

Application du clustering descriptif

Réalisation d'un premier prototype

mai 2022

Mathieu GUILBERT

Université d'Orléans - LIFO

Sommaire

1 Observations sur les données

- Rappel des données
- Premières expériences

2 Interface visuelle

3 Discussions

Rappel des données

Les fichiers de données contiennent les informations suivantes pour 645 molécules :

- nom des composés
- chemotype
- code smiles
- réactivité sur les kinases. Les valeurs sont des pourcentages d'inhibitions. Une molécule est jugée active quand la valeur est supérieure à 50.
- Poids moléculaire
- Propriétés structurelles (fingerprints)

Utilisation du fichier avec fingerprint ECFP4

Sommaire

1 Observations sur les données

- Rappel des données
- Premières expériences

2 Interface visuelle

3 Discussions

Vues sur les molécules

Deux vues sur les molécules

- description des molécules avec fingerprints
- activité biologique sur les kinases

2 manières d'utiliser le clustering descriptif:

Clustering avec fingerprints

- clustering sur les fingerprints (matrice de distances) + description sur les activités biologiques (binaires 1/0 si activité supérieure à 50 ou non)

Clustering avec fingerprints

- clustering sur les fingerprints (matrice de distances) + description sur les activités biologiques (binaires 1/0 si activité supérieure à 50 ou non)

avec MMNA et MMCTA, un seul élément dans le front de Pareto :

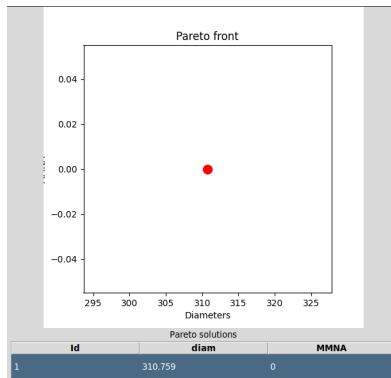


Figure: Front de Pareto pour clustering avec fingerprints

Fingerprint

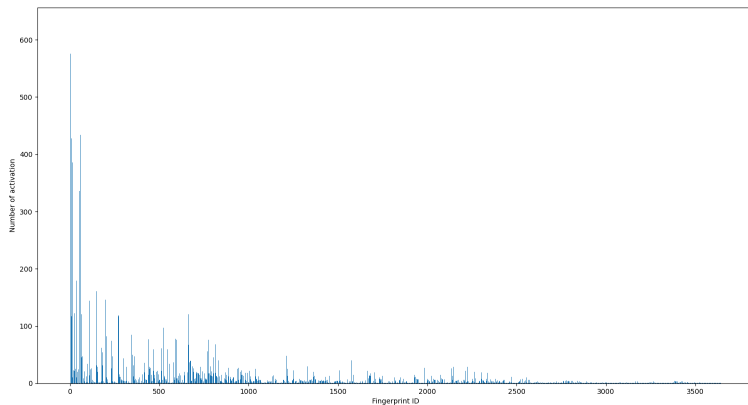


Figure: Nombre d'occurrences du fingerprint (ECFP4)

Clustering avec activité biologique

- clustering sur l'activité biologique (distance euclidienne) + description sur les fingerprints
 - maximisation MMCTA: problème de mémoire dans la version actuelle
 - maximisation MMNA: résultats complet obtenus avec $k=15$

Clustering avec activité biologique

- clustering sur l'activité biologique (distance euclidienne) + description sur les fingerprints

- maximisation MMCTA: problème de mémoire dans la version actuelle
- maximisation MMNA: résultats complet obtenus avec $k=15$

Objectif MMNA:

maximiser le nombre minimal de descripteurs en commun par paires de molécules de même cluster

+

minimisation du diamètre maximal de la partition

Sommaire

1 Observations sur les données

2 Interface visuelle

- **Front de Pareto**
- Détails de la partition
- Informations sur le cluster
- Descriptions des tags
- Changement seuil de description
- Détails de la partition
- Ajouts futurs

3 Discussions

Front de Pareto (k=15, MMNA)

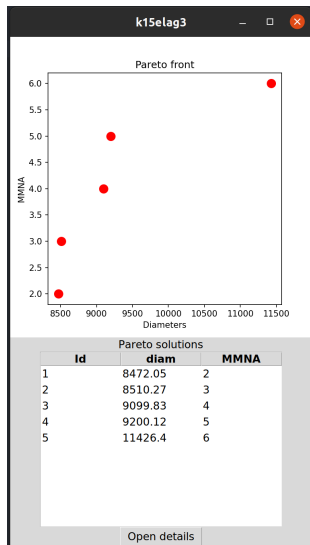


Figure: Première fenêtre

Sommaire

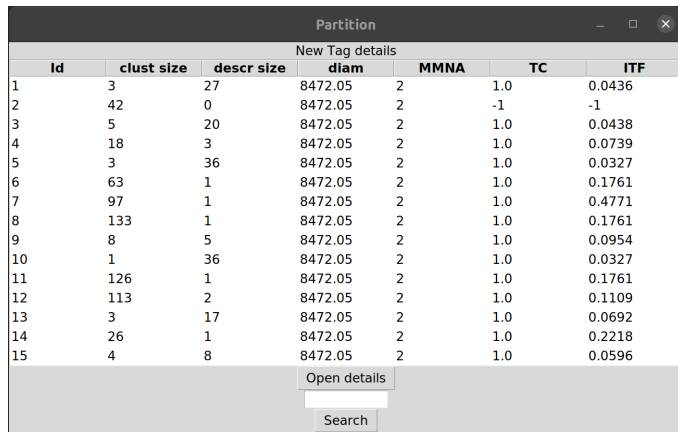
1 Observations sur les données

2 Interface visuelle

- Front de Pareto
- **Détails de la partition**
- Informations sur le cluster
- Descriptions des tags
- Changement seuil de description
- Détails de la partition
- Ajouts futurs

3 Discussions

Détails de la partition



Partition							
New Tag details							
Id	clust size	descr size	diam	MMNA	TC	ITF	
1	3	27	8472.05	2	1.0	0.0436	
2	42	0	8472.05	2	-1	-1	
3	5	20	8472.05	2	1.0	0.0438	
4	18	3	8472.05	2	1.0	0.0739	
5	3	36	8472.05	2	1.0	0.0327	
6	63	1	8472.05	2	1.0	0.1761	
7	97	1	8472.05	2	1.0	0.4771	
8	133	1	8472.05	2	1.0	0.1761	
9	8	5	8472.05	2	1.0	0.0954	
10	1	36	8472.05	2	1.0	0.0327	
11	126	1	8472.05	2	1.0	0.1761	
12	113	2	8472.05	2	1.0	0.1109	
13	3	17	8472.05	2	1.0	0.0692	
14	26	1	8472.05	2	1.0	0.2218	
15	4	8	8472.05	2	1.0	0.0596	

Open details

Search

Figure: Deuxième fenêtre

Mesures de qualité

$$TC(C_i) = \frac{1}{|D_i|} \sum_{d \in D_i} \frac{|\{(x, t) \in C_i : d \in t\}|}{|C_i|}$$

$$ITF(C_i) = \frac{1}{|D_i|} \sum_{d \in D_i} \log \frac{K}{\sum_{j=1}^K |d \in D_j|}$$

- TC mesure la pertinence d'un tag décrivant un cluster spécifique.
- ITF mesure la répétitivité d'un tag parmi une partition.

Sommaire

1 Observations sur les données

2 Interface visuelle

- Front de Pareto
- Détails de la partition
- **Informations sur le cluster**
- Descriptions des tags
- Changement seuil de description
- Détails de la partition
- Ajouts futurs

3 Discussions

Détails sur le cluster 9 avec seuil de 100%

The screenshot shows a window titled "Cluster details" with a dark header. Below the header is a "Cluster summary" table with columns: Id, clust size, descr size, diam, MMNA, TC, ITF, and tags. The values for cluster 9 are: Id: 9, clust size: 8, descr size: 5, diam: 8472.05, MMNA: 2, TC: 1.0, ITF: 0.0954, tags: 2, 3, 12, 49, 51. Below the summary are tabs for "Search", "See cluster tags", "Growth rate tags", and "TF-IDF tags". The main area is titled "Cluster details" and contains a table with columns: name, smile, descr size, and number of clust tags. The table lists 8 molecules with their names, SMILES, description sizes, and the number of cluster tags they belong to.

Id	clust size	descr size	diam	MMNA	TC	ITF	tags
9	8	5	8472.05	2	1.0	0.0954	2, 3, 12, 49, 51

name	smile	descr size	number of clust tags
PFE-PKIS 36	3-vinyl indazole	42	5
UNC10112774A	2-aminobenzimidazole	49	5
UNC10225038A	3-anilino-4-aryl maleimide	61	5
UNC10225081A	4-pyrimidinyl_ortho-aryl_azoles	62	5
UNC10225235A	2-anilino-4-pyrimidinopyrimidine	56	5
UNC10225280A	carboxamide pyrimidine	62	5
UNC10225437A	4-pyrimidinyl_ortho-aryl_azoles	55	5
UNC10225486A	4-pyridinyl_ortho-aryl_azoles	58	5

Figure: Liste des molécules d'un cluster

Mesures de qualités

$$GR_i(X) = \frac{|D|-|C_i|}{|C_i|} \times \frac{F(X, C_i)}{F(X, D) - F(X, C_i)}$$

C_i est le i -ème cluster. La fréquence d'un pattern X dans un dataset D notée $F(X, D)$ est le nombre de clusters de D contenant X .

Dans notre application, pour chaque instance x , $T(x)$ dénote l'ensemble de tags décrivant x .

Définition inspirée de: Condensed representation of emerging patterns (Soulet, Arnaud and Crémilleux, Bruno and Rioult, François)

Mesures de qualité

$$TF - IDF(d, C) = \frac{|\{x \in C \mid d \in T(x)\}|}{|C|} * \log\left(\frac{|X|}{|\{x \in X \mid d \in T(x)\}|}\right)$$

avec d un tag donnée, C un cluster, X un dataset, t l'ensemble des tags associés à l'instance x .

Sommaire

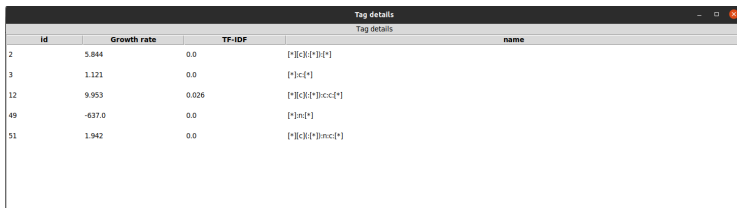
1 Observations sur les données

2 Interface visuelle

- Front de Pareto
- Détails de la partition
- Informations sur le cluster
- **Descriptions des tags**
- Changement seuil de description
- Détails de la partition
- Ajouts futurs

3 Discussions

Descriptions des tags



Tag details			
id	Growth rate	TF-IDF	name
2	5.844	0.0	*[[c];*]*[*]
3	1.121	0.0	*]:c[*]
12	9.953	0.026	*[[c];*]:cc[*]
49	-637.0	0.0	*]:n[*]
51	1.942	0.0	*[[c];*]:nc[*]

Figure: Descripteurs d'un cluster

Sommaire

1 Observations sur les données

2 Interface visuelle

- Front de Pareto
- Détails de la partition
- Informations sur le cluster
- Descriptions des tags
- **Changement seuil de description**
- Détails de la partition
- Ajouts futurs

3 Discussions

Sommaire

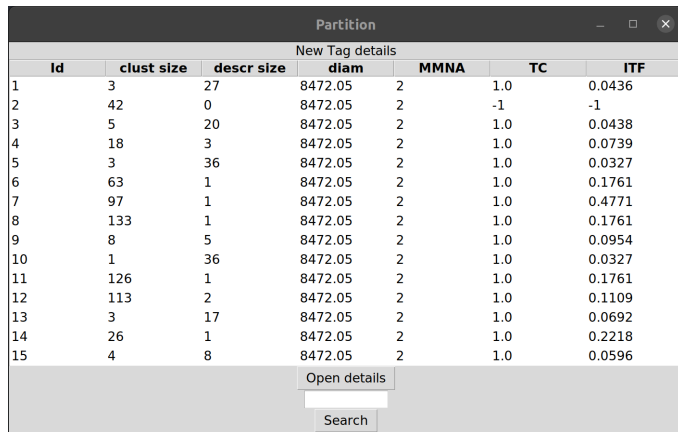
1 Observations sur les données

2 Interface visuelle

- Front de Pareto
- Détails de la partition
- Informations sur le cluster
- Descriptions des tags
- Changement seuil de description
- **Détails de la partition**
- Ajouts futurs

3 Discussions

Détails de la partition



The screenshot shows a window titled "Partition" with a "New Tag details" table. The table has 8 columns: Id, clust size, descr size, diam, MMNA, TC, and ITF. Below the table are buttons for "Open details" and "Search".

Id	clust size	descr size	diam	MMNA	TC	ITF
1	3	27	8472.05	2	1.0	0.0436
2	42	0	8472.05	2	-1	-1
3	5	20	8472.05	2	1.0	0.0438
4	18	3	8472.05	2	1.0	0.0739
5	3	36	8472.05	2	1.0	0.0327
6	63	1	8472.05	2	1.0	0.1761
7	97	1	8472.05	2	1.0	0.4771
8	133	1	8472.05	2	1.0	0.1761
9	8	5	8472.05	2	1.0	0.0954
10	1	36	8472.05	2	1.0	0.0327
11	126	1	8472.05	2	1.0	0.1761
12	113	2	8472.05	2	1.0	0.1109
13	3	17	8472.05	2	1.0	0.0692
14	26	1	8472.05	2	1.0	0.2218
15	4	8	8472.05	2	1.0	0.0596

Figure: Deuxième fenêtre

Description des clusters avec un seuil de fréquence

Partition									
New Tag details									
id	clust size	descr size	diam	MMNA	TC	ITF			
1	3	39	8472.05	2	0.8974	0.0122			
2	42	8	8472.05	2	0.7619	0.0497			
3	5	67	8472.05	2	0.794	0.0176			
4	18	33	8472.05	2	0.8165	0.0356			
5	3	46	8472.05	2	0.9275	0.0256			
6	63	10	8472.05	2	0.7968	0.0331			
7	97	10	8472.05	2	0.7608	0.0135			
8	133	8	8472.05	2	0.7763	0.0168			
9	8	16	8472.05	2	0.8047	0.0249			
10	1	36	8472.05	2	1.0	0.0327			
11	126	6	8472.05	2	0.8333	0.0104			
12	113	8	8472.05	2	0.8252	0.0121			
13	3	25	8472.05	2	0.8933	0.035			
14	26	17	8472.05	2	0.8032	0.0411			
15	4	21	8472.05	2	0.8452	0.0273			

Open details

Search

Figure: Même partition - changement sous-jacent de description des clusters - les tags décrivent au moins 60% des éléments du clusters

Sommaire

1 Observations sur les données

2 Interface visuelle

- Front de Pareto
- Détails de la partition
- Informations sur le cluster
- Descriptions des tags
- Changement seuil de description
- Détails de la partition
- Ajouts futurs

3 Discussions

Ajouts futurs

- Utilisation du critère du diamètre et de MMNA pour des raisons d'efficacité
 - ▶ Intégration d'autres critères
 - ★ Lesquels seraient intéressants?
 - ★ Efficacité
- Amélioration de l'interface
 - ▶ Afficher Diamètre et MMNA pour chaque cluster (facile)
 - ▶ Commentaires à propos des partitions (ou des clusters) → ajout de contraintes ?
 - ▶ Ajout de la visualisation des molécules ? Des pharmacophores ?
 - ▶ Amélioration du visuel
 - Desiderata des chemo-informaticiens
- Utilisation d'autres types de fingerprints?

Sujets

- Idées de rajouts dans visualisations ?

Sujets

- Idées de rajouts dans visualisations ?
- Ajouter l'absence de propriétés comme descripteurs ?

Sujets

- Idées de rajouts dans visualisations ?
- Ajouter l'absence de propriétés comme descripteurs ?
- Changer nombre de clusters que l'on cherche ?

Sujets

- Idées de rajouts dans visualisations ?
- Ajouter l'absence de propriétés comme descripteurs ?
- Changer nombre de clusters que l'on cherche ?
- Utilisation d'autres critères de qualité ?

Sujets

- Idées de rajouts dans visualisations ?
- Ajouter l'absence de propriétés comme descripteurs ?
- Changer nombre de clusters que l'on cherche ?
- Utilisation d'autres critères de qualité ?
- Lorsqu'on utilise la distance euclidienne sur les kinases, faudrait-il pondérer pour qu'une famille ne l'emporte pas sur les autres ?

Sujets

- Idées de rajouts dans visualisations ?
- Ajouter l'absence de propriétés comme descripteurs ?
- Changer nombre de clusters que l'on cherche ?
- Utilisation d'autres critères de qualité ?
- Lorsqu'on utilise la distance euclidienne sur les kinases, faudrait-il pondérer pour qu'une famille ne l'emporte pas sur les autres ?
- Valider les petits clusters qui pourraient ensuite être transformés en contraintes Must-Link ?

Sujets

- Idées de rajouts dans visualisations ?
- Ajouter l'absence de propriétés comme descripteurs ?
- Changer nombre de clusters que l'on cherche ?
- Utilisation d'autres critères de qualité ?
- Lorsqu'on utilise la distance euclidienne sur les kinases, faudrait-il pondérer pour qu'une famille ne l'emporte pas sur les autres ?
- Valider les petits clusters qui pourraient ensuite être transformés en contraintes Must-Link ?
- Générer des contraintes à partir de partitions précédemment obtenues et/ou à partir de similitude connues entre molécules ? → Discussions nécessaires

Sujets

- Idées de rajouts dans visualisations ?
- Ajouter l'absence de propriétés comme descripteurs ?
- Changer nombre de clusters que l'on cherche ?
- Utilisation d'autres critères de qualité ?
- Lorsqu'on utilise la distance euclidienne sur les kinases, faudrait-il pondérer pour qu'une famille ne l'emporte pas sur les autres ?
- Valider les petits clusters qui pourraient ensuite être transformés en contraintes Must-Link ?
- Générer des contraintes à partir de partitions précédemment obtenues et/ou à partir de similitude connues entre molécules ? → Discussions nécessaires
- Utiliser les propriétés hiérarchiques des fingerprints pour l'exploration et les explications ? → Priorité