Application du clustering descriptif Réalisation d'un premier prototype mai 2022

Mathieu GUILBERT

Université d'Orléans - LIFO

Sommaire

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

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 :

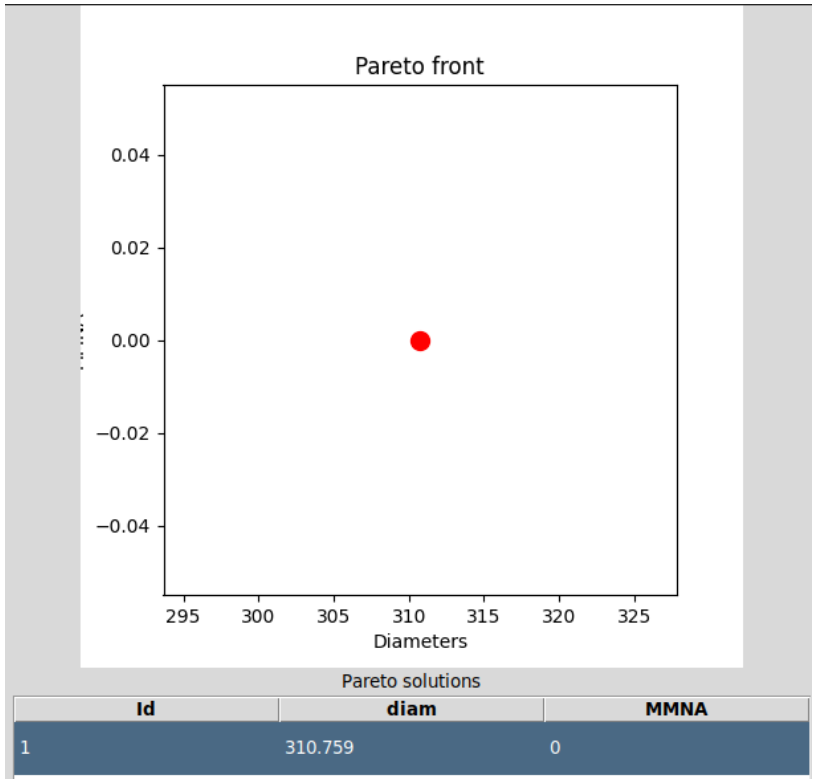


Figure: Front de Pareto pour clustering avec fingerprints

Fingerprint

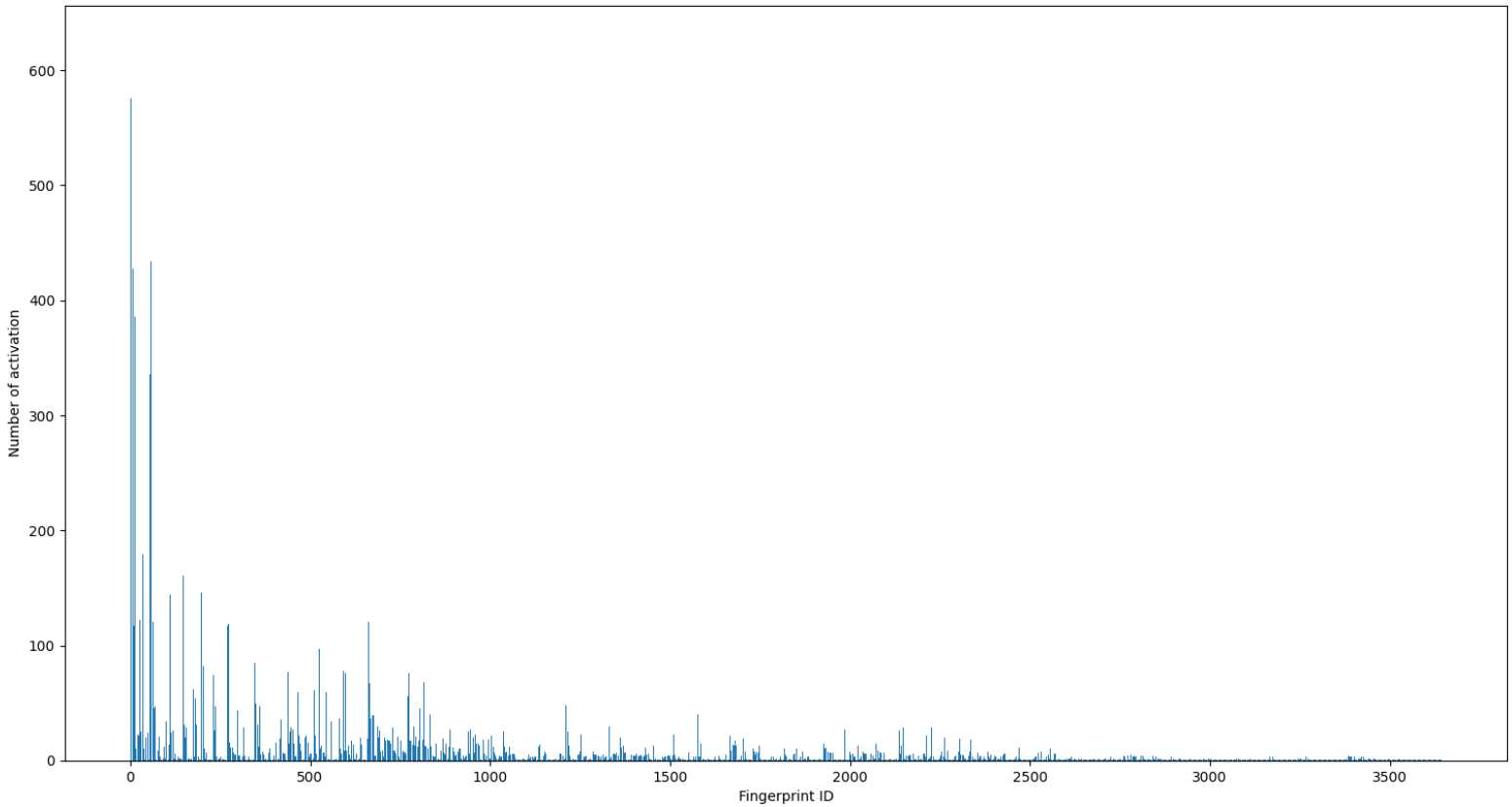


Figure: Nombre d'occurrences du fingerprint (ECFP4)

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

Front de Pareto (k=15, MMNA)

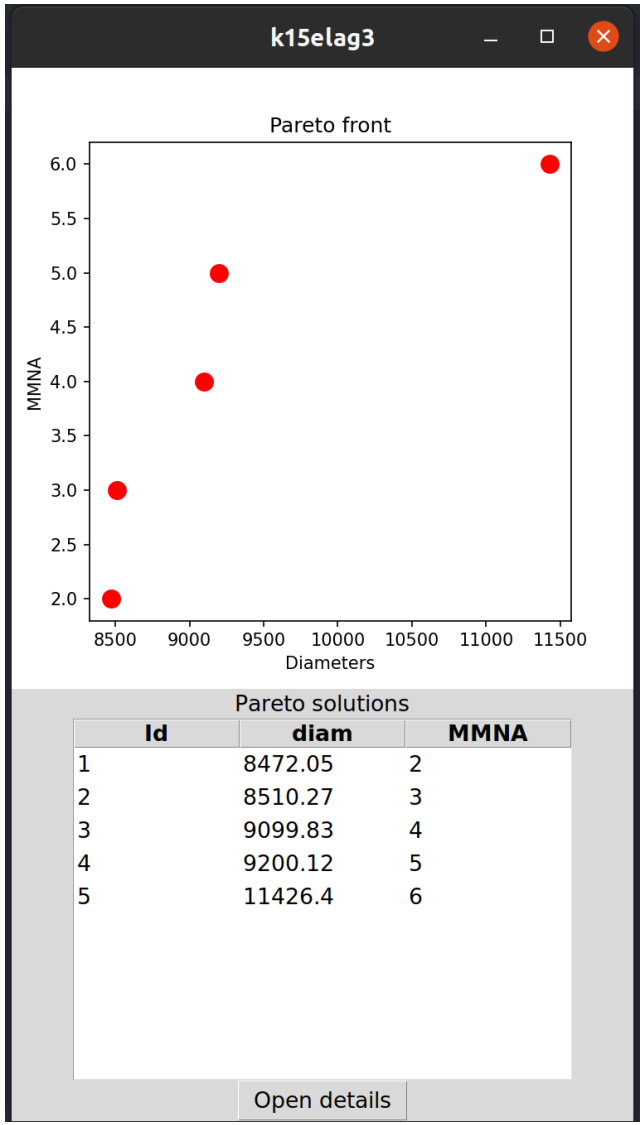


Figure: Première fenêtre

Sommaire

Détails de la partition

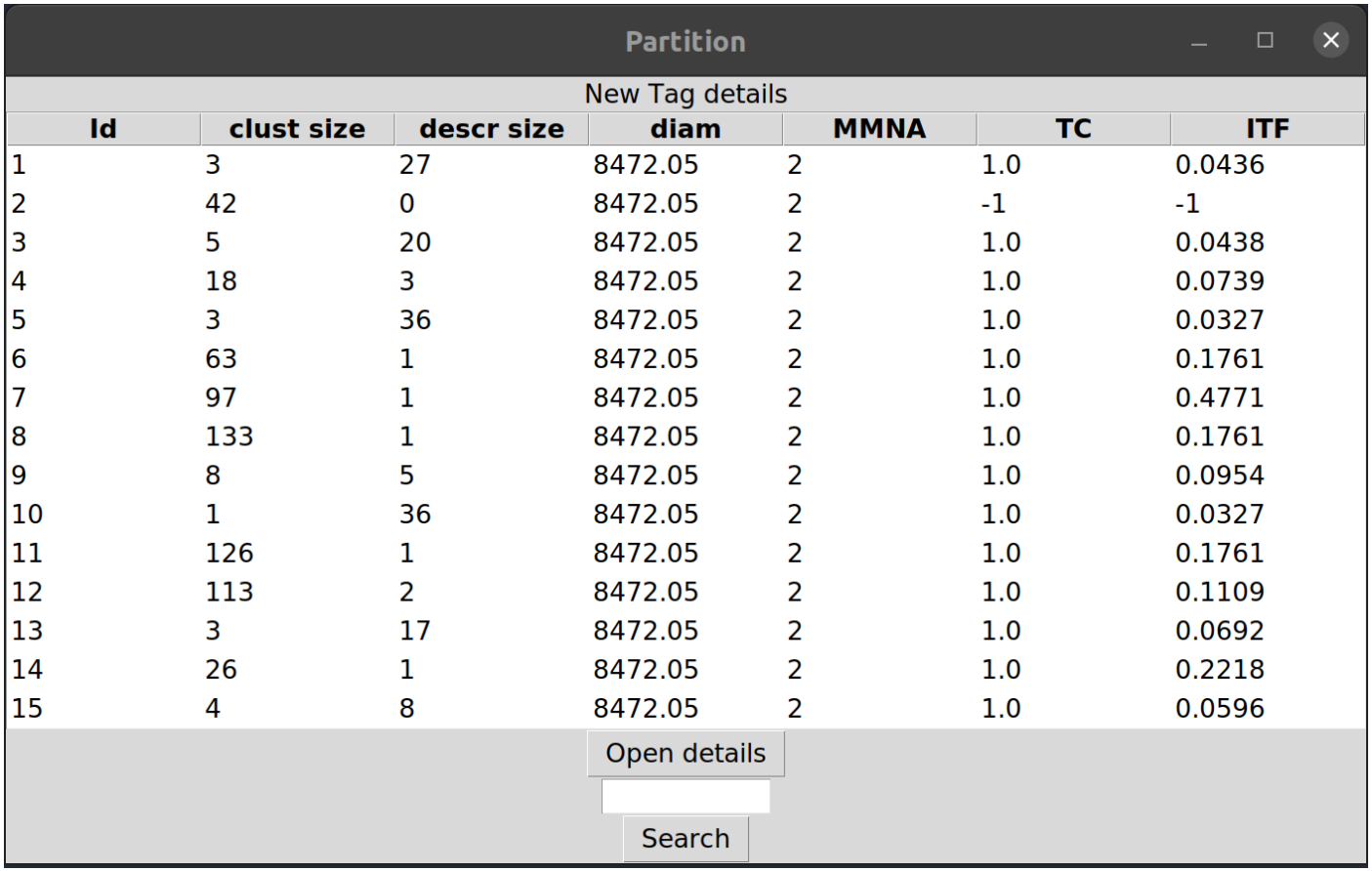


Figure: Deuxième fenêtre

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

Détails sur le cluster 9 avec seuil de 100%

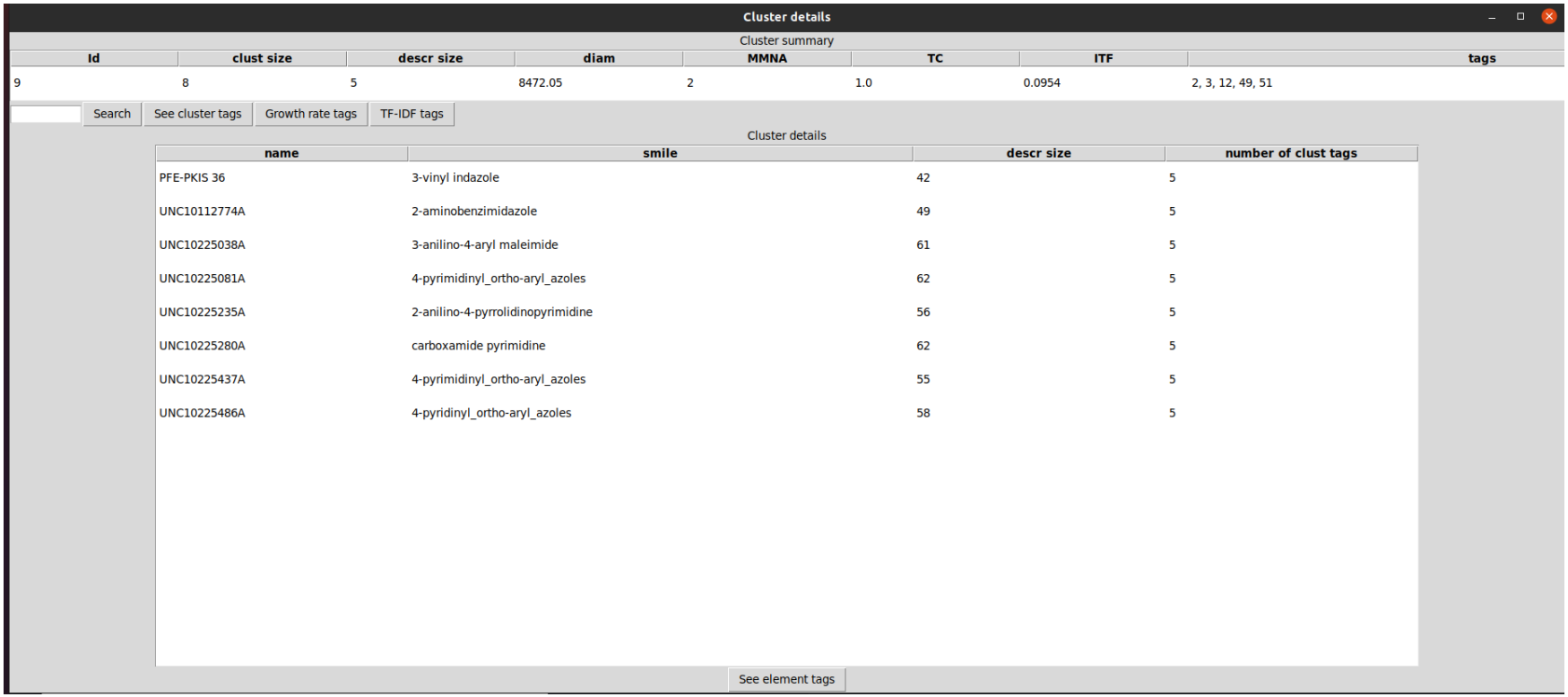


Figure: Liste des molécules d'un cluster

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 contenants 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

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

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

Sommaire

Descriptions des tags

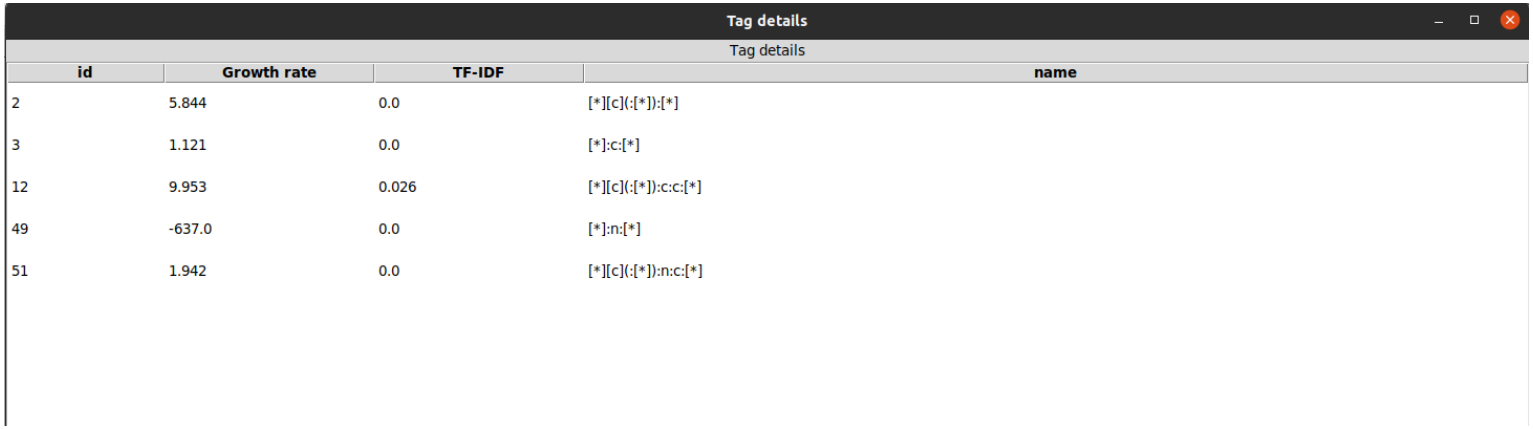


Figure: Descripteurs d'un cluster

Sommaire

Sommaire

Détails de la partition

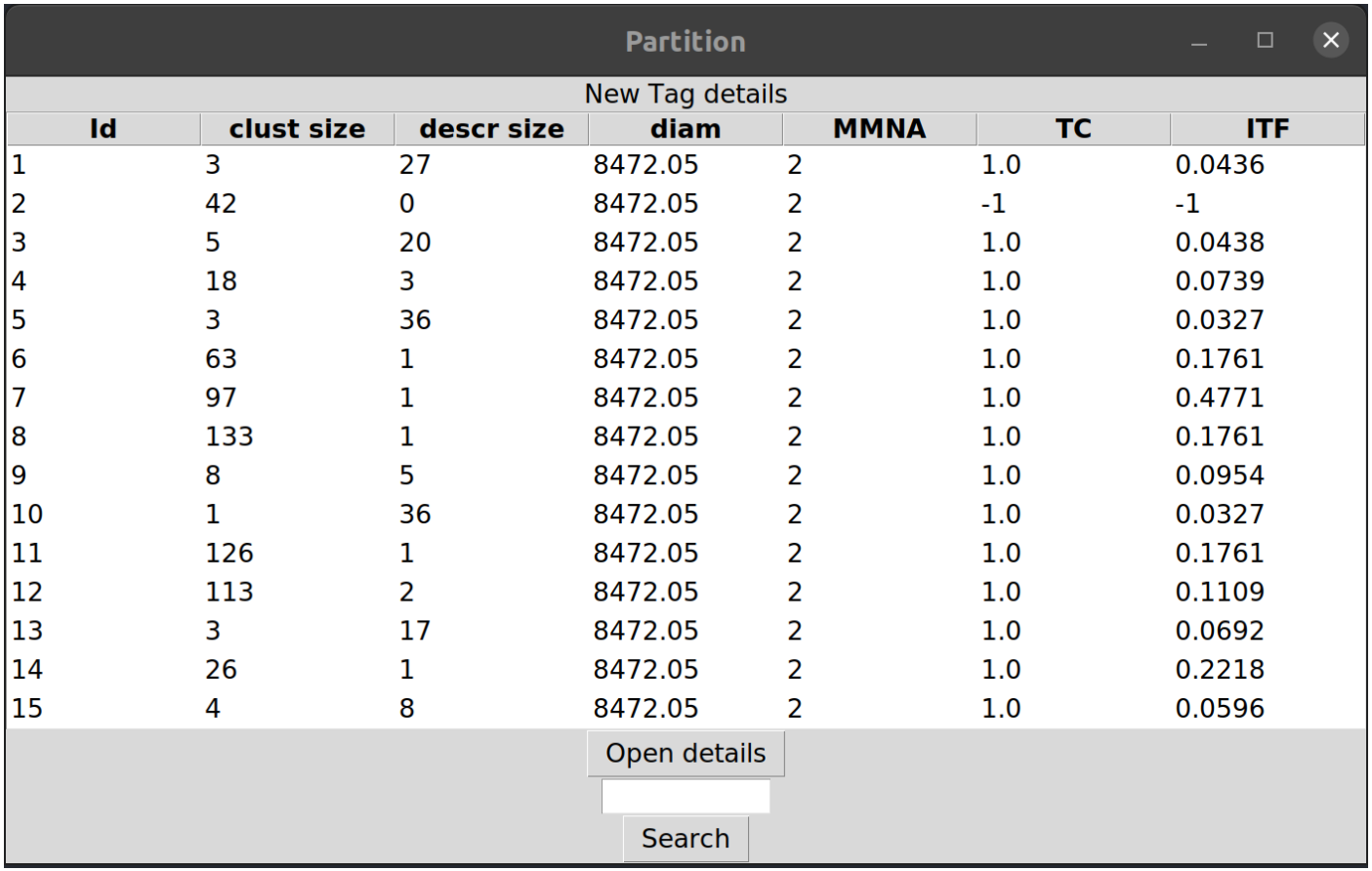
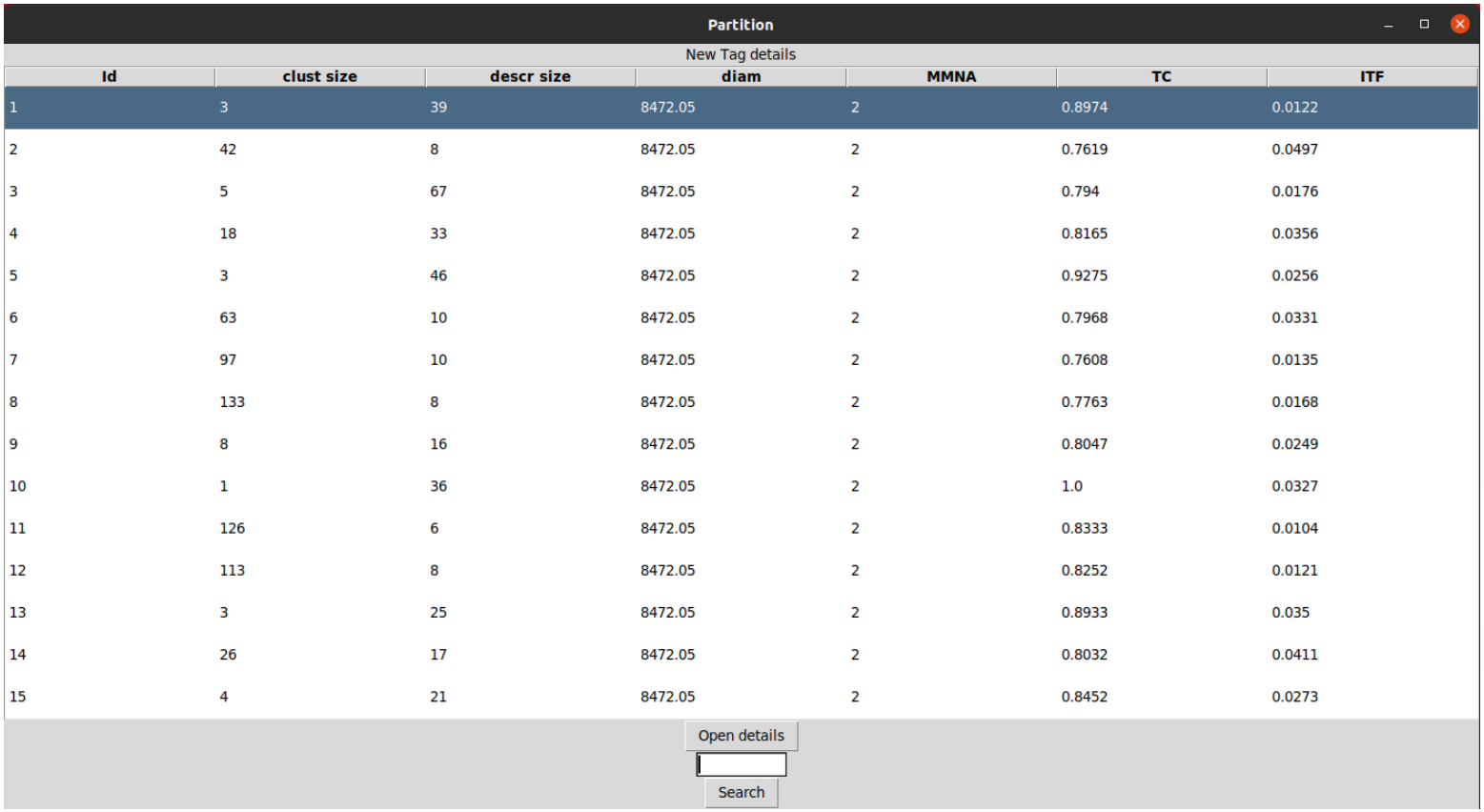


Figure: Deuxième fenêtre

Description des clusters avec un seuil de fréquence



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

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

Sommaire

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) $\rightarrow$ ajout de contraintes ?
  - Ajout de la visualisation des molécules ? Des pharmacophores ?
  - Amélioration du visuel

  $\rightarrow$ Desiderata des chemo-informaticiens
- Utilisation d'autres types de fingeprints?

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 ? $\rightarrow$ 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 ? $\rightarrow$ Discussions nécessaires
- Utiliser les propriétés hiérarchiques des fingerprints pour l'exploration et les explications ? $\rightarrow$ Priorité