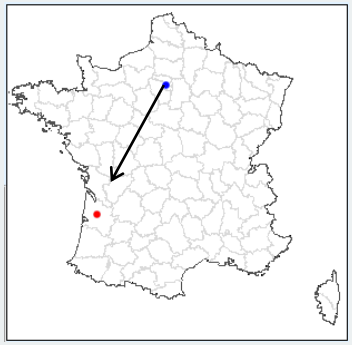
?

Méthodes à noyaux, calcul de similarité

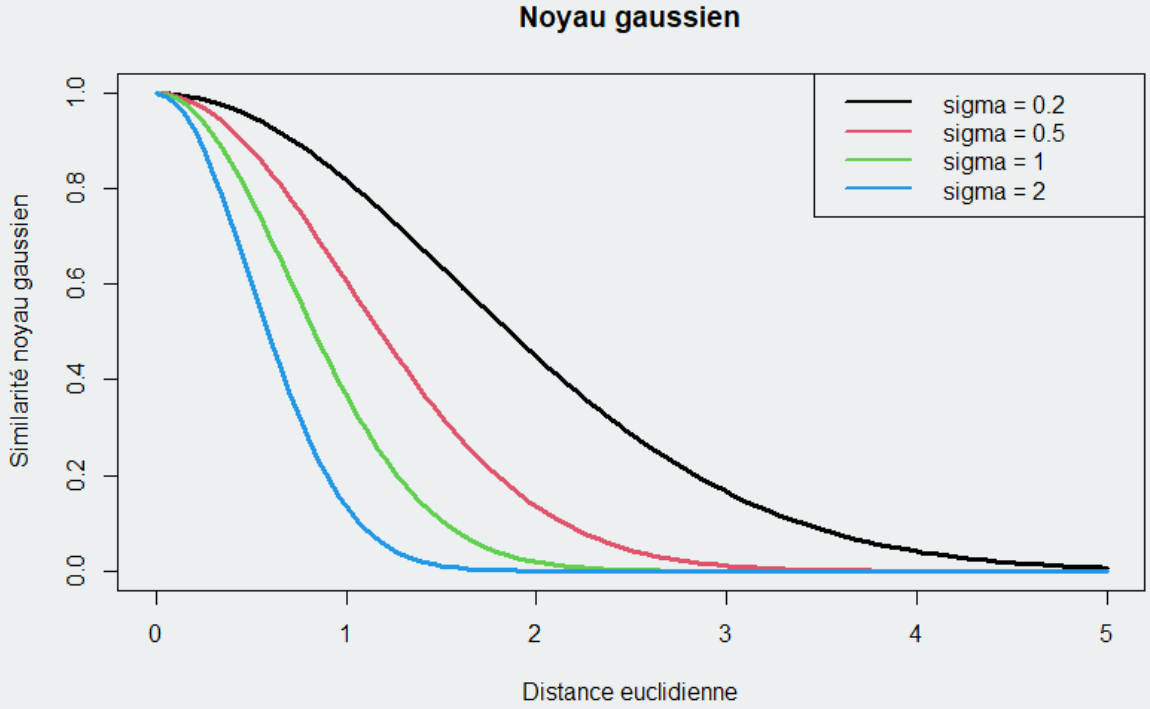
Paris-Lyon :
Orthodromie : 392 km



Paris-Bordeaux :
Orthodromie : 498 km



Exemple de noyau : le noyau Gaussien

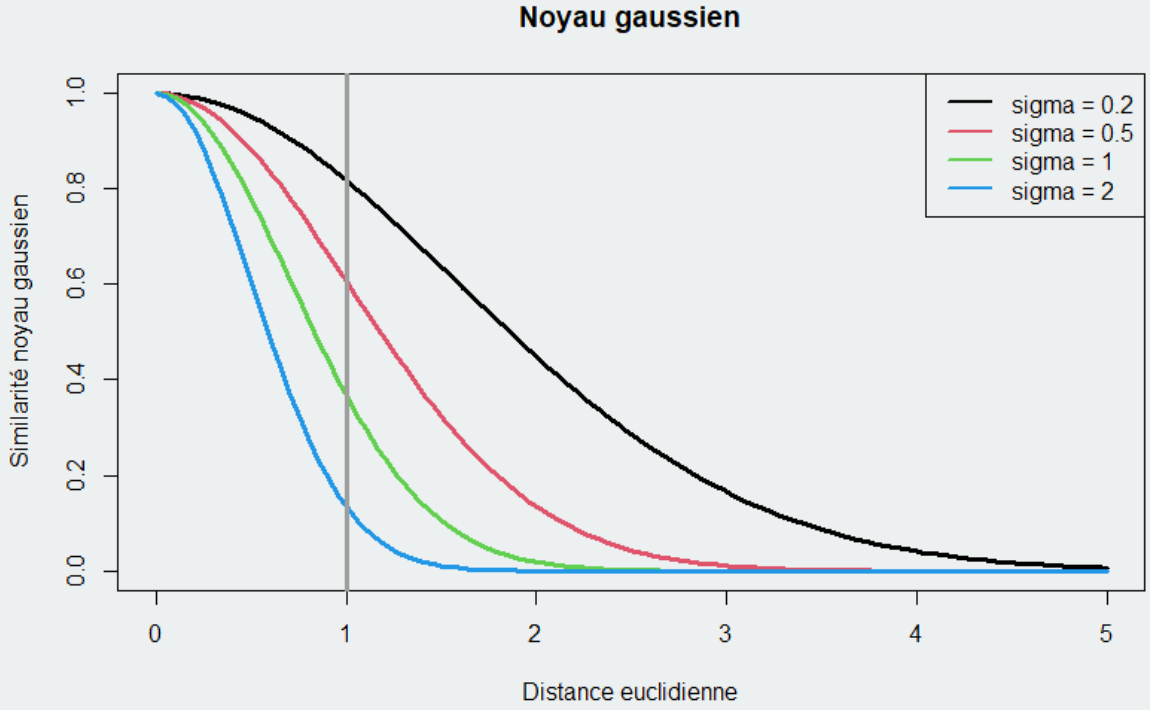


$K(x, x') = \exp^{-\sigma||x-x'||^2}$ avec x et x′ défini dans un espace arbitraire X.

Effet du paramètre sigma : Plus sigma est élevé, plus vite la similarité diminue avec la distance euclidienne.

Exemple d'application : exclure des "outliers".

Exemple de noyau : le noyau Gaussien

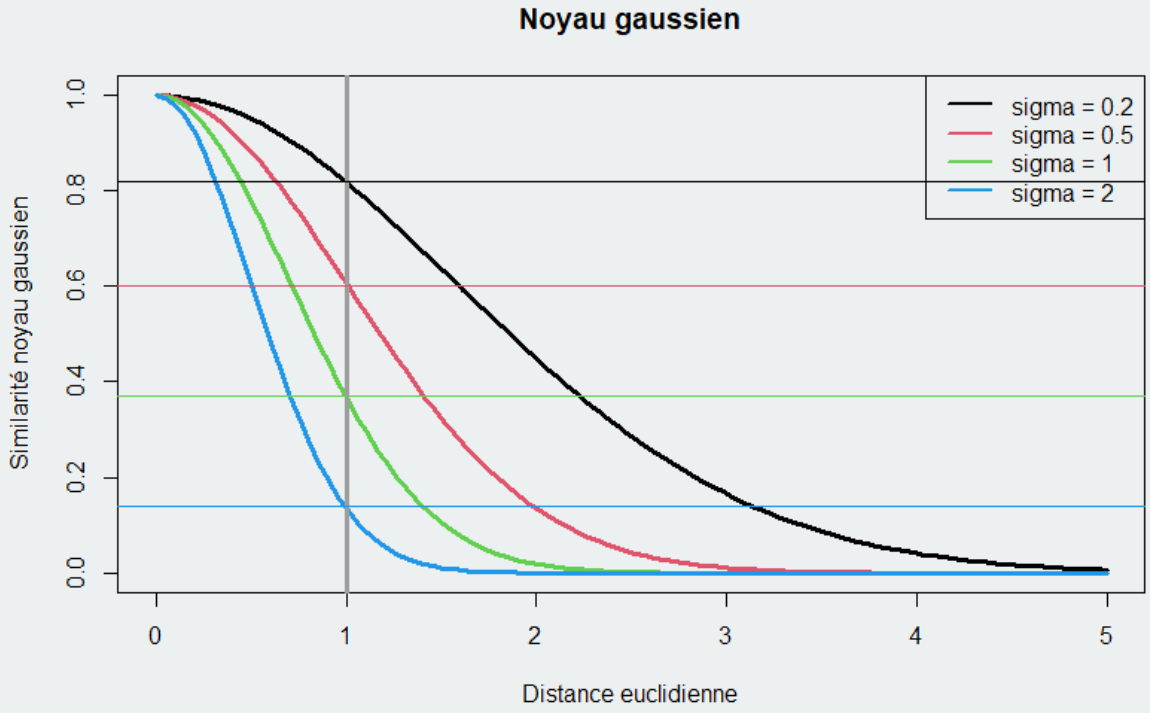


$K(x, x') = \exp^{-\sigma||x-x'||^2}$ avec x et x′ défini dans un espace arbitraire X.

Effet du paramètre sigma : Plus sigma est élevé, plus vite la similarité diminue avec la distance euclidienne.

Exemple d'application : exclure des "outliers".

Exemple de noyau : le noyau Gaussien



$K(x, x') = \exp^{-\sigma||x-x'||^2}$ avec x et x′ défini dans un espace arbitraire X.

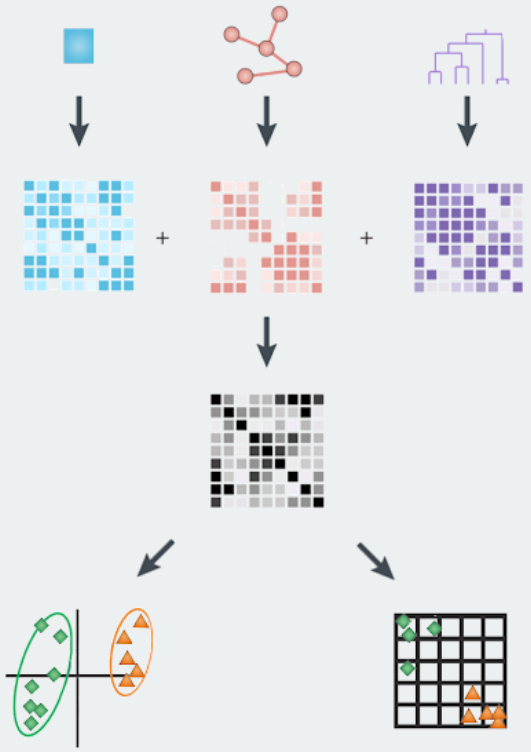
Effet du paramètre sigma : Plus sigma est élevé, plus vite la similarité diminue avec la distance euclidienne.

Exemple d'application : exclure des "outliers".

Objectifs et problématique :

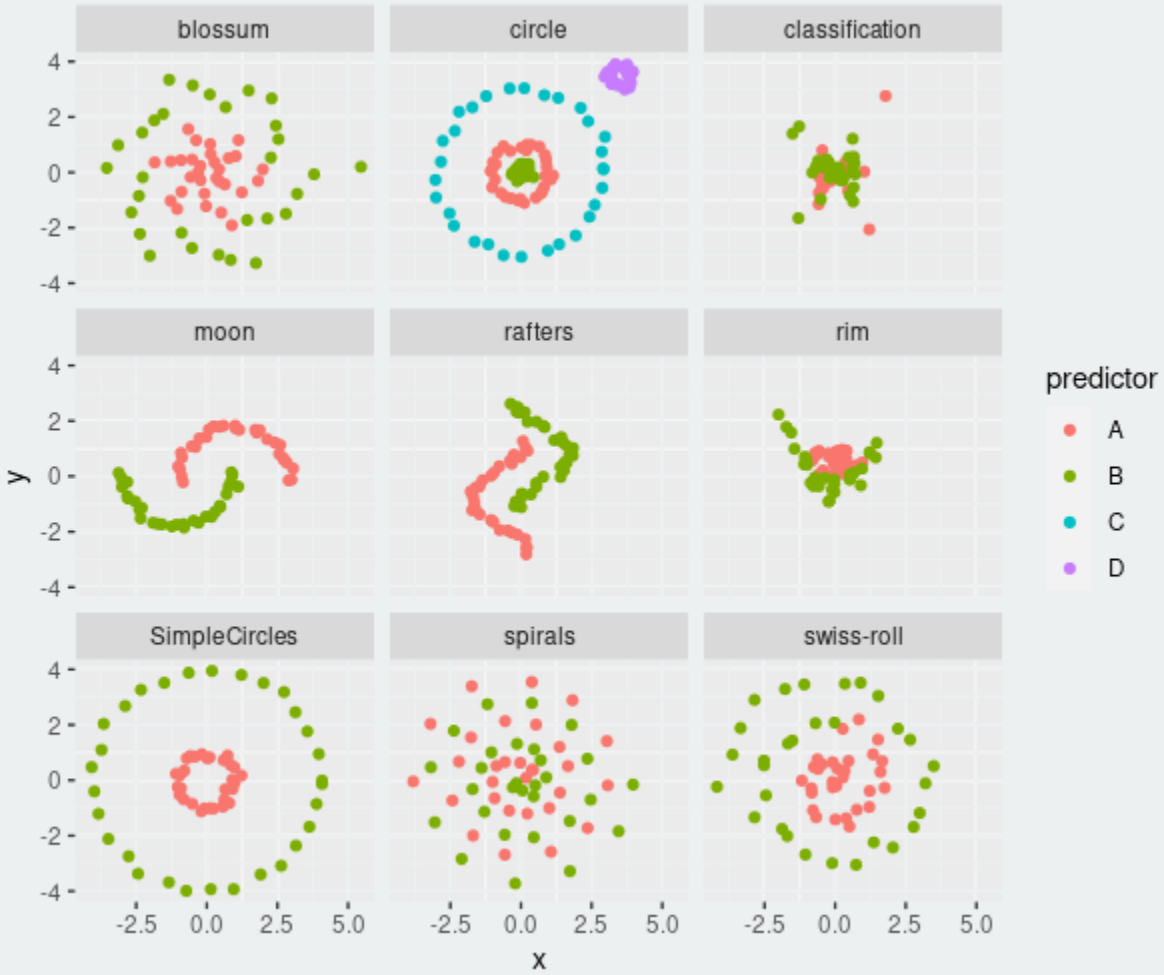
Comment utiliser les méthodes à noyau pour l'analyse et l'intégration de données de biologie non numériques ?

- Développer des méthodes à noyaux adaptés aux données biologiques non numériques.
- Transformer avec ces noyaux les données en matrices de similarités.
- Intégrer ces matrices de similarités afin de prendre en considération toute l'information biologique dans l'analyse.



Jeux de données simulés classiques

Simulation de jeux de données simples afin de se familiariser avec les méthodes à noyaux.

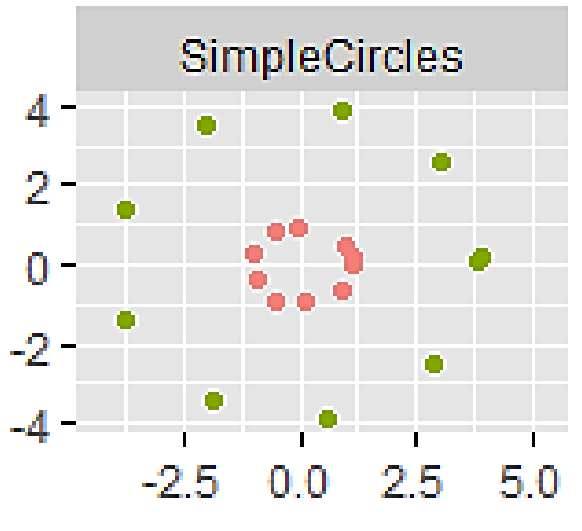


Matrice de similarités

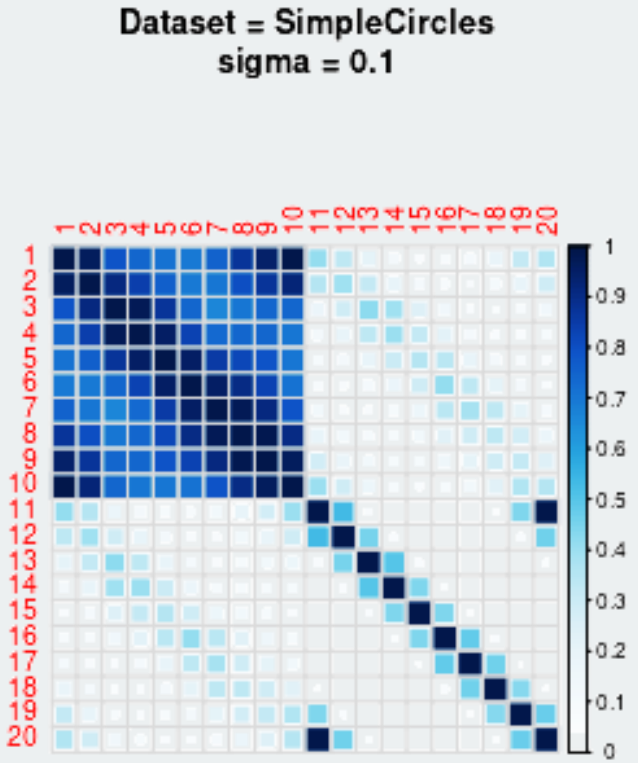
Utilisation du ‘‘package’’ R kernlab.

Karatzoglou *et al.* (2004). “kernlab – An S4 Package for Kernel Methods in R.” *Journal of Statistical Software*, **11**(9), 1–20.

Jeu de données original



Noyau Gaussien

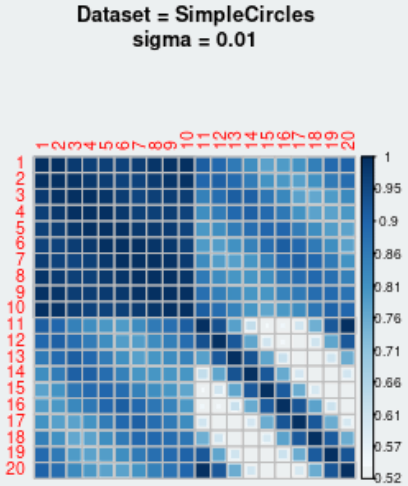


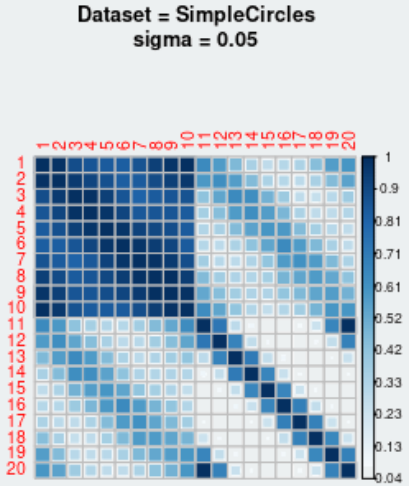
A : observations 1 à 10

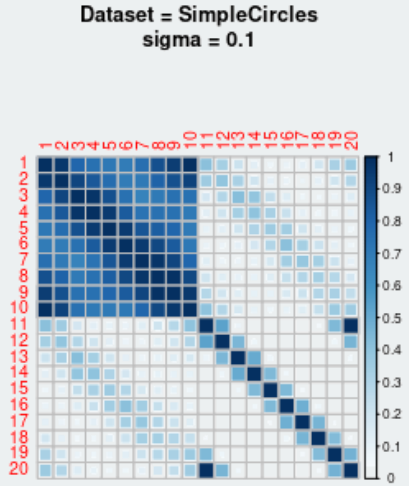
B : observations 11 à 20

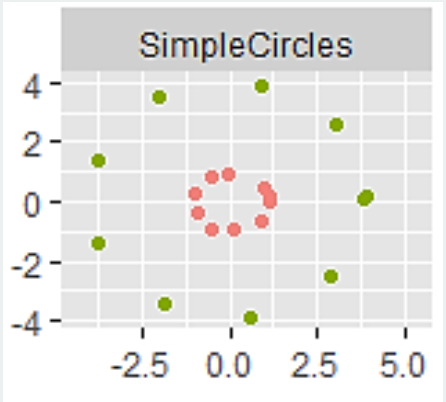
Corrélogramme des similarités calculées par le noyau Gaussien

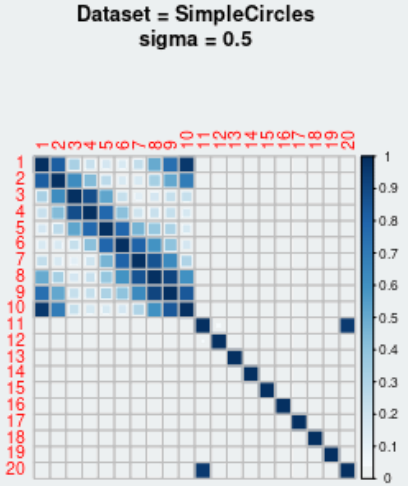
Influence du paramètre sigma

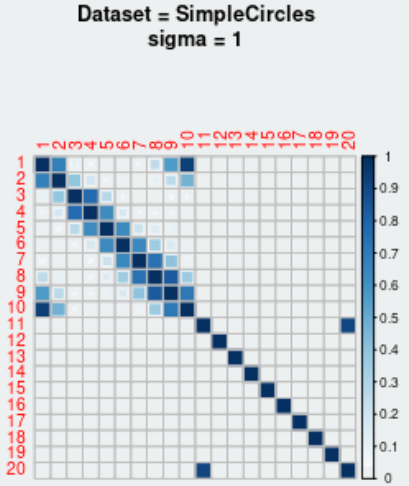


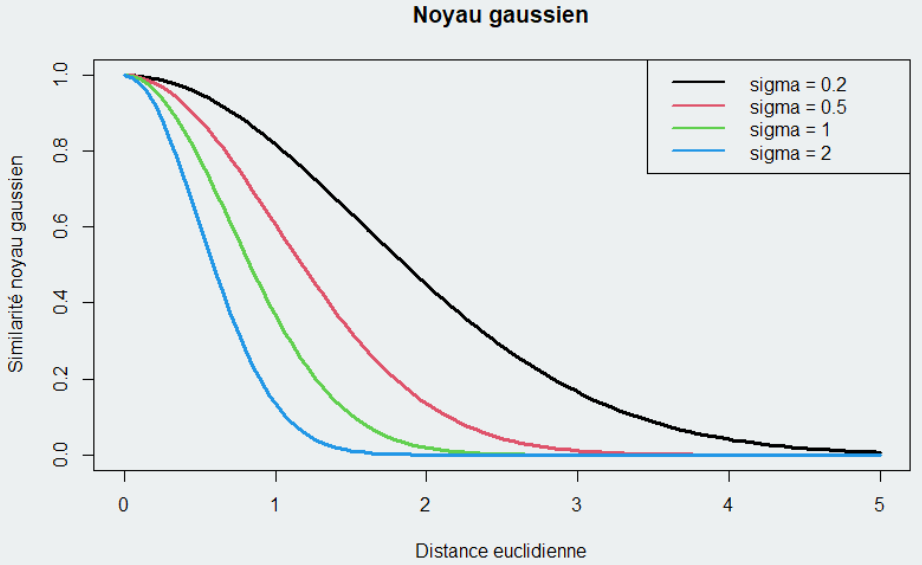










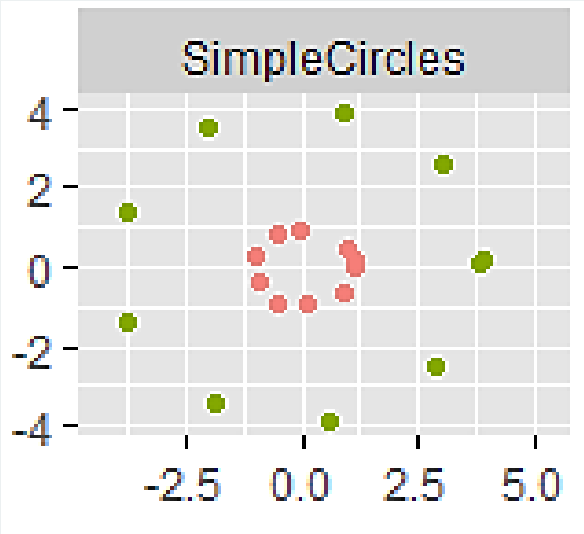


Heuristique pour déterminer la valeur de sigma

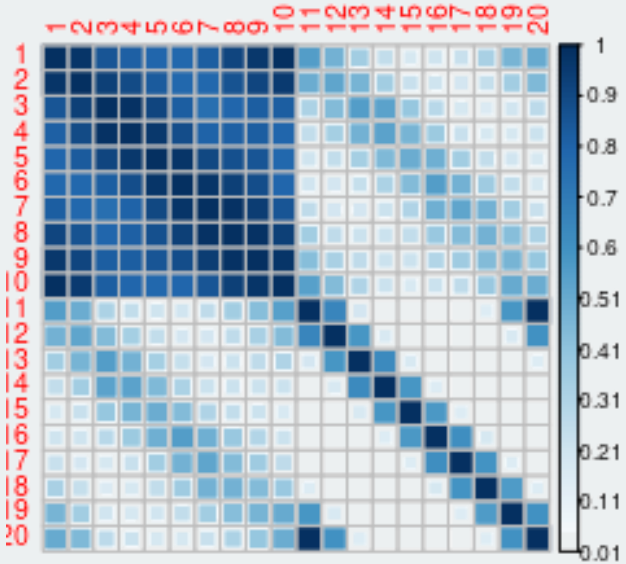
$$\sigma = \text{median}\left(\frac{1}{||x - x'||^2}\right)$$

Dataset = SimpleCircles
sigma = 0.067

Jeu de données original



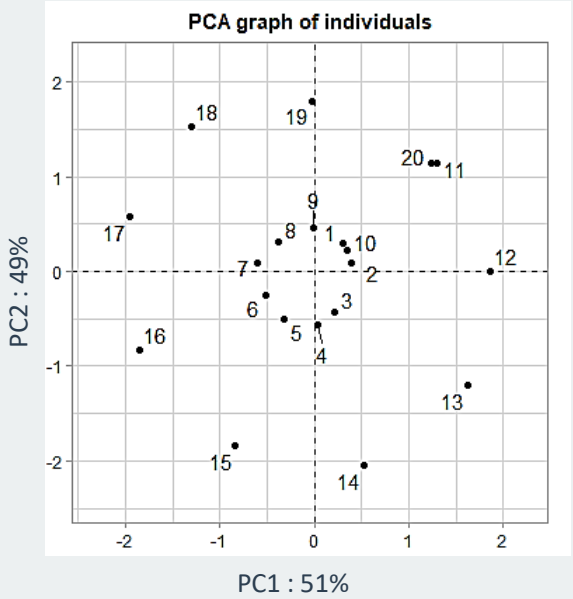
Noyau Gaussien



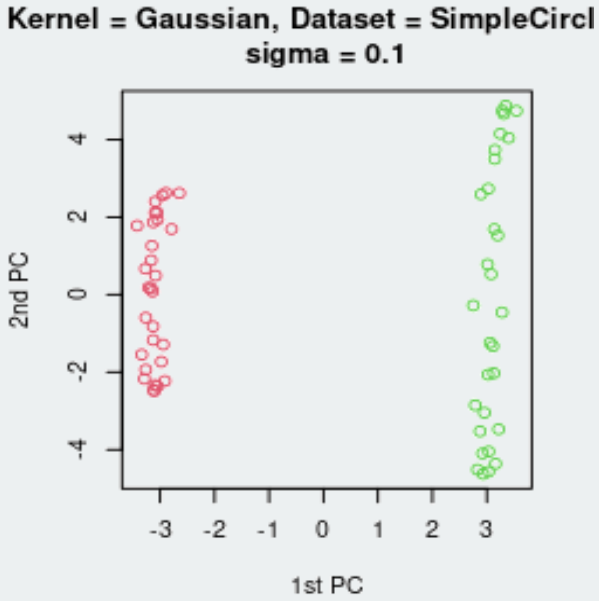
A : observations 1 à 10
B : observations 11 à 20

Analyse en composante principale à noyau (KPCA)

Analyse en composante Principale



Analyse en composante Principale à Noyau



La KPCA permet d'identifier les deux groupes de notre jeu de données,
Les observations du petit cercle d'un côté et celles du grand cercle de l'autre,
Sur le premier axe de la KPCA.

Conclusion sur les jeux de données simulés

- Ces premières approches à noyaux ont permis de se familiariser avec ces méthodes,
- Définir des outils de représentations,
- De mettre au point un protocole d'utilisation.

- Grâce à cette première manipulation des méthodes à noyau,
- Développement de méthodes pour des données de biologie,
- Noyaux de génotypages.

Noyaux de génotypage

Utilisation du ‘‘package’’ R SKAT.

ARTICLE

Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test

Michael C. Wu,[1,5] Seunggeun Lee,[2,5] Tianxi Cai,[2] Yun Li,[1,3] Michael Boehnke,[4] and Xihong Lin[2,*]

- G la matrice de génotypage,
- m le nombre de marqueurs,

G_{ij} prend pour valeurs :

- 0 si le marqueur est identique à la référence,
- 1 s’il est hétérozygote “muté”,
- et 2 s’il est homozygote “muté”,

“muté” par rapport à la référence.

Noyau dit linéaire :

$$K(G_i, G_{i'}) = \sum_{j=1}^{m} G_{ij} G_{i'j}$$

Noyau dit quadratique :

$$K(G_i, G_{i'}) = (1 + \sum_{j=1}^{m} G_{ij} G_{i'j})^2$$

Noyau dit additif :

$$K(G_i, G_{i'}) = \sum_{j=1}^{m} (2 - |G_{ij} - G_{i'j}|)$$

Jeux de données *Arabidopsis thaliana*

- 22 écotypes d'*Arabidopsis thaliana* originaires des Pyrénées,
- + Sha : écotype de haute altitude du Tadjikistan,
- Références : Col0,
- 338 marqueurs SNP ("Single Nucléotide polymorphism").

Objectif de l'étude : décrire les mécanismes d'adaptation d'A. thaliana aux conditions de températures en altitude. (Duruflé *et al.* 2019)

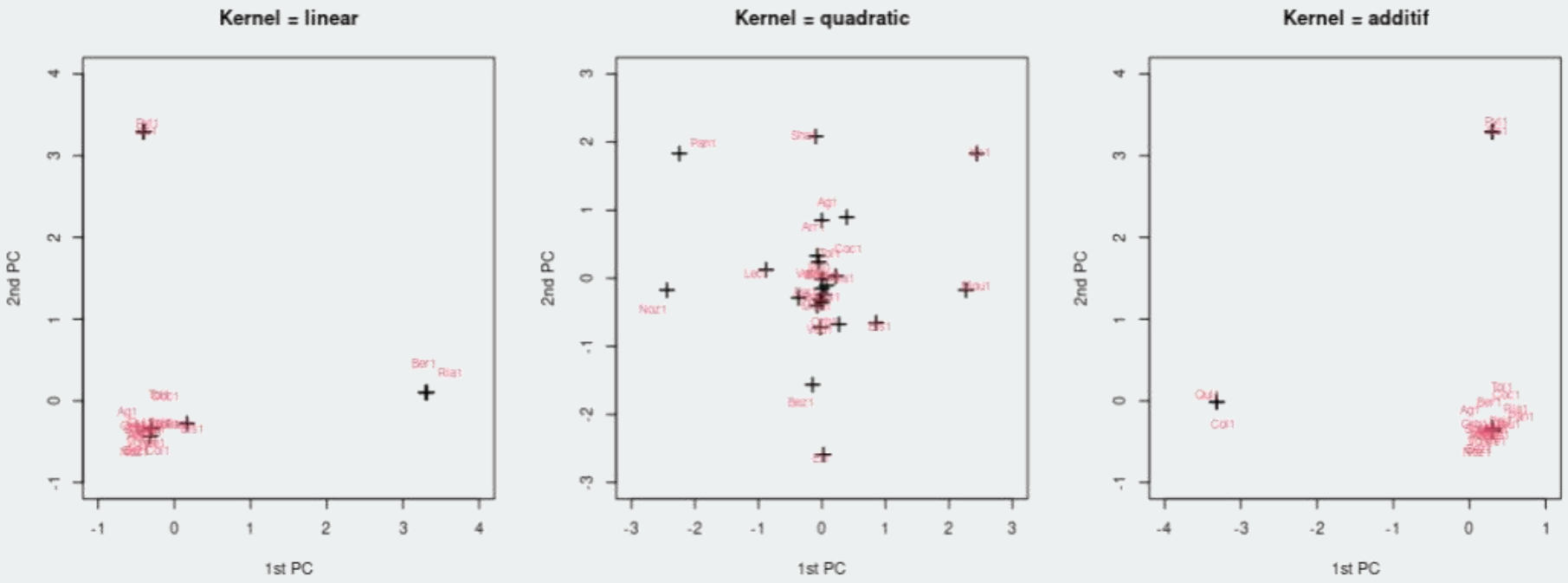
Données à disposition :

- Matrice de génotypage sur les 24 écotypes, 338 marqueurs.

Calcul des noyaux et visualisation des résultats.

KPCA : données de génotypage

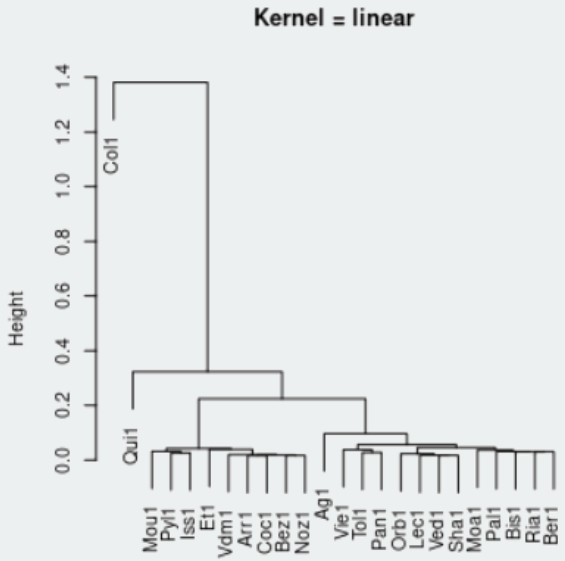
Résultats de la KPCA (2 premiers axes) avec les 3 noyaux de génotypages :

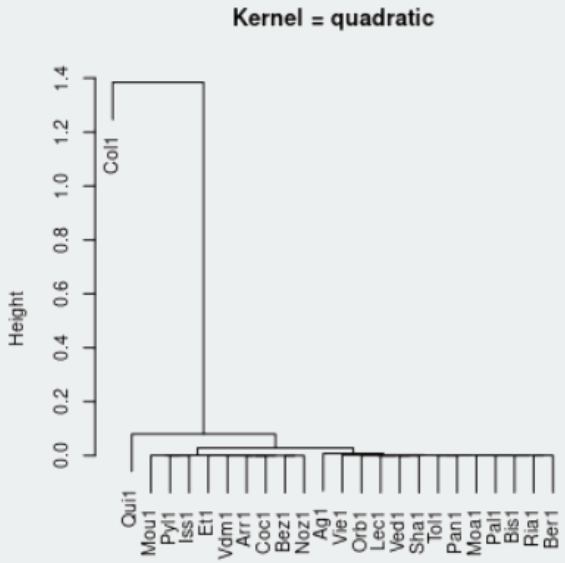


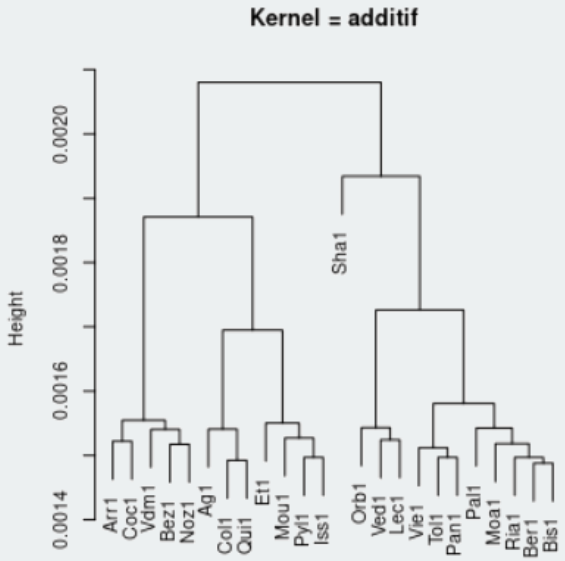
Chacun des noyaux calcule une similarité différentes entre les individus en fonction de leur variations génétiques.

Visualisation de la similarité calculée par les différents noyaux

Représentation des similarités :







Conclusion sur les noyaux de génotypage

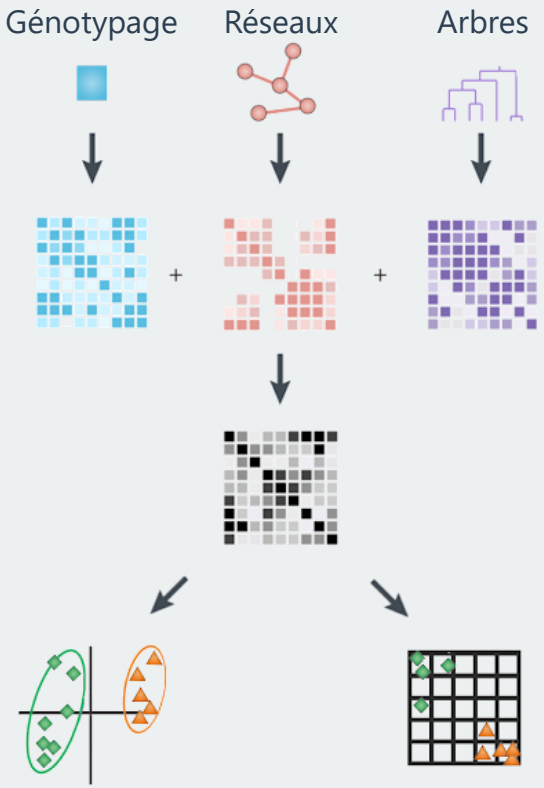
- Méthodologie de calcul de similarité entre individus à partir d'une matrice de génotypage grâce aux méthodes à noyaux,

- Quelques outils de visualisation de cette similarité, notamment par la formation de groupes d'individus,

- Méthodes à noyaux mettent en valeur des variations génétiques différentes,

- Il est désormais nécessaire de se rapprocher de l'expert de ces données pour vérifier la pertinence des résultats,
- Introduire les données de phénotypes associées.

Perspectives vers l'intégration des données.



Vers la biologie des systèmes…

Remerciements

Merci pour votre attention.

Sébastien Déjean

Jérôme Mariette

MASTER SCIENCES
ET NUMERIQUE POUR LA SANTÉ

Benchmarking