



HAL
open science

Analyse des reads orphelins

Sophie S. Schbath

► **To cite this version:**

Sophie S. Schbath. Analyse des reads orphelins. Ecole-chercheur Métaprogramme Métaomique des Ecosystèmes Microbiens, Feb 2011, Paris, France. pp.20. hal-02802862

HAL Id: hal-02802862

<https://hal.inrae.fr/hal-02802862v1>

Submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Analyse des reads orphelins

Sophie Schbath

INRA - MIG



Métagénomique des Eco-systèmes Microbiens, Paris, février 2011

Quels types d'analyses ?

Objectif

- Trouver une structure dans ce grand ensemble de séquences.

Idée 1 : étudier la composition *globale* en “mots” de ces reads

- Mots fréquents (signature) ?
- Mots *significativement* fréquents ? Cad on s'affranchit par ex. de la composition en nucléotides (voire en di-, tri-nucl.).
→ Problème de l'hétérogénéité des compositions : regrouper les reads de compositions homogènes/proches.

Idée 2 : faire des groupes de reads qui se “ressemblent”

- Classification *non supervisée* : aucun a priori sur les groupes et sur leur nombre.
- Groupes les plus homogènes possibles, groupes bien séparés.

Deux familles de méthodes de classification

Méthodes de partitionnement à base de distances. [Cadre géométrique]

- Définir une distance entre 2 séquences.
- Regrouper les séquences “proches”.

Méthodes à base de modèles. [Cadre probabiliste].

On suppose que

- chaque séquence est gouvernée par un modèle,
- les séquences d’un même groupe sont gouvernées par le même modèle
→ un modèle différent par groupe.

Principe des méthodes à base de distances

On dispose de

- n séquences S_1, \dots, S_n ,
- une distance entre 2 séquences S_i et S_j : $d(S_i, S_j)$.
(on peut définir le barycentre d'un ensemble de séquences).

Décomposition de l'inertie du nuage pour K groupes :

$$I = \sum_{i=1}^n d^2(S_i, G) = \underbrace{\sum_{k=1}^K \sum_{i \in \text{groupe } k} d^2(S_i, C_k)}_{\text{Inertie intra-groupe}} + \underbrace{\sum_{k=1}^K n_k d^2(C_k, G)}_{\text{Inertie inter-groupes}}$$

avec G barycentre de S_1, \dots, S_n , C_k barycentre du groupe k et n_k taille du groupe k

Objectif = trouver la partition à K groupes qui

- minimise l'inertie intra-groupe (= maximise l'inertie inter-groupe)

Distances entre séquences

Il existe beaucoup de distances entre séquences ; Dai et al. (2008) en recense 9.

Elles sont le plus souvent construites à partir des comptages des q -mers de chaque séquence :

$$\mathbf{N}_i = [N_i(\text{aa}\cdots\text{aa}) \quad N_i(\text{aa}\cdots\text{ac}) \quad \dots \quad N_i(\text{tt}\cdots\text{tt})] \in \mathbb{N}^{4^q}.$$

- Distance euclidienne entre \mathbf{N}_i/ℓ_i et \mathbf{N}_j/ℓ_j .
- Coefficient de corrélation : Fichant & Gautier (1987)
- Distance de Kullback-Leibler : Wu et al. (2001).
- Score D_2 : $D_2 = \sum_{w \in q\text{-mers}} N_i(w)N_j(w)$, Lippert et al.(2002)
- etc.

Ce large éventail pose le problème du **choix de la distance**.

Partition optimale

Comment trouver la partition de n individus en K groupes qui minimise l'inertie intra-groupe ?

Impossible de considérer toutes les partitions :

- $n = 19$ et $K = 4 \rightarrow$ plus de 10^{10} possibilités
- $n = 100$ et $K = 2 \rightarrow$ plus de 10^{29} possibilités

Solution : ne visiter qu'un nombre restreint de partitions (solution approchée)

- Algorithme des K -means
- Classification Ascendante Hiérarchique

Algorithme des K -means

Nombre de groupes K fixé.

Initialisation : on choisit K individus comme centres des K groupes

Algorithme

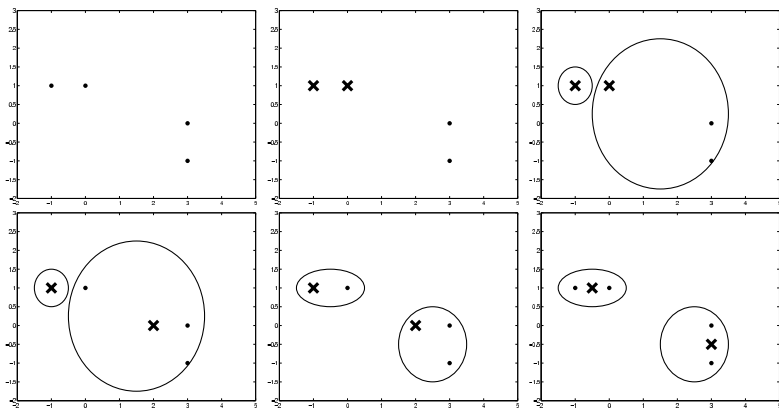
- ➊ Affectation : chaque individu est affecté au centre le plus proche
→ on définit ainsi K groupes G_1, \dots, G_k
- ➋ Calcul des nouveaux centres (barycentre) de chaque groupe G_j .

Condition d'arrêt

- Nb fixé d'itérations.
- Les groupes restent les mêmes entre 2 itérations (convergence).

Propriété : l'inertie intra-groupe diminue à chaque itération.

K-means : exemple



K-means : avantages/inconvénients

Avantages :

- simplicité,
- convergence rapide de l'algorithme.

Inconvénients

- instabilité par rapport à l'initialisation,
- minimum local,
- choix du nombre de groupes.

Classification Ascendante Hiérarchique

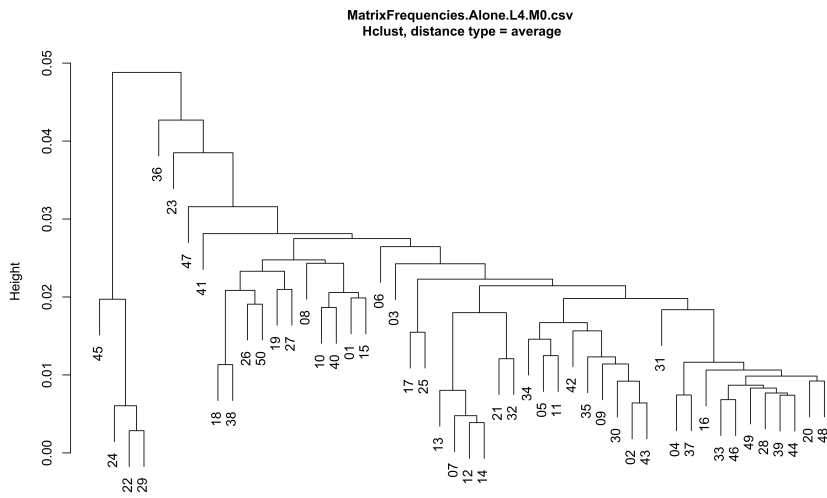
Objectif : construire une suite de partitions emboîtées en n groupes, puis $(n - 1)$ groupes, ..., en 1 groupe.

Initialisation : les individus constituent des groupes à eux seuls (n groupes)

Algorithme itératif :

- les 2 individus les + “proches” sont réunis en 1 groupe,
- on calcule les distances entre ce nouveau groupe et les $(n - 2)$ autres individus
→ nécessité de définir une distance entre 2 groupes G_i et G_j
- les 2 “individus” les + “proches” sont réunis en 1 groupe,
- etc. jusqu’à n’obtenir qu’un seul groupe.

CAH : dendrogramme

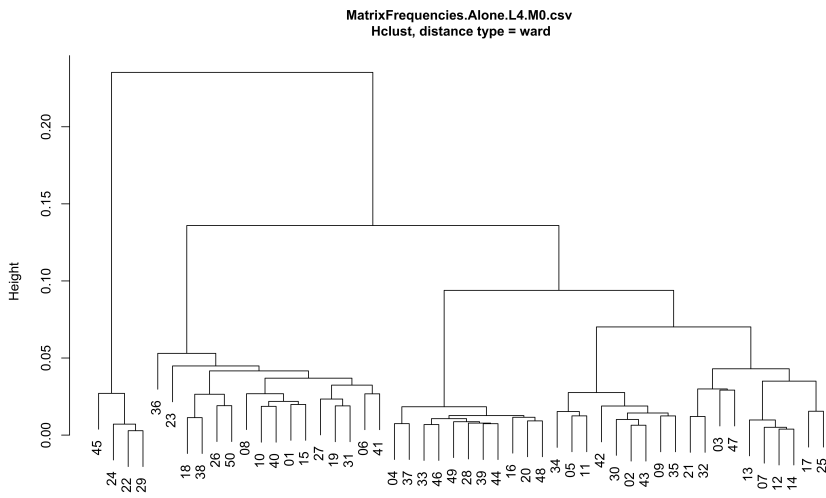


CAH : distance entre groupes

Il existe aussi plusieurs distances entre 2 groupes A et B :

- **lien simple** : $\min\{d(a, b) \text{ avec } a \in A, b \in B\}$
- **lien complet** : $\max\{d(a, b) \text{ avec } a \in A, b \in B\}$
- **lien moyen** : $\frac{1}{|A| \cdot |B|} \sum_{a,b} d(a, b)$
- **Ward** : $d^2(C_A, C_B) \times \frac{|A| \cdot |B|}{|A| + |B|}$
(on regroupe les deux groupes qui entraînent le plus faible gain d'inertie intra-groupe)

CAH : dendrogramme



CAH : avantages/inconvénients

Avantages

- stabilité (pas d'initialisation),
- pas besoin de choisir le nombre de groupes à l'avance. Il peut être choisi *a posteriori* à partir du dendrogramme.

Inconvénients

- lent dès que n est grand,
(possibilité de réduire le nombre d'individus en faisant un K -means avant.)
- dépend du choix de la distance entre groupe choisie.

Principe du modèle de mélange

On dispose de n séquences s_1, \dots, s_n .

Mélange de distributions

- Les séquences sont issues de $K < n$ modèles (paramètres θ_k) :

$$(S_i \mid S_i \in \text{groupe } k) \sim f(\cdot; \theta_k), \quad k = 1, \dots, K$$

- Distribution de S_j :

$$f(s_j) = \sum_{k=1}^K \underbrace{\alpha_k}_{P(S \in \text{groupe } k)} f(s_j; \theta_k)$$

Objectifs :

- estimer les paramètres (α_k, θ_k)
- calculer les probabilités *a posteriori* d'appartenance aux groupes :

$$P(S_j \in \text{groupe } k \mid S_j = s_j) = \frac{\alpha_k f(s_j; \theta_k)}{\sum_{\ell} \alpha_{\ell} f(s_j; \theta_{\ell})}$$

Exemples de modélisation

Modèles sur la composition \mathbf{N}_i de S_i en q -mers

- $\mathbf{N}_i \ell_i$ modélisé par une gaussienne multi-dimensionnelle $\theta_k = (\mathbf{m}_k, \Sigma_k)$.
- \mathbf{N}_i modélisé par un vecteur de variables de Poisson (pb de dépendance).

Modèle sur la séquence S_i

- Analyser \mathbf{N}_i ou la chaîne de Markov stationnaire d'ordre $(q - 1)$ qui s'ajuste sur S_i est équivalent car \mathbf{N}_i est la *statistique exhaustive* d'une CM($q - 1$).
- S_i est modélisée par une chaîne de Markov stationnaire d'ordre $(q - 1)$ de probabilités de transition $\Pi_k (= \theta_k)$.

Chaîne de Markov d'ordre $q - 1$

On considère une **séquence aléatoire** $S = X_1 X_2 X_3 \cdots X_\ell$,
 $X_i \in \mathcal{A} = \{a, c, g, t\}$

S est une chaîne de Markov d'ordre $q - 1$ ssi

- les lettres X_i ne sont pas indépendantes,
- la loi de la lettre X_i dépend des $(q - 1)$ lettres précédentes,
 $i = 1, \dots, \ell$,
- la loi de X_i sachant les $(q - 1)$ lettres précédentes $X_{i-q+1} \cdots X_{i-1}$ est donnée par les probabilités de transition

$$\pi(a_1 \cdots a_{q-1}, a_q) = P(X_i = a_q \mid X_{i-q+1} \cdots X_{i-1} = a_1 \cdots a_{q-1}).$$

Pour ajuster une $CM(q - 1)$ sur une séquence d'ADN :

$$\hat{\pi}(a_1 \cdots a_{q-1}, a_q) = \frac{N(a_1 \cdots a_{q-1} a_q)}{N(a_1 \cdots a_{q-1} +)}$$

Estimation par maximum de vraisemblance

On définit $Z_i = k$ si $S_i \in$ groupe k (label).

Si les Z_i étaient observés :

- calculer la vraisemblance $\mathcal{L}(\mathbf{S}, \mathbf{Z}) = P(S_i = s_i, Z_i = z_i, \forall i)$

$$\mathcal{L}(\mathbf{S}, \mathbf{Z}) = \exp \left(\sum_{i=1}^n \sum_{k=1}^K 1_{\{Z_i = k\}} \log(\alpha_k f(s_i; \theta_k)) \right)$$

- maximiser cette vraisemblance en $(\alpha_k, \theta_k) \Rightarrow (\hat{\alpha}_k, \hat{\theta}_k)$.

Mais les Z_i sont “cachés” :

- on remplace $1_{\{Z_i = k\}}$ par $P(Z_i = k \mid S_i = s_i)$,
- on maximise la pseudo-vraisemblance avec l’algorithme EM,

$$\Rightarrow (\hat{\alpha}_k, \hat{\theta}_k) \text{ et } \hat{P}(S_i \in \text{groupe } k \mid S_i = s_i) = \frac{\hat{\alpha}_k f(s_i; \hat{\theta}_k)}{\sum_{\ell} \hat{\alpha}_{\ell} f(s_i; \hat{\theta}_{\ell})}.$$

Mélanges : avantages/inconvénients

Avantages

- algorithme EM est simple et converge vite,
- probabilités *a posteriori* d'appartenance à chacune des K classes,
→ classement via la règle du Maximum A Posteriori, (MAP)
- critères théoriques de sélection du nombre K de groupes (ex. ICL)
→ pénalisation de la vraisemblance par la taille du modèle.

Inconvénients

- instabilité par rapport à l'initialisation,
- minimum local.

Classification

Pourquoi classer ?

- un objectif en soi : signification des groupes ?
- ou une première étape de l'analyse
exemple : y a-t-il des motifs sur-représentés dans les groupes de reads sachant leur composition homogène en q -mers ?

Verrous généraux

- choix des distances,
- choix du nombre K de groupes (question de recherche).

Verrous liés aux données métagénomiques : séquences courtes (q doit rester petit) en très grand nombre

- complexité calculatoire,
- fléau de la dimension.