

Clustering problems in gene data analysis

Description

Le but est de réaliser une étude de gènes cherchant à isoler des groupes évoluant de manière similaire

→ C'est donc un problème de clustering (classification) de données !

Pour permettre une étude efficace et agréable, une application a été réalisée permettant :

- la recherche et la découpe en clusters
- sauvegarde des données et des résultats
- l'affichage graphique des données et des résultats

Intérêt

Ce travail peut être utilisé pour :

- aider à la compréhension du fonctionnement des gènes
- aider dans le diagnostic des états d'une maladie
- indiquer les effets des traitements médicaux

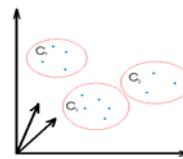
Les contraintes

Le défi était la grande quantité de données à analyser, ce qui fait que les algorithmes de clustering sont extrêmement difficiles à calculer avec une méthode exhaustive.

De plus, il n'y a aucune information sur le nombre de clusters à obtenir ou leur structure.

Les clusters

Un cluster est un ensemble d'objets ayant les caractéristiques les plus semblables possibles.



Exemple :
regroupement en 3 clusters
des objets présentés.

Nous utilisons la distance euclidienne entre les objets dans notre algorithme pour définir la similarité entre les objets.

L'évolution au sein d'un cluster

L'objectif est de diviser les gènes en clusters de manière à ce que le maximum de valeurs de chaque groupe évolue simultanément



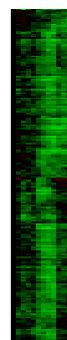
Cas désiré



Cas à éviter

Les Résultats

Données : 200 gènes pour 13 expériences



Matrice de données



Matrice réorganisée en clusters

Nombre de cluster : 3
Homogénéité globale : 0.772
Séparation globale : 0.182