

Utiliser la diversité moléculaire pour mieux explorer le vivant

Julie Dubois-Chevalier

Cascimodot 11 décembre 2009

Directeurs : Luc Morin-Allory et Christel Vrain

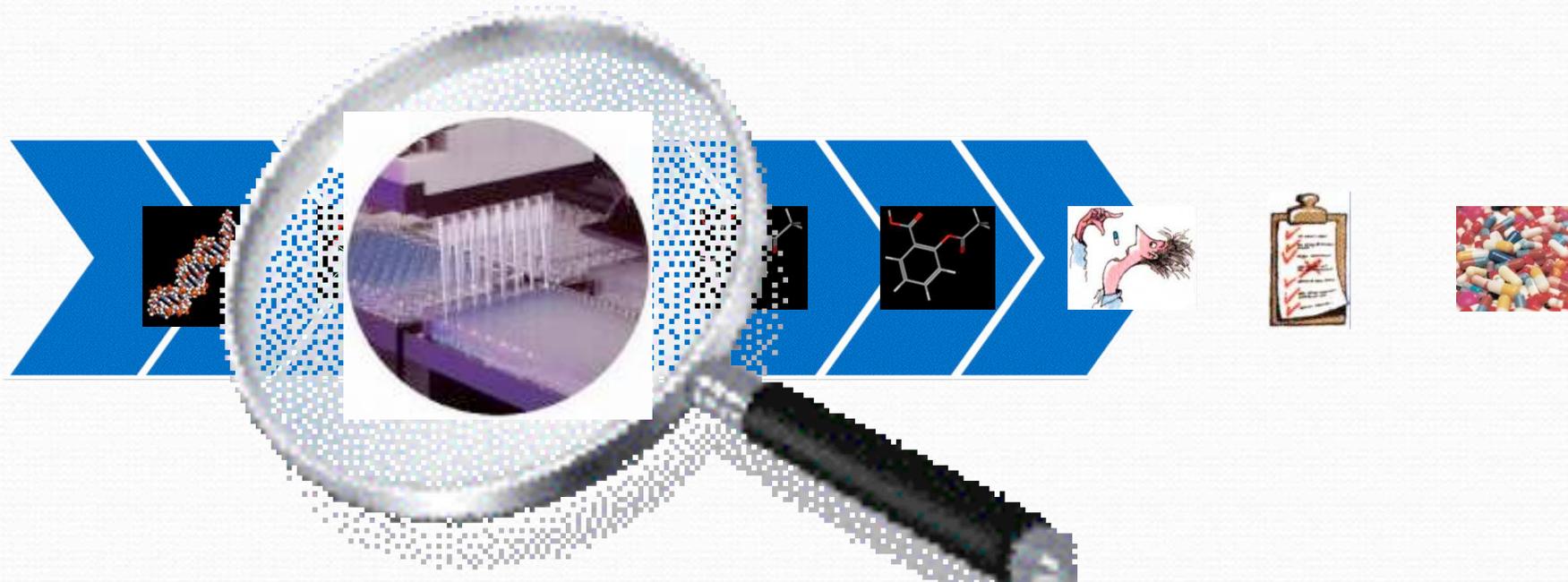
Laboratoires : ICOA / LIFO

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

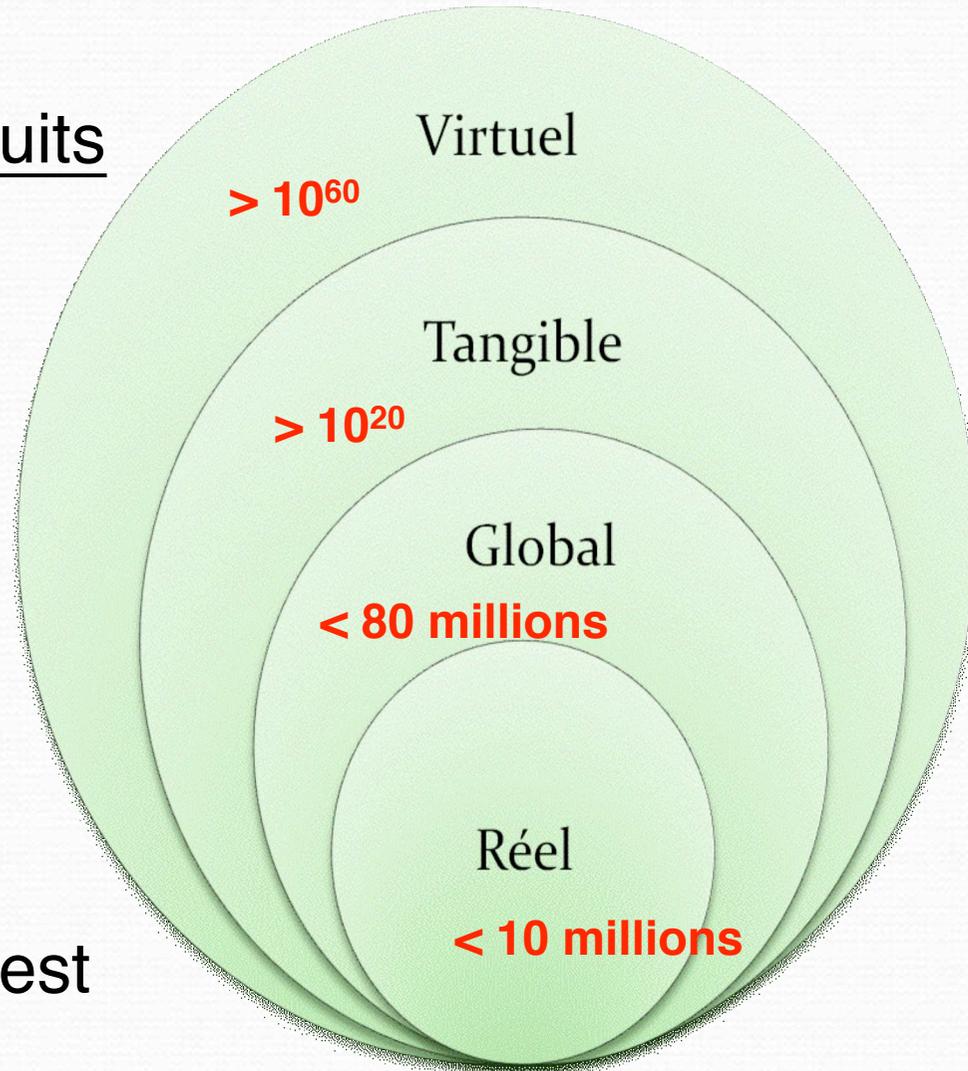
- 800 millions à 1,5 milliards d'euros pour le développement total d'un médicament
- 15 ans pour arriver à sa mise sur le marché



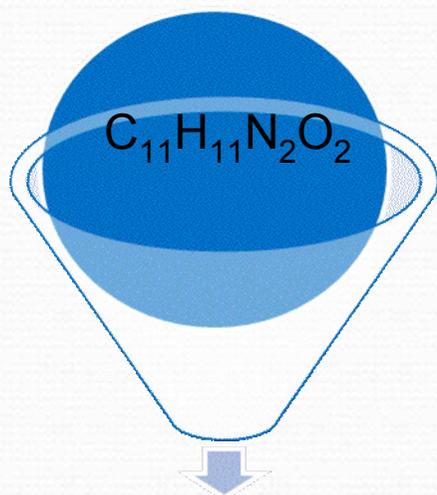
Pourquoi intervenir avant le criblage ?

Ensembles des produits chimiques

1 grande entreprise
10-20 tests bio/an
100-500 000 molec/test



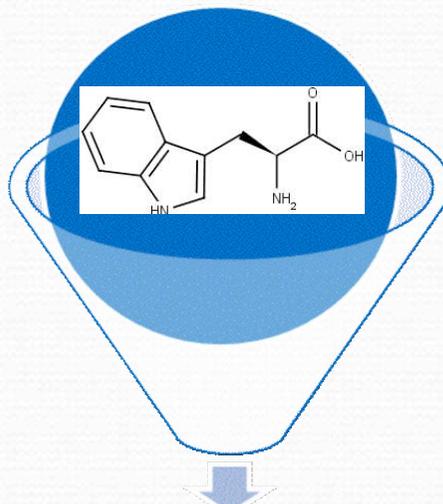
Vocabulaire et définitions



Poids moléculaire

Présence/absence de
certains éléments

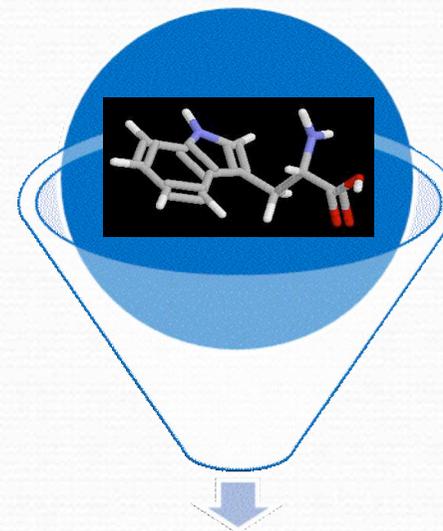
Nombre d'atomes



LogP

Présence/absence de
fragments

Graphe de fonctions



Surface électronique

Conformations

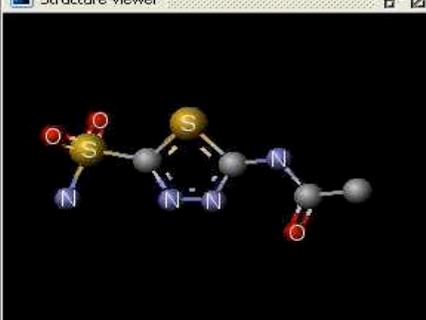
- Molécules= données= observations
 - Descripteurs= variables=features
 - Collection de composés=chimiothèque=librairie
- + 4 millions de molécules disponibles**
+ 2000 descripteurs calculables

Outil logiciel : Screening Assistant

ScreeningAssistant

Database Configure ?

Structure Viewer



NCI-Open_09-03_N5C_145177(NCI)
OY52473179(TimTec)

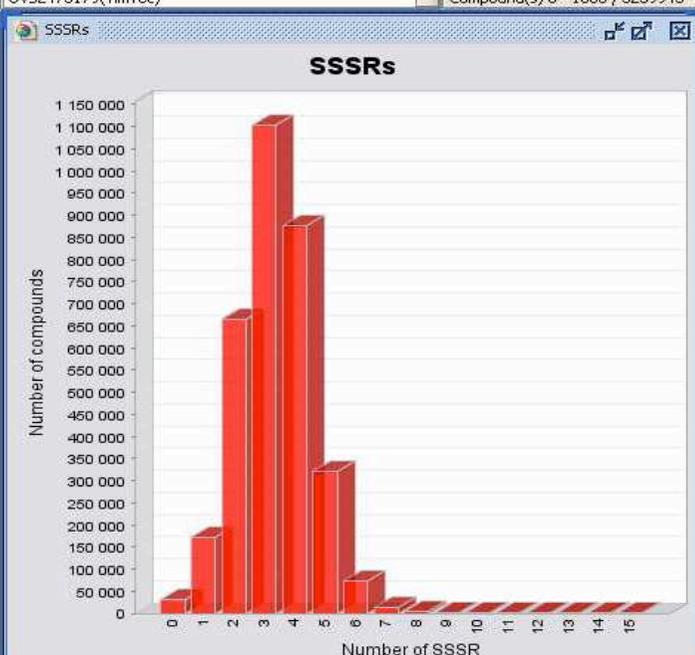
DataBase Viewer: OrganicCompounds4

Database Charts Compute

id	md5ichi	mw	logP	TPSA	Hba	Hbd	rotatable_b...	halogens	single_chain	perfluorinat...
1	1cd63405ff6...	159.21	0.1546	71.47	4	1	3	0	0	0
2	c353d6da30...	299.398	2.134	51.48	5	2	6	0	0	0
3	5c6a61dd3d...	324.402	3.2906	100.72	6	2	7	0	0	0
4	260c2b4921...	222.251	0.2247	151.66	7	2	5	0	0	0
5	7727523b83...	646.754	0.586	154.65	12	4	20	0	0	0
6	1ac16dbfa3f...	225.208	-1.3318	119.05	6	3	6	0	0	0
7	442997432a...	312.433	3.7035	30.74	3	1	11	0	0	0
8	6a3f9f1f0fe...	352.434	3.1167	55.76	4	2	4	0	0	0
9	46fd7e9ad1...	426.665	6.3808	18.46	3	0	20	0	0	0
10	997a1c785c...	873.09	5.3774	170.06	14	3	21	0	0	0

Compound(s) 0 - 1000 / 3259943

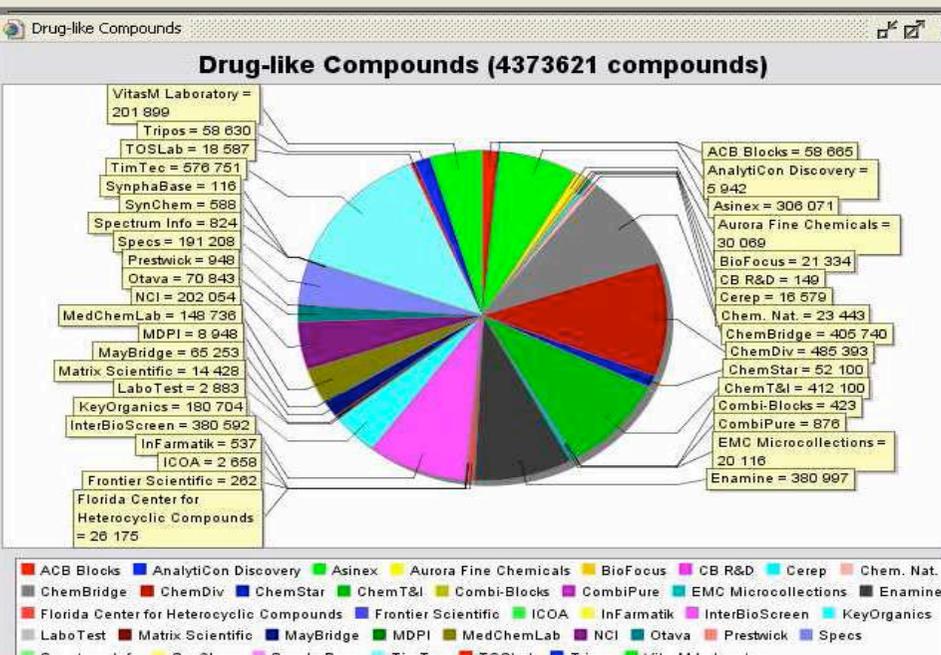
SSSRs



Number of compounds

Number of SSSR

Drug-like Compounds (4373621 compounds)



VitasM Laboratory = 201 899
Tripos = 58 630
TOSLab = 18 587
TimTec = 576 751
SynphaBase = 118
SynChem = 588
Spectrum Info = 824
Specs = 191 208
Prestwick = 948
Otava = 70 843
NCI = 202 054
MedChemLab = 148 738
MDPI = 8 948
MayBridge = 65 253
Matrix Scientific = 14 428
LaboTest = 2 883
KeyOrganics = 180 704
InterBioScreen = 380 592
InFarmatik = 537
ICOA = 2 658
Frontier Scientific = 262
Florida Center for Heterocyclic Compounds = 26 175

ACB Blocks = 58 665
AnalytiCon Discovery = 5 942
Asinex = 306 071
Aurora Fine Chemicals = 30 069
BioFocus = 21 334
CB R&D = 149
Cerep = 18 579
Chem. Nat. = 23 443
ChemBridge = 405 740
ChemDiv = 485 393
ChemStar = 52 100
ChemT&I = 412 100
Combi-Blocks = 423
CombiPure = 876
EMC Microcollections = 20 116
Enamine = 380 997

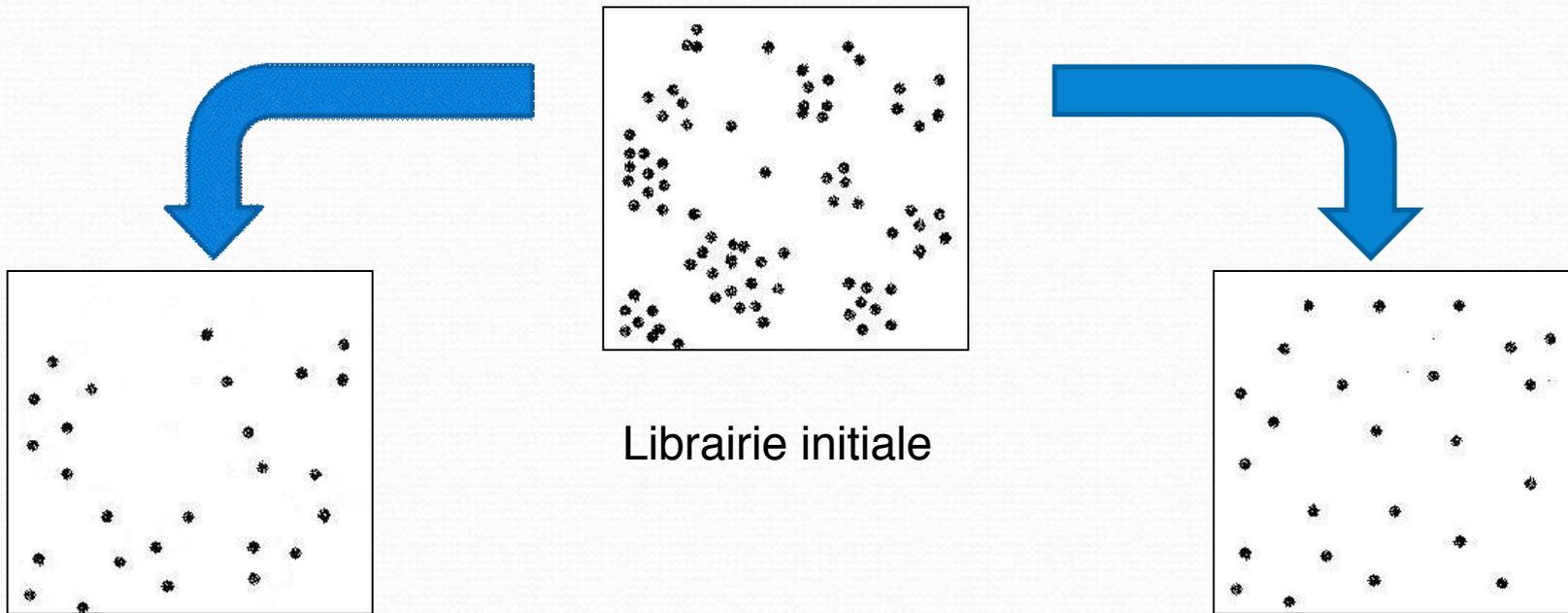
Legend:

- ACB Blocks
- AnalytiCon Discovery
- Asinex
- Aurora Fine Chemicals
- BioFocus
- CB R&D
- Cerep
- Chem. Nat.
- ChemBridge
- ChemDiv
- ChemStar
- ChemT&I
- Combi-Blocks
- CombiPure
- EMC Microcollections
- Enamine
- Florida Center for Heterocyclic Compounds
- Frontier Scientific
- ICOA
- InFarmatik
- InterBioScreen
- KeyOrganics
- LaboTest
- Matrix Scientific
- MayBridge
- MDPI
- MedChemLab
- NCI
- Otava
- Prestwick
- Specs
- Spectrum Info
- SynChem
- SynphaBase
- TimTec
- TOSLab
- Tripos
- VitasM Laboratory

Les problématiques posées

(1)

Sélection par diversité par rapport à une librairie de départ



Sous-ensemble
représentatif

Librairie initiale

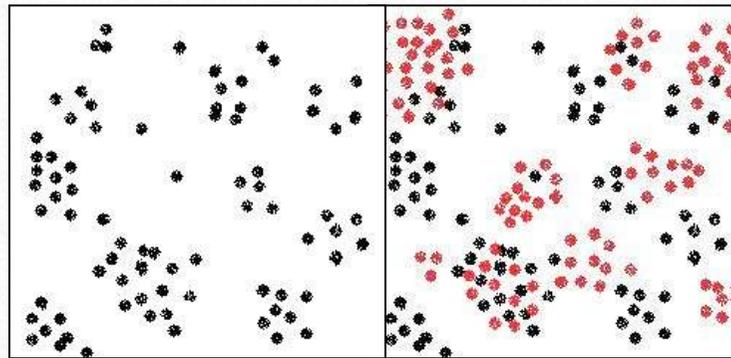
Sous-ensemble
divers

On sait déjà plus ou moins bien le faire

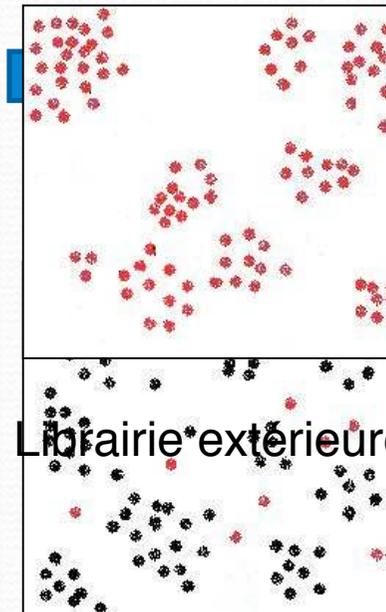
Les problématiques posées

(2)

Complément de librairie par diversité



Librairie initiale



Librairie extérieure

Librairie finale

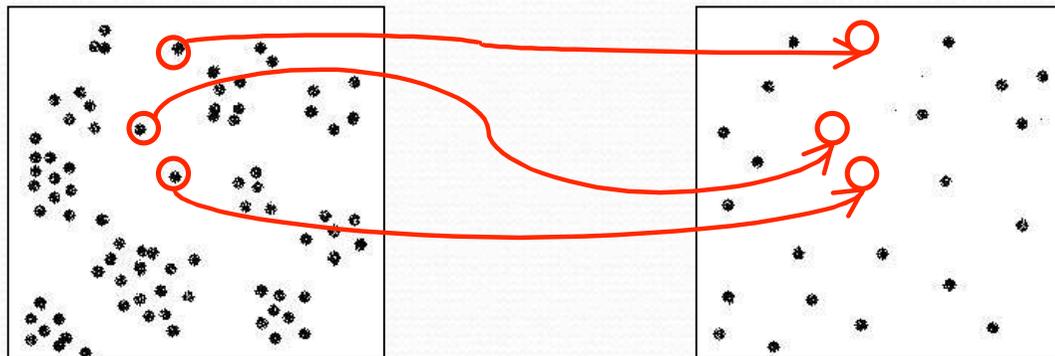
Aucune méthode efficace connue

Les problématiques posées

(3)

Les produits sélectionnés doivent avoir des voisins

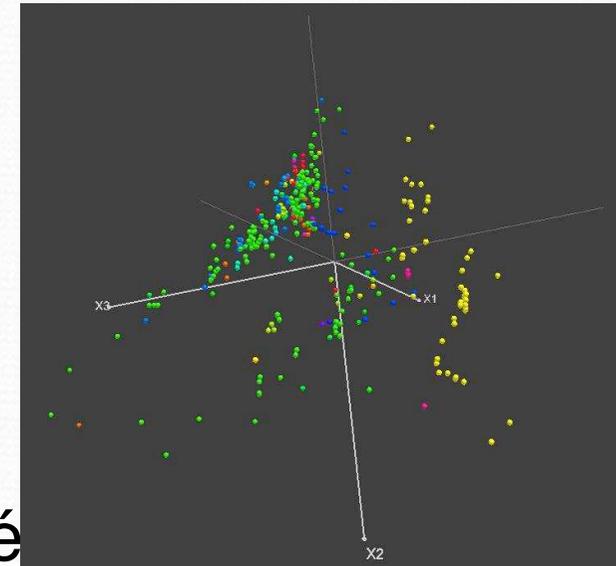
Utile pour faire des tests rapidement avec des produits internes



Problème jamais étudié

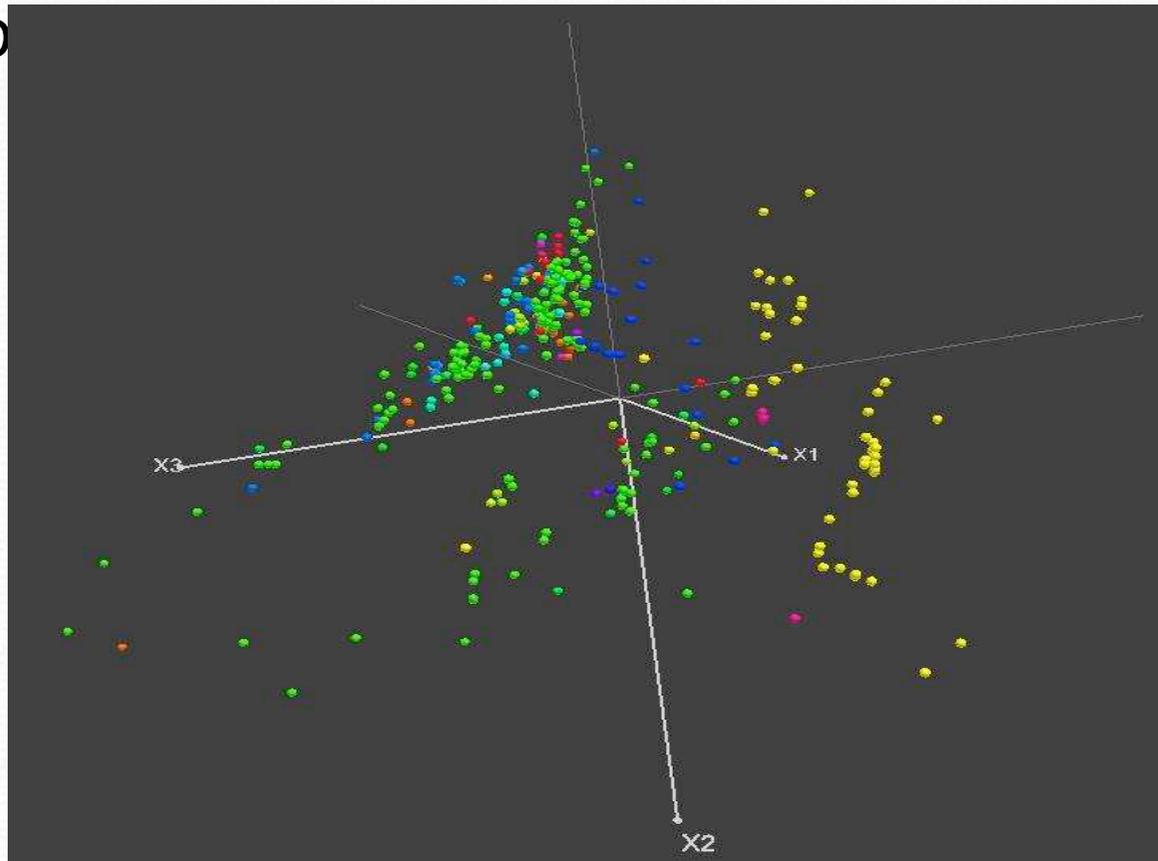
Ce qu'il faut définir pour y répondre

- 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)
- Amélior



Choix des descripteurs :

types

- Types de descripteurs

Fingerprints	<ul style="list-style-type: none">• Vecteurs de bits compressés
Propriétés Physico-chimique	<ul style="list-style-type: none">• Variables numériques continues• Variables numériques discrètes
Particularités	<ul style="list-style-type: none">• Binaires (is_peptide)
Topologie	<ul style="list-style-type: none">• 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 descripteurs¹
 - Shannon-Entropy (contenu en information des descripteurs)²
 - Sélection aléatoire basée sur une pondération des descripteurs les plus importants³ (en vue d'une ACP)
- Validation du jeu réduit de descripteurs
 - Entropie avec la similarité⁴
 - Shannon-Entropy
 - Representation Entropy¹

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-Theoretic 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**

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

où P_k 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 $S_{i1,i2} = e^{-\alpha \times \text{Distance}_{i1,i2}}$

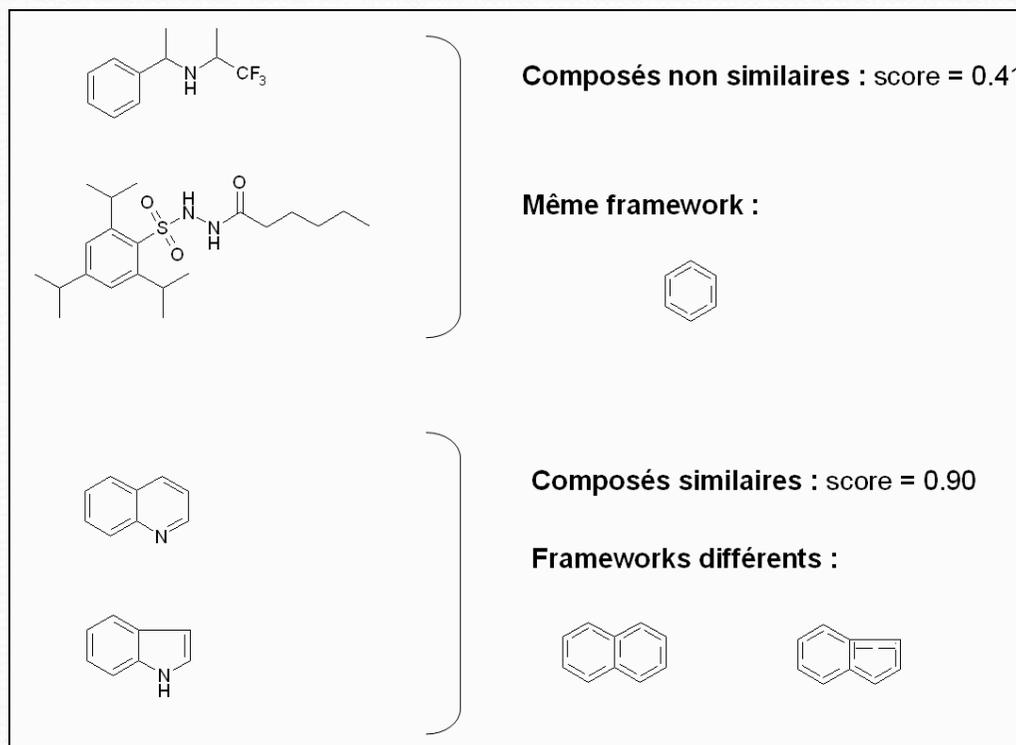
- **Entropie modifiée de Mitra et al.**

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

λ_j valeur propre de la matrice de covariance m x m

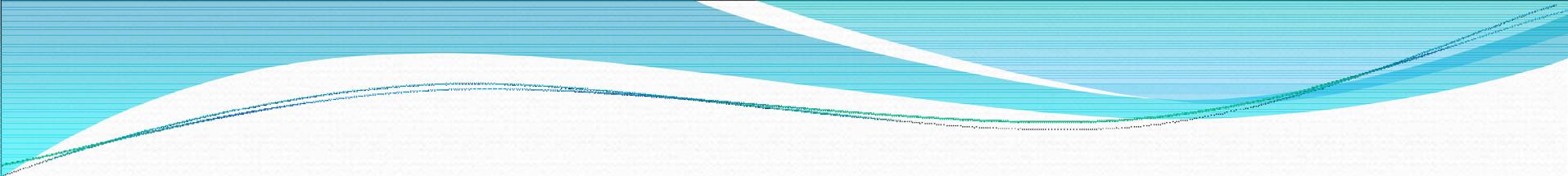
Mesure de similarité

- 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 bibliothèques



Merci de votre attention

Avez-vous des questions ?

Pour toute suggestion, piste, remarque n'hésitez pas

julie.dubois@univ-orleans.fr