MUSÉUM NATIONAL D'HISTOIRE NATURELLE

SEMIN-R

Introduction à la classification hiérarchique

Loïc PONGER
MNHN, Régulation et Dynamique des Génomes

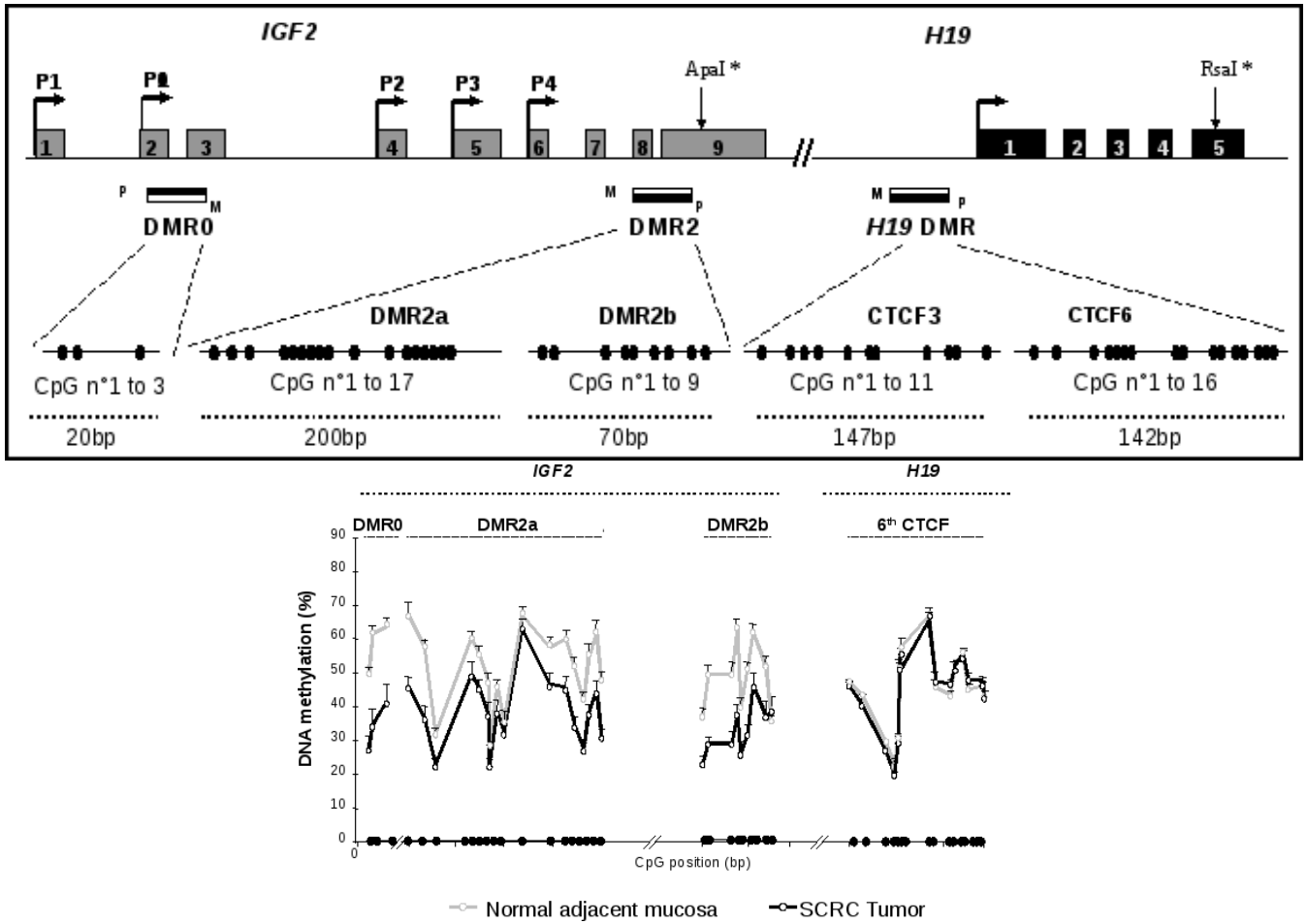
Une introduction à la classification hiérarchique

Loïc PONGER

ponger@mnhn.fr
USM 503 - Régulation et Dynamique des Génomes
Muséum National d'Histoire Naturelle
43 rue Cuvier
75005 Paris

23 juin 2009

Problème biologique : la méthylation des cytosines

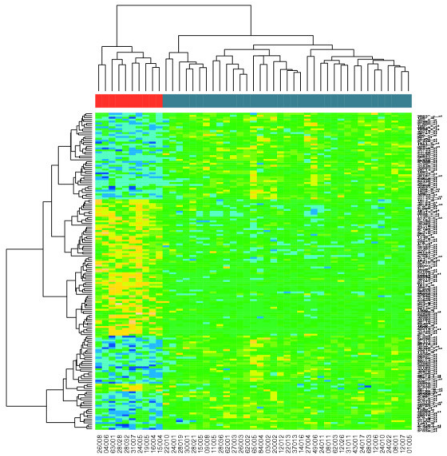


Est-ce qu'il y a des clusters de CpG dont le profil de méthylation est similaire ? Est-ce qu'ils correspondent aux 4/5 différentes régions étudiées ?

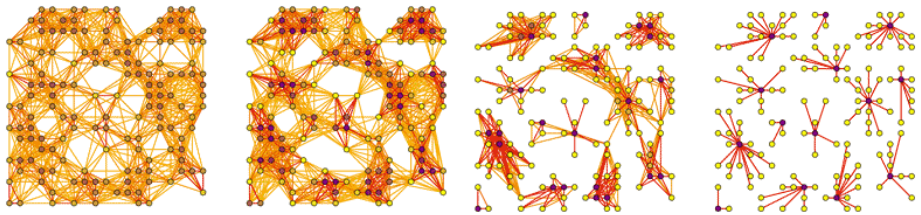
Autres problèmes pour le clustering ...

Clustering ...

- de gènes en fonction de données d'expression dans différents tissus



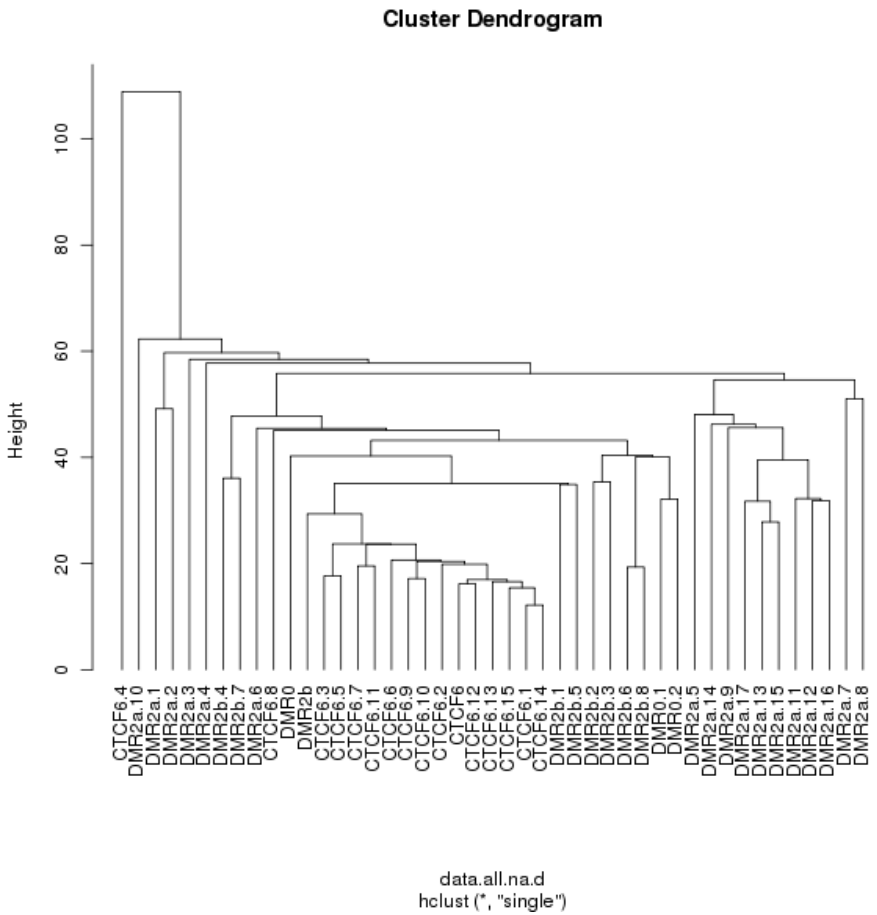
- de protéines en fonction de leur similarité (orthologues)



Introduction

Classification hiérarchique

- Objectif : regrouper les individus proches en k-classes
- k n'étant pas fixé *a priori* (Partitionnement)
- Résultat : un dendrogramme
- ascendante vs. descendante
- Fonction : hclust()



Algorithme

1. choix des distances et de la méthode de regroupement
2. calcul de distances entre toutes les paires d'individus (matrice)
3. chaque individu est considéré comme un cluster

Algorithme

1. choix des distances et de la méthode de regroupement
2. calcul de distances entre toutes les paires d'individus (matrice)
3. chaque individu est considéré comme un cluster
4. recherche des deux clusters à regrouper (cf. méthode de regroupement)
5. fusion des deux clusters et mise à jour de la matrice de distances

Algorithme

1. choix des distances et de la méthode de regroupement
2. calcul de distances entre toutes les paires d'individus (matrice)
3. chaque individu est considéré comme un cluster
4. recherche des deux clusters à regrouper (cf. méthode de regroupement)
5. fusion des deux clusters et mise à jour de la matrice de distances
6. répétition à partir de l'étape 4 jusqu'à n'avoir qu'un cluster

Choisir ...

... une distance entre individus

- mesure de dissimilarité entre N individus
- fonctions : dist, dist.binary(), dist.quant(), dist.genet(), dist.prop(), dist.dudi(), dist.neig() (package ade4)
- résultat : matrice de distances (NxN)

Choisir ...

... une méthode de regroupement des clusters

- lien simple : $min(d(i,j), i \in C_i, j \in C_j)$ (arbres applatis)

Choisir ...

... une méthode de regroupement des clusters

- lien simple : $min(d(i,j), i \in C_i, j \in C_j)$ (arbres applatis)
- lien complet : $max(d(i,j), i \in C_i, j \in C_j)$ (arbres allongés)

Choisir ...

... une méthode de regroupement des clusters

- lien simple : $min(d(i,j), i \in C_i, j \in C_j)$ (arbres applatis)
- lien complet : $max(d(i,j), i \in C_i, j \in C_j)$ (arbres allongés)
- lien moyen (UPGMA) : $\frac{\sum_{i,j} d(i,j)}{n_1 x n_2}, i \in C_i, j \in C_j$

Choisir ...

... une méthode de regroupement des clusters

- lien simple : $min(d(i,j), i \in C_i, j \in C_j)$ (arbres applatis)
- lien complet : $max(d(i,j), i \in C_i, j \in C_j)$ (arbres allongés)
- lien moyen (UPGMA) : $\frac{\sum_{i,j} d(i,j)}{n_1 x n_2}, i \in C_i, j \in C_j$
- distance des centres de gravités : $d_{eucl}(g_1, g_2)$

Choisir ...

... une méthode de regroupement des clusters

- lien simple : $min(d(i,j), i \in C_i, j \in C_j)$ (arbres applatis)
- lien complet : $max(d(i,j), i \in C_i, j \in C_j)$ (arbres allongés)
- lien moyen (UPGMA) : $\frac{\sum_{i,j} d(i,j)}{n_1 x n_2}, i \in C_i, j \in C_j$
- distance des centres de gravités : $d_{eucl}(g_1, g_2)$
- distance de Ward : $\sqrt{\frac{n_1 x n_2}{n_1 + n_2}} d_{eucl}(g_1, g_2)$

Questions

- meilleur dendrogramme ?
- où couper le dendrogramme (déterminer k) ?
- validité des clusters ?

Validité des clusters

- package pvclust et pvclust() (Suzuki, 2006)

Validité des clusters

- package pvclust et pvclust() (Suzuki, 2006)
- cohérence des clusters par rapport aux données

Validité des clusters

- package pvclust et pvclust() (Suzuki, 2006)
- cohérence des clusters par rapport aux données
- Approximately Unbiased p-value (multiscale bootstrap resampling, Shimodaira)

Validité des clusters

- package pvclust et pvclust() (Suzuki, 2006)
- cohérence des clusters par rapport aux données
- Approximately Unbiased p-value (multiscale bootstrap resampling, Shimodaira)
- Bootstrap p-value
- nombre de bootstrap : N=1000 puis N=10000

Validité des clusters

- package pvclust et pvclust() (Suzuki, 2006)
- cohérence des clusters par rapport aux données
- Approximately Unbiased p-value (multiscale bootstrap resampling, Shimodaira)
- Bootstrap p-value
- nombre de bootstrap : N=1000 puis N=10000
- pvrect() : rectangles autour des clusters significatifs

Validité des clusters

- package pvclust et pvclust() (Suzuki, 2006)
- cohérence des clusters par rapport aux données
- Approximately Unbiased p-value (multiscale bootstrap resampling, Shimodaira)
- Bootstrap p-value
- nombre de bootstrap : N=1000 puis N=10000
- pvrect() : rectangles autour des clusters significatifs
- seplot() : affiche le sd des AU p-values

Pour finir

- dendrogram : as.dendrogram(), rev(), cut()

Pour finir

- dendrogram : as.dendrogram(), rev(), cut()
- package cluster : agnes (similar hclust), diana (CHD), ...