Utiliser la diversité moléculaire pour mieux explorer le vivant

Julie Dubois-Chevalier

Cascimodot 11 décembre 2009

Directeurs : Luc Morin-Allory et Christel Vrain

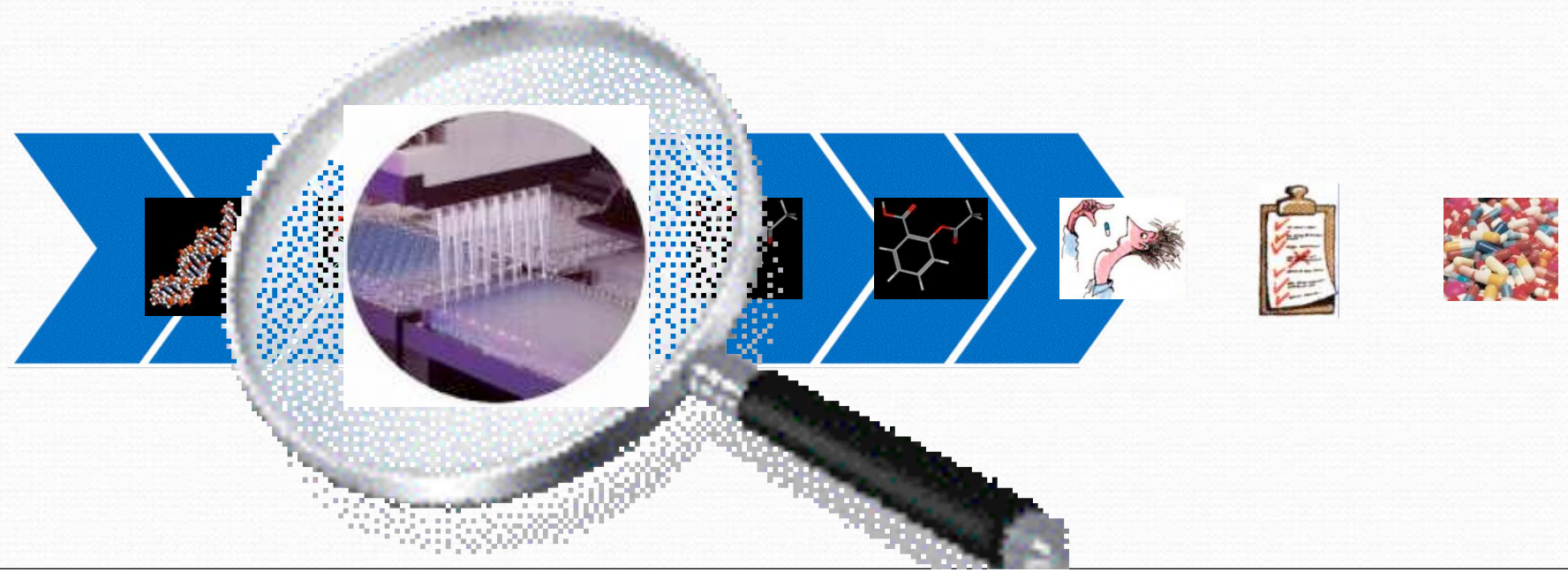
Laboratoires : ICOA / LIFO

Introduction

- Contexte de la thèse : la recherche de médicaments
- Outil logiciel : Screening Assistant
- Problématiques posées
- Comment y répondre :
  - Référentiel
  - Choix des descripteurs
  - Mesure de similarité
  - Nos pistes de travail

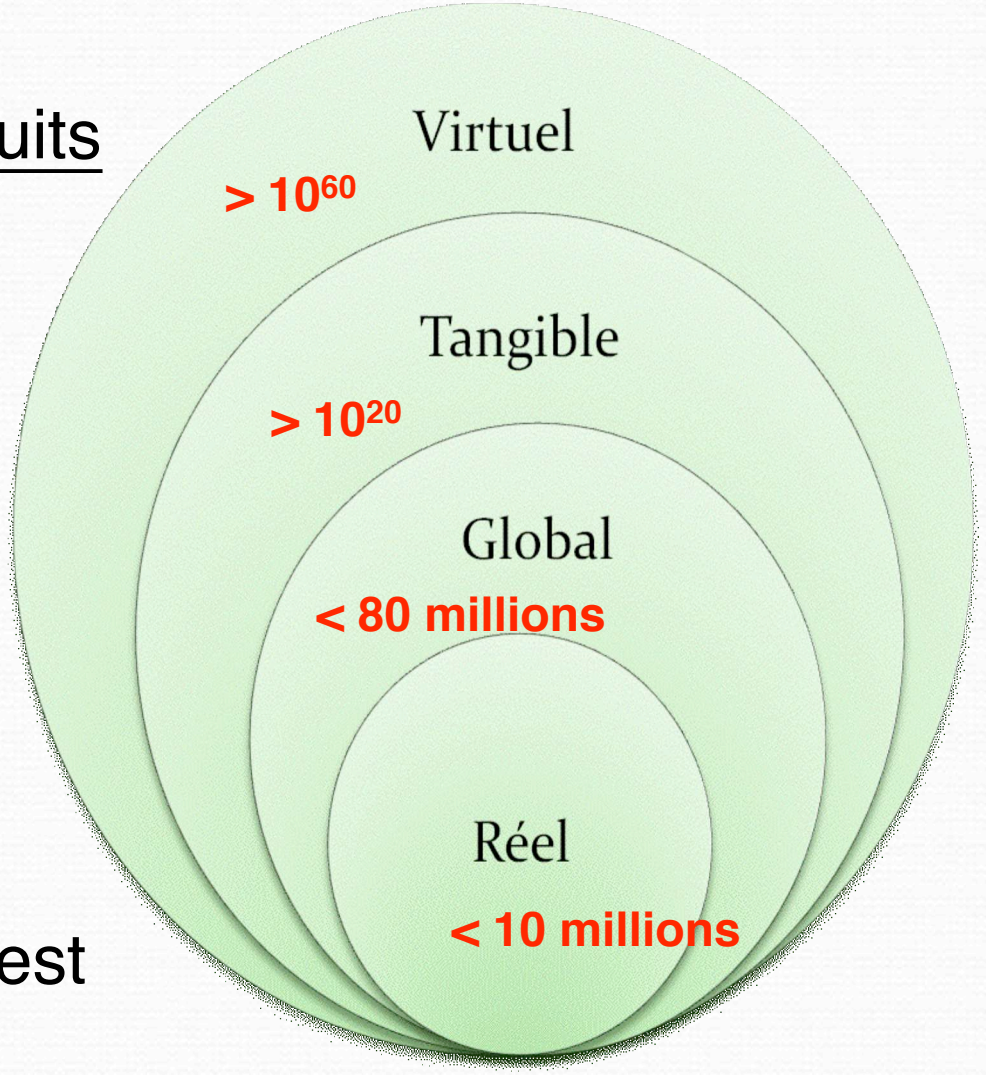
La recherche de médicaments

- 800 millions à 1,5 milliards d'euros pour le développement total d'un médicament
- 15 ans pour arriver à sa mise sur le marché



Pourquoi intervenir avant le criblage ?

Ensembles des produits chimiques

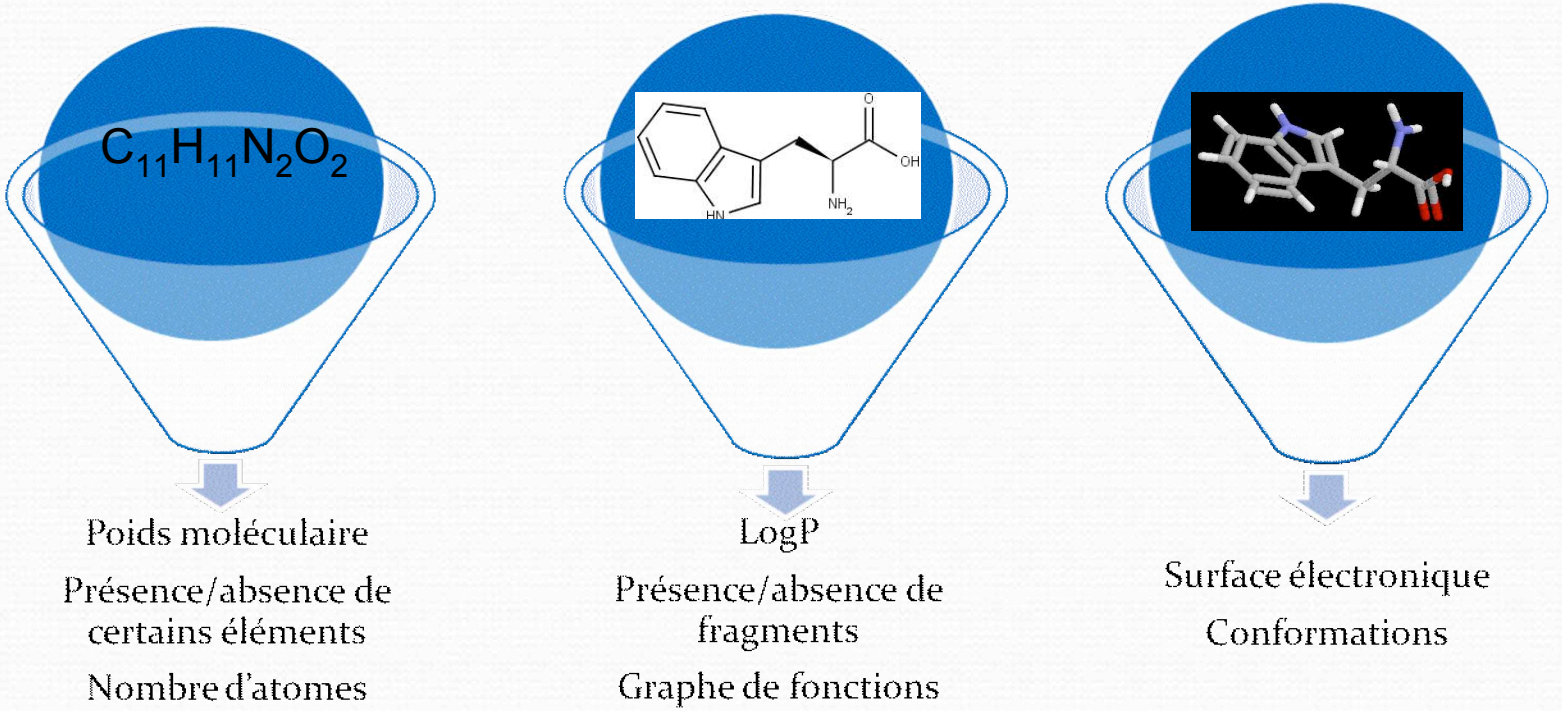


1 grande entreprise

10-20 tests bio/an

100-500 000 molec/test

Vocabulaire et définitions

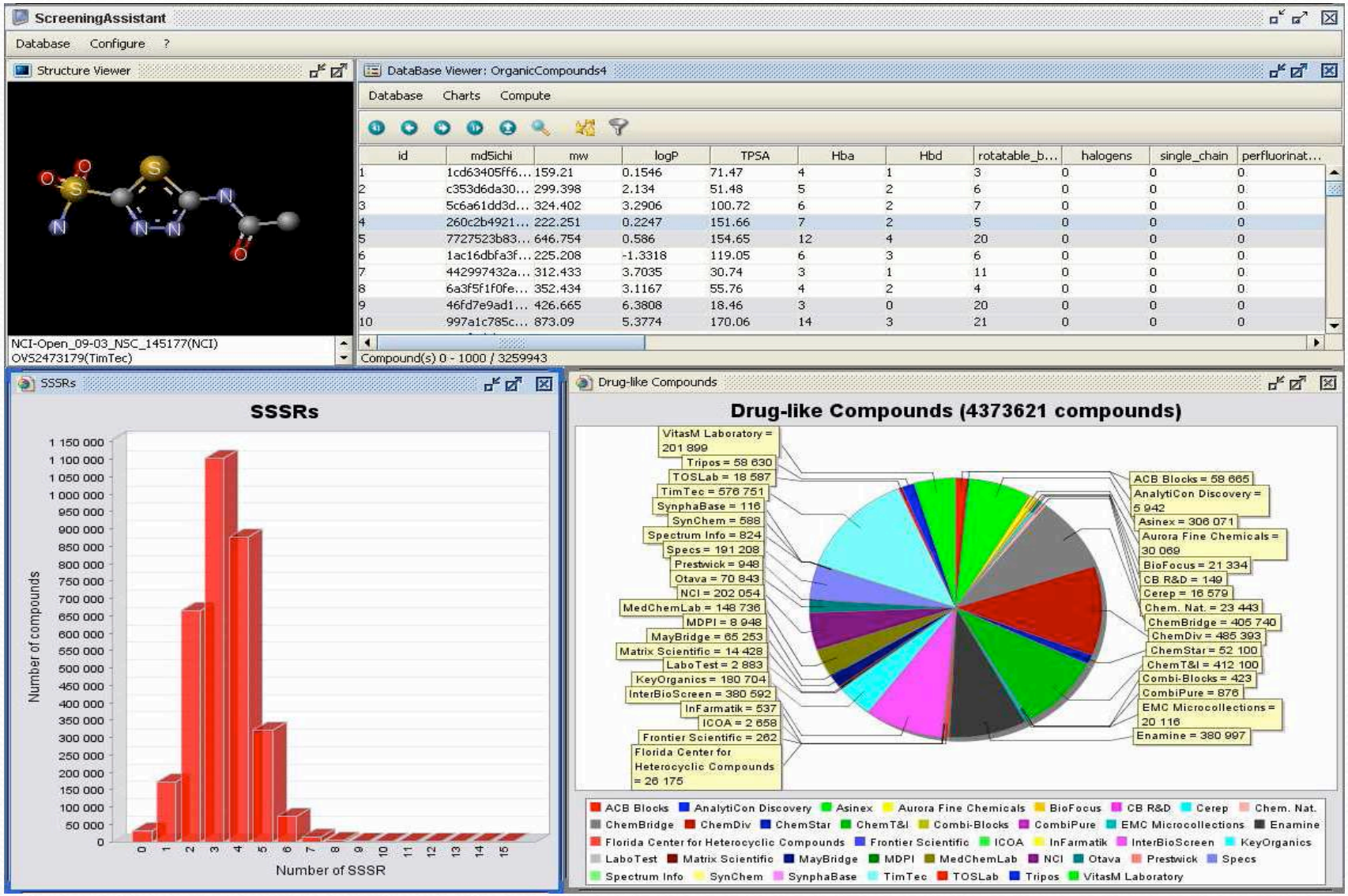


- Molécules= données= observations
- Descripteurs= variables=features
- Collection de composés=chimiothèque=librairie

+ 4 millions de molécules disponibles

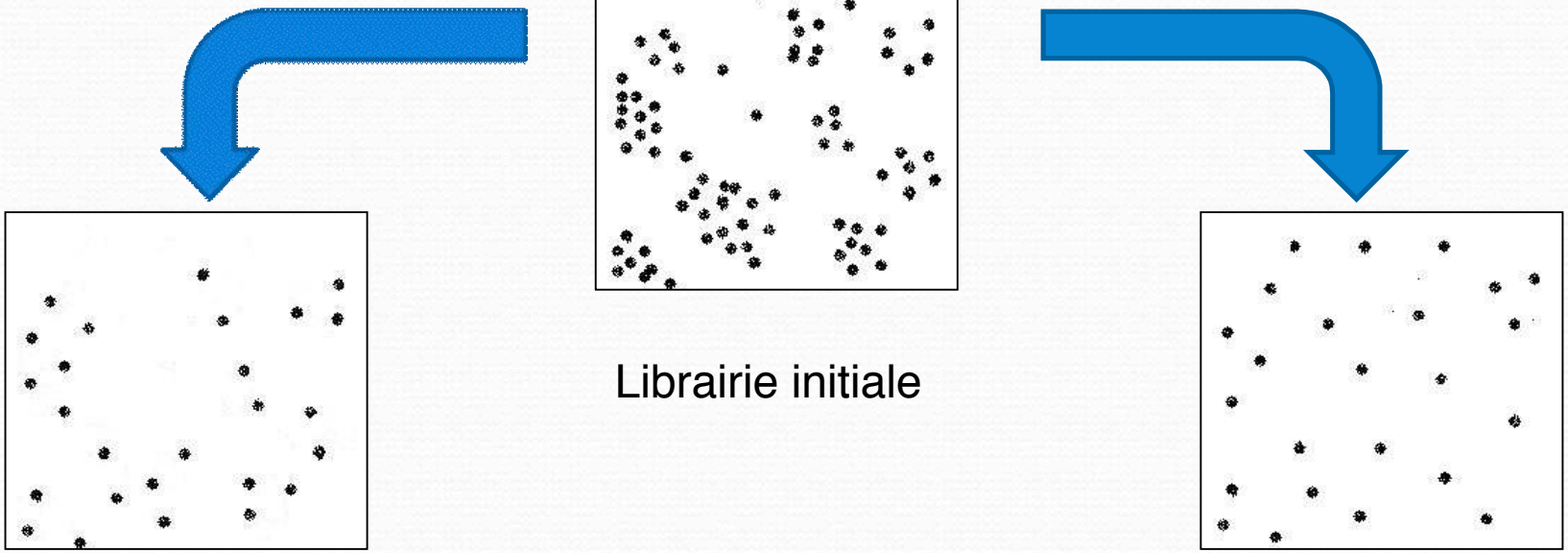
+ 2000 descripteurs calculables

Outil logiciel : Screening Assistant



Les problématiques posées (1)

Sélection par diversité par rapport à une librairie de départ



Librairie initiale

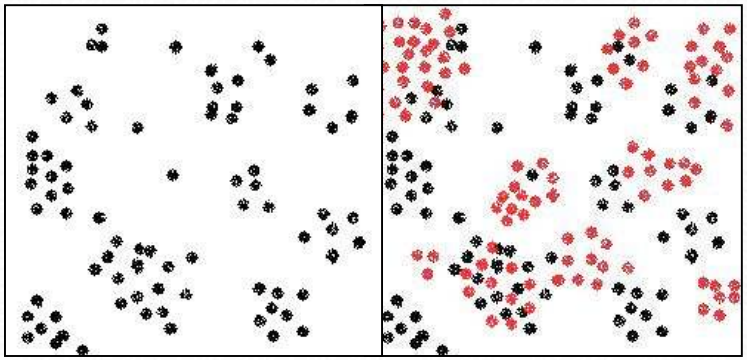
Sous-ensemble représentatif

Sous-ensemble divers

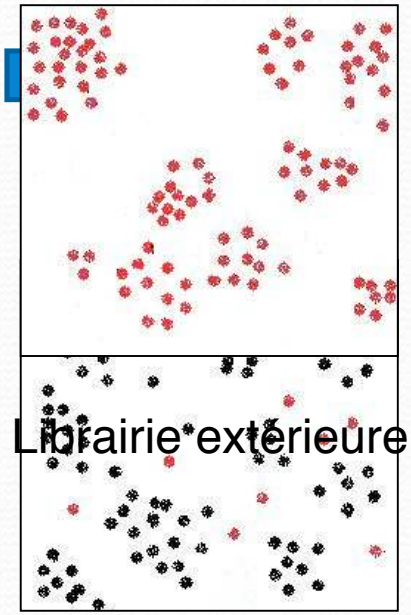
On sait déjà plus ou moins bien le faire

Les problématiques posées (2)

Complément de librairie par diversité



Librairie initiale



Librairie extérieure

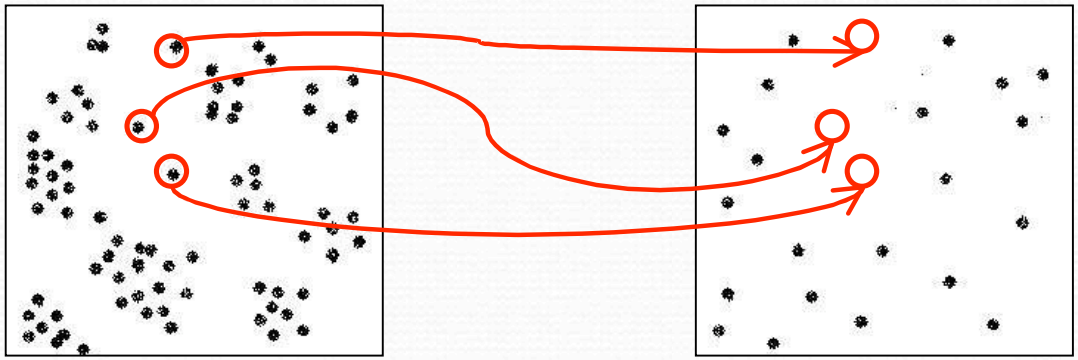
Librairie finale

Aucune méthode efficace connue

Les problématiques posées (3)

Les produits sélectionnés doivent avoir des voisins

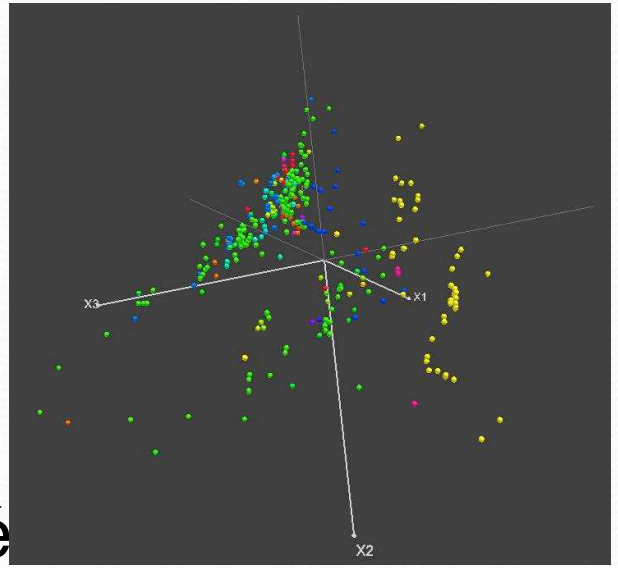
Utile pour faire des tests rapidement avec des produits internes



Problème jamais étudié

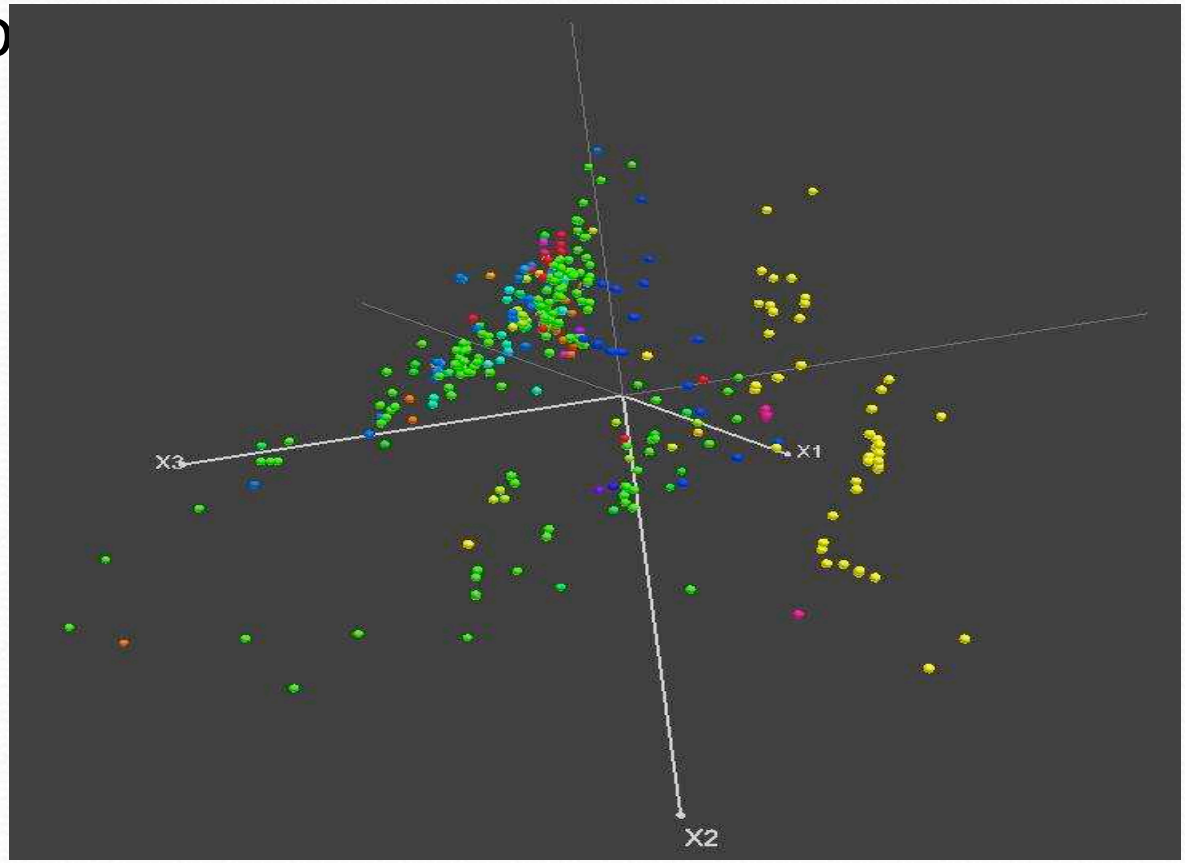
Ce qu'il faut définir pour y répondre

- Espace de référence
- Descripteurs à utiliser
- Choix de la mesure de similarité
- Définition de la diversité
- Critère d'évaluation de cette diversité
- Algorithme de sélection efficace sur de grands volumes de données



L’espace de référence

- Espace chimique = espace à n dimensions (= n descripteurs)
- Amélio



Choix des descripteurs : types

- Types de descripteurs

Fingerprints	• Vecteurs de bits compressés
Propriétés Physico-chimique	• Variables numériques continues • Variables numériques discrètes
Particularités	• Binaires (is_peptide)
Topologie	• Graphe • Matrice de connectivité

- Cadre de l'étude : non supervisé (peu étudié)

↳ Pas d'information de classe pour vérifier la pertinence de la sélection

Choix des descripteurs : sélection

- Sélection à partir de 2000 descripteurs
  - Clustering de descripteurs[1]
  - Shannon-Entropy (contenu en information des descripteurs)[2]
  - Sélection aléatoire basée sur une pondération des descripteurs les plus importants[3] (en vue d'une ACP)
- Validation du jeu réduit de descripteurs
  - Entropie avec la similarité[4]
  - Shannon-Entropy
  - Representation Entropy[1]

1. **Mitra et al.** : Unsupervised Feature Selection Using Feature Similarity, *IEEE transactions on pattern analysis and machine intelligence*, **2002**,24,n°3

2. Godden and Bajorath : An information-Theoritic Approach to Descriptor Selection for Database Profiling and QSAR Modeling, *QSAR Comb. Sci. ,* **2003**, *22*, p487-497

3.Boutsidis et al. : Unsupervised Feature Selection for Principal Component analysis, 2008, *KDD'08*

4. Dash and Liu : Feature Selection for Clustering, *Proceedings of fourth pacific-asia conference on knowledge discovery and data-mining,*2000

Choix des descripteurs : l'entropie

Soit l'ensemble I de n molécules : (i1,i2…,in)
L'ensemble D de m descripteurs : (d1,d2…,dm) pouvant chacun prendre l valeurs

- **Entropie de Shannon pour un descripteur**

$$H(d) = -\sum_{k=1}^{l} Pk \log Pk$$

où Pk est la probabilité d'apparition de la valeur k pour le descripteur d

- **Entropie modifiée de Dash et Liu**

$$H(D) = -\sum_{i1=1}^{n}\sum_{i2=1}^{n}(S_{i1,i2} \times \log S_{i1,i2} + (1 - S_{i1,i2}) \times \log(1 - S_{i1,i2}))$$

Où $S_{i1,i2} = e^{-\alpha \times Dis\tan ce_{i1,i2}}$

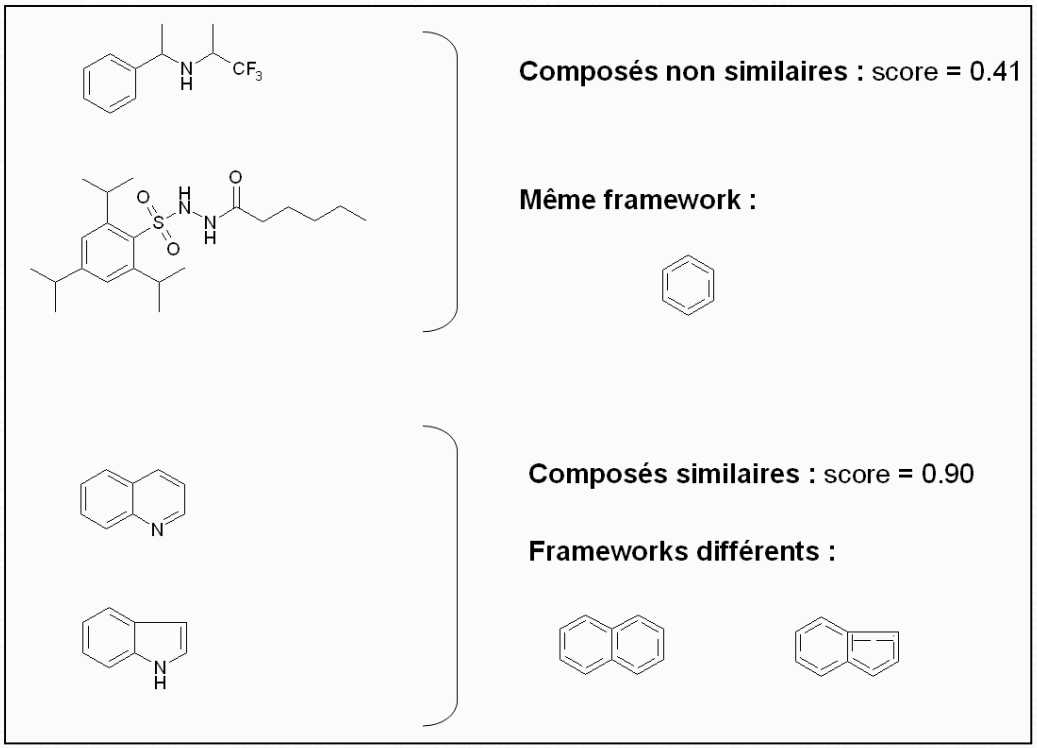
- **Entropie modifiée de Mitra et al.**

$$H_R(D) = -\sum_{j=1}^{m} \tilde{\lambda}_j \times \log \tilde{\lambda}_j \qquad \text{où} \quad \tilde{\lambda}_j = \frac{\lambda_j}{\sum_{j=1}^{m} \lambda_j}$$

λ_j valeur propre de la matrice de covariance m x m

Mesure de similarité

- Pour nos descripteurs, les distances (euclidienne, de Hamming, Tanimoto…) sont plus ou moins adaptées
- Prise en compte de sous-structures, de graphes



Pistes de travail actuelles

- Divers critères de validation basés sur
  - la représentation des squelettes
  - des indices de taux de dissimilarité (_ dist)
- Méthodes de sélection existantes :
  - Cluster-based selection (k-means)
  - Partition-based selection
  - Dissimilarity-based selection
  - Optimization-based selection
  - Autres approches avec algorithmes génétiques, réseaux de neurones…
- Aucune méthode poussée sur le complément de librairies

Merci de votre attention

Avez-vous des questions ?

Pour toute suggestion, piste, remarque n'hésitez pas

julie.dubois@univ-orleans.fr