Chimiothèque : vers une approche rationnelle pour la sélection de sous-chimiothèques

Julie Dubois Dubois-Chevalier

PUCVL
Pôle Universités
Centre Val de Loire

UNIVERSITÉ D'ORLÉANS

UNIVERSITE D'ORLEANS

ÉCOLE DOCTORALE SCIENCES ET TECHNOLOGIES

LABORATOIRES : ICOA, UMR CNRS 6005 et LIFO, EA 4022

THÈSE présentée par :

Julie DUBOIS-CHEVALIER

soutenue le : **7 décembre 2011**

pour obtenir le grade de : **Docteur de l'université d'Orléans**

Discipline : **Chimie**

Chimiothèques ; vers une approche rationnelle pour la sélection de sous-chimiothèques

THÈSE dirigée par :

Luc MORIN-ALLORY	Professeur, Université d'Orléans
Christel VRAIN	Professeur, Université d'Orléans

RAPPORTEURS :

Ronan BUREAU	Professeur, Université de Caen
Antoine CORNUEJOLS	Professeur, AgroParisTech

JURY :

Ronan BUREAU	Professeur, Université de Caen, Président du jury
Antoine CORNUEJOLS	Professeur, AgroParisTech
Luc MORIN-ALLORY	Professeur, Université d'Orléans
Christel VRAIN	Professeur, Université d'Orléans

A Guillaume, Axelle, Clément, ma famille et mes amis.

Remerciements

Je tiens tout d'abord à exprimer ma profonde reconnaissance au Professeur Luc Morin-Allory et au Professeur Christel Vrain qui ont dirigé cette thèse. Merci pour leur implication, leur confiance, leur grande disponibilité et leurs conseils scientifiques qui m'ont permis de mener ce projet à terme.

J'adresse ma gratitude au Professeur Ronan Bureau et au Professeur Antoine Cornuéjols pour avoir accepté de juger ce travail de thèse et pour le temps qu'ils y ont consacré.

Je remercie également l'équipe de chémoinformatique et l'équipe d'apprentissage pour leur accueil, leurs conseils et leur bonne humeur : Stéphane Bourg, Laurent Robin, Vincent Le Guilloux, Lionel Colliandre, Véronique Hamon, Christophe Marot, Matthieu Exbrayat, Lionel Martin, Guillaume Cleuziou, Frédéric Moal, Matthieu Lopez, Jacques-Henri Sublemontier et Sylvie Billot. Je tiens à adresser un merci particulier à Stéphane Bourg pour l'aide qu'il m'a apporté notamment pour la rédaction de mon premier article et pour le temps qu'il y a consacré ; ainsi qu'à Lionel Martin pour sa collaboration sur ce projet, pour les très longues discussions et surtout les questionnements qui ont fait avancer notre définition de la diversité et pour le développement de l'algorithme.

Enfin, merci aux membres de l'ICOA et du LIFO pour leur accueil chaleureux et leur sympathie.

Table des matières

Introduction

Le processus de découverte et de développement d'un médicament est constitué de nombreuses étapes :

- Découverte de la cible (gène, protéine ou cellule) à atteindre
- Criblage à haut débit (HTS, c'est à dire tests de plusieurs centaines à plusieurs milliers de molécules chimiques sur la cible)
- Découverte de molécules chimiques actives
- Recherche de molécules similaires non toxiques et les plus actives possibles sur la cible
- Tests cliniques
- Validation
- Mise sur le marché

Ce processus complet coûte aujourd'hui environ 800 millions d'euros et peut prendre une quinzaine d'années. C'est pourquoi, la recherche pharmaceutique tente d'optimiser chacune de ces étapes. Pour cela elle fait appel à un nouveau domaine de recherche : la chémoinformatique. En effet la chémoinformatique propose de bons outils pour diminuer le coût et le temps de plusieurs étapes du processus avec notamment les techniques de modélisation moléculaire, de docking, de QSAR... Ces techniques utilisent souvent des méthodes d'apprentissage artificiel.

Le travail que nous allons présenter se situe en amont de l'étape de criblage à haut débit et dans le cas où peu ou pas d'information sur la cible sont connues. En effet, si la cible exacte n'est pas connue, on ne peut extrapoler aucune information sur la structure ou l'activité potentielle que devra avoir la molécule médicament. Dans l'idéal, on pourrait vouloir tester toutes les molécules chimiques possibles pour en trouver une active sur la cible. Cependant il existe plus de 10 millions de molécules disponibles et un nombre quasiment infini de produits chimiques envisageables. Il est donc impossible de les tester tous. Le biologiste souhaitera alors tester un nombre restreint de molécules couvrant autant d'activités biologiques que possible. Dans ce cas, on parle alors de sous-ensemble de molécules divers. Il s'agit donc d'extraire un sous-ensemble divers à partir d'une grande quantité de molécules. De nombreux travaux se sont penchés sur ce problème. Cependant, ils ne sont pas toujours applicables aux grands jeux de données, et la notion de diversité n'est pas définie formellement. Le but de cette thèse est d'approfondir la notion de diversité, de créer des critères que les techniques de sélection par diversité peuvent optimiser et de proposer une nouvelle méthode de sélection par diversité. Ce travail a été réalisée en collaboration avec l'équipe de chémoinformatique de l'ICOA et avec l'équipe d'apprentissage du LIFO. Une des difficultés de la thèse a résidé dans la communication entre ces deux domaines aux techniques et aux vocabulaires souvent très différents. Les travaux présentés par la suite résultent de la synergie des expertises apportées par la chémoinformatique et par l'apprentissage artificiel.

Dans un premier chapitre, nous présenterons les notions liées aux molécules et à la diversité qui seront utilisées dans le reste de la thèse. Nous définirons notamment ce qu'est une molécule, les différentes façon de les collecter et de les classer. Puis nous évoquerons les techniques de descriptions des composés chimiques, ainsi que les différents filtres pouvant leur être appliqués. Nous décrirons quelques techniques permettant de visualiser les produits dans des espaces de faible dimension. Enfin, les notions de similarité/dissimilarité seront présentées pour ensuite définir la diversité et expliquer les différentes méthodes déjà existantes dans ce domaine.

Le deuxième chapitre nous permettra de décrire les données utilisées pour réaliser les travaux de la thèse. Nous présenterons tout d'abord comment les molécules ont été collectées puis traitées. Ensuite nous parlerons des descripteurs choisis pour décrire les données. Enfin la création des différents jeux servant aux tests sera présentée.

Les chapitres trois et quatre constituent le coeur de ce travail et se concentreront sur la sélection par diversité.

Dans le chapitre trois nous donnerons une définition de la diversité que nous formaliserons par un critère. Nous le comparerons à d'autres critères similaires découverts dans la littérature en apprentissage artificiel. Puis chaque méthode comparée par la suite sera décrite. Enfin nous développerons notre propre implémentation du problème de sélection par diversité.

Dans le chapitre quatre nous présenterons notre plan d'expérimentation ainsi que les différents critères d'évaluation utilisés pour comparer et valider nos résultats. Ensuite, les résultats de sélection de chaque méthode seront étudiés séparement puis comparés entre eux. Nous comparerons également les résultats pour différents jeux de données présentant des propriétés identiques ou différentes et pour différentes tailles de sélection.

Enfin nous conclurons en résumant les apports et les perspectives qu'offre cette thèse.

Chapitre 1

Définitions et Etat de l'art

Dans un premier temps nous définissons ce qu'est une molécule, comment l'identifier et enfin nous verrons que les molécules peuvent être organisées en collections de différents types dans le but de faciliter leur manipulation.

Les molécules ainsi définies, nous pourrons expliquer comment les décrire et nous verrons qu'il existe de nombreuses possibilités dans ce domaine. Nous verrons ensuite que ces descriptions conduisent à la définition d'espaces mathématiques dans lesquels on peut projeter et situer les molécules les unes par rapport aux autres.

Ainsi nous exposerons les différents modes de comparaison des molécules (mesure de similarité/dissimilarité). Enfin nous tenterons de définir la diversité d'un ensemble de molécules et nous donnerons les méthodes existantes pour effectuer la sélection par diversité.

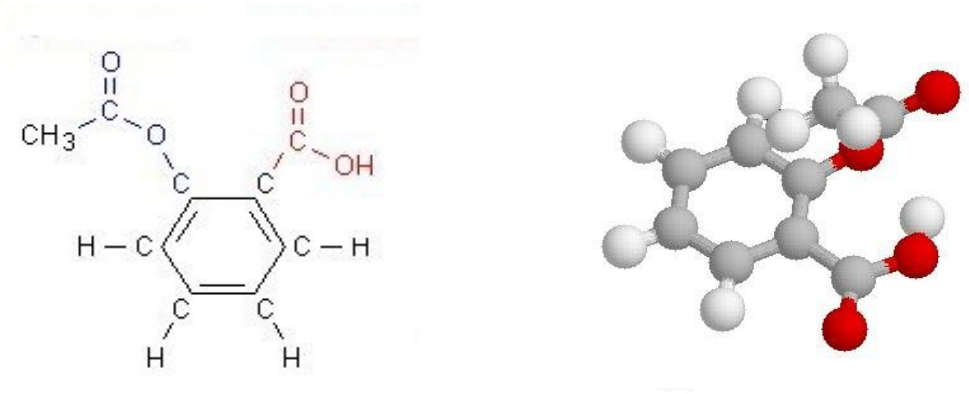
1.1 Les molécules

1.1.1 Définition et identification

Les atomes sont les éléments à partir desquels une molécule est constituée. Les atomes de carbone, d'hydrogène, d'oxygène et d'azote sont les principaux éléments des molécules du vivant. Chacun de ces atomes est identifié par un symbole : C pour le carbone, H pour l'hydrogène, O pour l'oxygène et N pour l'azote. Ceux-ci permettent d'écrire la formule chimique de toute molécule. En effet, la constitution d'une molécule est représentée par la formule brute (1 dimension, soit : C9H8O4 pour l'aspirine ou acide acétylsalicylique) et par sa formule semi-developpée (2 dimensions, cf. Figure 1.1(a)). La molécule peut également être représentée en 3 dimensions (cf. Figure 1.1(b)) en sélectionnant une de ses conformations possibles dans l'espace.

Enfin, à chaque molécule est associé un nom unique suivant la nomenclature IUPAC [1]. Pour la molécule de la figure 1.1, il s'agit de l'acide 2-acétyloxybenzoïque. Cependant ce nom peut devenir très complexe avec la longueur de la molécule et n'est pas adapté aux traitements automatiques. Des codes ont alors été mis en place pour nommer chaque composé chimique. Nous trouvons notamment le code SMILES [2] [1] (cf. Tableau 1.1) et le code InChI [3] [2]. Le code InChI identifie de façon unique chaque composé chimique et permet de manipuler les informations concernant ces produits plus facilement au sein de grands ensembles de molécules.

1. Union Internationale de Chimie Pure et Appliquée
2. Simplified Molecular Input Line Entry System
3. The IUPAC International Chemical Identifier



(a) Formule semi-developpée (b) Une conformation 3D

FIGURE 1.1: Représentations 2D et 3D de l'aspirine

Code SMILES	CC(=O)OC1=CC=CC=C1C(=O)O
Code InChI	1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)

TABLE 1.1: Tableau de quelques codes utilisés pour représenter la molécule d'aspirine

1.1.2 Classement

Chaque jour, plus de 12 000 nouvelles substances chimiques (organiques ou inorganiques) sont ajoutées [3] à la plus célèbre des collections de molécules : le CAS Registry. Il devient donc important, d'une part de les caractériser dans des ensembles reflétant leur disponibilité sur le marché, d'autre part de collecter les informations les concernant et de les classer de manière intelligente.

1.1.2.1 Caractérisation de la disponibilité : Ensembles chimiques

Tout d'abord, tous les composés chimiques peuvent être classés dans des ensembles chimiques. Chacun de ces ensembles caractérise des groupes de molécules en fonction de leur disponibilité ou de leur faisabilité (cf. Figure 1.2). Hann et Oprea [4] distinguent quatre ensembles chimiques :

- Ensemble Virtuel : l'ensemble des molécules possibles, imaginables de taille raisonnable (500 à 1000 Dalton, ensemble estimé à 10^{60} produits) [5]. Il n'y aurait pas assez de matière dans l'univers pour fabriquer une molécule de chacun de ces produits.
- Ensemble Tangible : l'ensemble des produits possiblement synthétisables (avec les connaissances actuelles), le nombre de molécules de cet espace est estimé entre 10^{20} et 10^{24} [6].
- Ensemble Global : l'ensemble des composés qui ont déjà été synthétisés. Il est impossible de connaître leur nombre exact car beaucoup de molécules sont confidentielles [7].
- Ensemble Réel : l'ensemble des produits actuellement existants et réellement disponibles, c'est à dire l'ensemble des collections existantes chez les fournisseurs.

Dans la suite de notre étude, nous nous intéressons à l'ensemble réel, et nous utiliserons donc des produits réellement disponibles, mais l'étude est applicable à tout type d'ensemble de molécules.

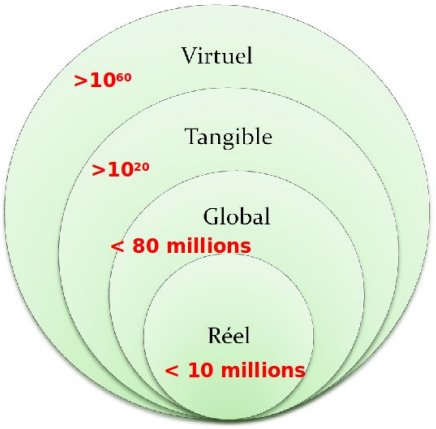


FIGURE 1.2: Les ensembles chimiques et le nombre approximatif de molécules qu'ils contiennent

1.1.2.2 Collections

Nous avons vu que l'espace réel contient plusieurs millions de composés chimiques. Les chimistes et autres utilisateurs de ces molécules ont donc dû créer des collections pour organiser cette masse de données. Ces collections sont aussi appelées chimiothèques. Celles-ci peuvent être virtuelles ou réelles.

Les chimiothèques réelles sont souvent sous forme de plaques de puits (cf. Figure 1.3) contenant chacun un produit différent. Ces plaques sont donc prêtes à être testées.

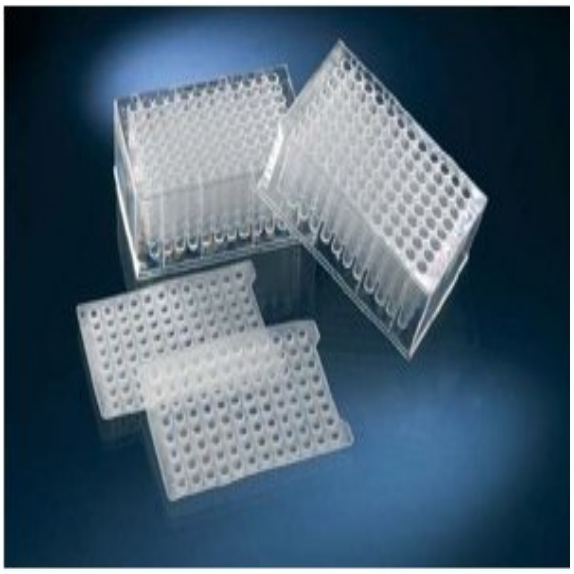


FIGURE 1.3: Exemple de plaques 96 puits (société Nunc)

Les chimiothèques virtuelles sont un ensemble d'informations plus ou moins organisées et hiérarchisées regroupant tout ou partie des données concernant un ensemble de compo-

sés chimiques. Elles peuvent être de plusieurs types et sont classées différemment dans la littérature. Par exemple, la review [8] classe les collections de molécules selon leur contenu en 3 types : bases de littérature (données bibliographiques), bases factuelles (propriétés physico-chimiques, spectres, description des projets de recherche) et bases structurales (informations topologiques, réactions associées aux molécules). Nous adoptons ici le classement proposé par Gasteiger [9] qui divise les chimiothèques en bases bibliographiques et en bases de structures chimiques. Nous définissons chacun de ces types de collection tout en sachant que la frontière entre eux n'est pas absolue.

Les bases bibliographiques Ces bases contiennent les publications originales concernant chaque composé de la base. On y trouve également la littérature liée aux molécules et parfois des données expérimentales. Parmi les plus connues nous comptons : CAS Registry [10] avec plus de 60 millions de substances organiques et inorganiques, Belstein [11] avec 10 millions de produits et 320 millions de données expérimentales. La revue "Basic overview of Chemoinformatics" [8] donne une liste exhaustive des bases bibliographiques disponibles.

Les bases de structures chimiques Ces librairies peuvent provenir de plusieurs sources (distributeurs commerciaux, entrepôts académiques). Dans ce type de librairies, les composés apparaissent dans de multiples représentations (2D, SMILES, fingerprints : cf. paragraphe 1.2.2.1) et la plupart contiennent un large panel de valeurs caractérisant les molécules (appelées aussi descripteurs : section 1.2). On peut donc aussi appeler ces bases, des librairies annotées. Parmi les bases de structures chimiques on trouve notamment : iResearch de ChemNavigator [12] avec 15 millions de structures uniques provenant de 179 fournisseurs, Zinc [13], la chimiothèque nationale [14], ChemDB [15], ChemIDplus [16]. Chaque fournisseur commercial propose également sa propre librairie de produits chimiques (nous en donnons une liste des principaux dans le chapitre 2). Il existe enfin des bases formant un catalogue de ces produits commerciaux accompagnés d'une information sur les fournisseurs les proposant : e-molecules [17], ACD [18], CHEMCATS [19] ou encore ChemSource [20].

Certaines librairies sont annotées avec des informations sur des cibles biologiques potentielles. c'est le cas notamment de PubChem [21], de la base du NCI [22], ChemBank [23], ChemMine [24], ChEBI [25]. Certaines d'entre elles réunissent uniquement des molécules médicaments comme la Drug Bank [26] qui recense les cibles biologiques liées à chacun des composés de la librairie, ou encore Prestwick [27].

Parmi les librairies annotées, on compte également les bases contenant des informations thermodynamiques. Enfin certaines librairies sont particulières car elles ne contiennent pas que des composés existants. En effet, les librairies combinatoriales appartiennent plutôt à "l'ensemble global" défini précédemment. Les librairies combinatoires réelles proviennent de la chimie combinatoire et elles sont considérées comme des chimiothèques réelles (comme définies plus haut). En revanche les librairies combinatoires virtuelles sont basées sur un ensemble de fragments moléculaires et de réactions que l'on sait réalisables. Les composés contenus dans ces librairies résultent d'un grand nombre de combinaisons possibles et synthétisables entre ces fragments. Pour plus d'informations sur la construction de ces librairies on peut se référer aux travaux suivants [28, 29, 30]. A titre d'exemple, récemment Fink et Reymond [31] ont construit une très grande librairie combinatoire de 26,4 millions de molécules respectant les règles de biodisponibilité de Lipinski (voir paragraphe 1.3).

Enfin, les librairies ciblées contiennent des composés aux caractéristiques spécifiques en relation avec une ou quelques cibles biologiques particulières. Par exemple certains tra-

vaux [32] décrivent une technique pour construire une librairie ciblée enrichie en molécules bioactives.

Les sous-librairies Les librairies sont souvent beaucoup trop grandes pour effectuer par exemple une recherche de médicament ciblée ou des tests HTS in vitro [4]. Il est en effet impossible de pratiquer des tests biologiques à l'échelle de plusieurs millions de composés, pour des raisons de temps et de coût. Il est donc indispensable de réduire la taille de ces librairies pour obtenir des sous-librairies qui font partie de la classe des bases de structures chimiques. Celles-ci sont sélectionnées selon différents critères, par exemple :

- sélection par diversité pour effectuer un HTS (criblage à haut débit) réel ou virtuel sans connaissance de la cible
- sélection par similarité à un composé donné potentiellement actif, pour trouver un composé le plus actif possible et/ou non toxique

1.2 Les descripteurs

1.2.1 Définition

Toute molécule est décrite par des valeurs que l'on appelle descripteurs. Ces valeurs peuvent être obtenues expérimentalement mais le plus souvent, elles sont calculées à partir de la structure de la molécule. Par exemple, le poids moléculaire est calculé à partir de celui des atomes constituant le composé chimique.

De nombreux travaux [33, 34] étudient le calcul ou la prédiction de certains descripteurs. Plusieurs logiciels sont disponibles pour effectuer ces calculs, comme : QSARIS [35], Cerius2 [36], VolSurf [37], Dragon [38]. Enfin des librairies informatiques telles que le CDK [39] ont été développées. Elles permettent à des utilisateurs avertis d'intégrer le calcul de leurs descripteurs à une plateforme informatique déjà existante.

Il est possible de classer les milliers de descripteurs existants [40] de différentes façons. Nous verrons ces classements, puis nous présenterons quelques méthodes permettant de manipuler un nombre réduit de descripteurs.

1.2.2 Classement des descripteurs

En effet on peut organiser les descripteurs ainsi :

- selon leur type : quantitatif, qualitatif, structuré
- selon la dimensionnalité de la représentation moléculaire à partir de laquelle sont déduits les descripteurs
- selon la nature des propriétés décrites

1.2.2.1 Type numérique/structuré du descripteur

Nous décrivons ici quatre types principaux de formes numériques ou géométriques que peuvent prendre les descripteurs.

4. In vitro signifie "dans le verre" par opposition à in vivo (dans l'organisme, dans la cellule). L'expression in vitro désigne une recherche ou un test effectué en tube à essai ou plus généralement en dehors de l'organisme vivant, dans des conditions artificielles.

Les variables simples Il s'agit de la majorité des descripteurs. Ils peuvent prendre une valeur numérique dite quantitative ou une valeur qualitative.

Les variables qualitatives ou symboliques comme par exemple la couleur des yeux, sont des variables ne prenant que des valeurs non numériques appelées caractéristiques ou modalités. Elles peuvent être :

- ordinales, dans ce cas les modalités (ou valeurs) que prennent les variables peuvent être classées par ordre croissant ou décroissant (par exemple la variable taille peut prendre les modalités petit, moyen, grand). Les variables ordinales peuvent être numériques (par exemple : 1 = petit, 2 = moyen, 3 = grand) mais les opérations mathématiques (telles que la somme ou la moyenne) n'ont pas de sens sur ces valeurs.
- nominales, dans ce cas les modalités des variables ne sont pas ordonnables (exemple : la couleur peut prendre les modalités rouge, jaune, bleu).

Lorsque ces variables ne prennent que deux modalités, elles sont dites dichotomiques (exemple : jour/nuit). En informatique une variable dichotomique est dite variable binaire et elle est exprimée par les valeurs 0 et 1.

Enfin lorsque les données sont décrites avec des variables numériques, ces dernières sont dites quantitatives ou mesurables. Une variable est quantitative continue lorsqu'elle peut prendre l'une des valeurs d'un intervalle comme par exemple le poids moléculaire. Une variable est quantitative discontinue ou discrète lorsqu'elle ne prend que des valeurs isolées les unes des autres, par exemple le nombre d'atomes d'une molécule), le nombre de valeurs qu'elle peut prendre est alors dénombrable. Les valeurs des variables discrètes sont généralement des valeurs entières.

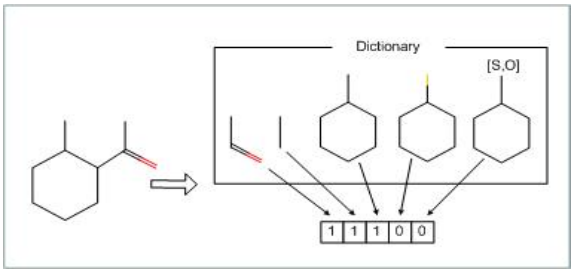


FIGURE 1.4: Exemple de fingerprint

Les fingerprints (cf. Figure 1.4) Les fingerprints sont des vecteurs de bits (ou valeurs binaires). Chaque bit prend la valeur 0 ou 1. Cette valeur code l'absence ou la présence de certains fragments structuraux [5] ou d'autres caractéristiques. Parfois, un ou plusieurs bits peuvent coder une propriété physico-chimique. Un exemple de la construction de fingerprint est donné dans [41]. Les plus couramment utilisés sont les fingerprints MACCS et Daylight [42], il existe aussi MolPrint2D [43], le fingerprint BCI [44], le fingerprint standard 2D de tripos [45] ou encore le fingerprint de Scitegic [46]. Les fingerprints restent l'un des types de descripteurs les plus utilisés en chemo-informatique pour décrire les molécules et les comparer entre elles. Ils sont d'ailleurs souvent utilisés comme seul type

5. On appelle fragment structural, un sous-ensemble d'une molécule ; c'est à dire une partie de l'enchaînement des atomes. Par exemple un cycle composé de 5 carbones est un fragment structural pouvant se trouver au sein d'une molécule

de descripteur. C'est pourquoi leur efficacité à caractériser correctement les molécules a été étudiée à plusieurs reprises [47, 48].

Vecteurs et matrices Les descripteurs peuvent être de forme plus complexe comme des vecteurs ou des matrices. Il peut s'agir de matrices de connectivité indiquant pour chaque couple d'atomes la présence ou non d'une liaison.

Graphes (cf. Figure 1.5) Ce type de descripteurs permet de représenter chaque molécule par un ensemble de nœuds et d'arêtes. Chaque nœud peut représenter une sous-structure de la molécule, une fonction ou un atome. Ces nœuds sont connectés entre eux par des arêtes symbolisant les liaisons existantes de la molécule. Cette représentation permet d'utiliser des algorithmes de comparaison de graphes pour calculer la similarité entre molécules chimiques.

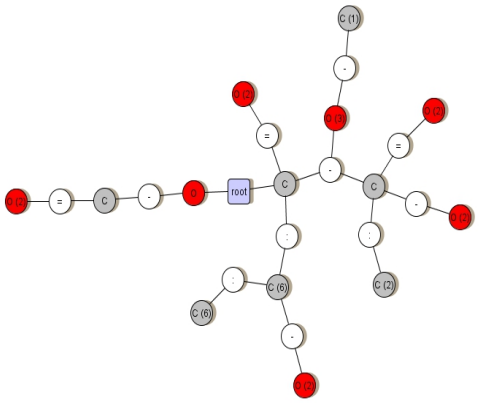


FIGURE 1.5: Exemple de graphe tiré de [49]

1.2.2.2 Classement selon la dimensionnalité

Nous avons vu que la molécule peut être représentée en une, deux ou trois dimensions. Les descripteurs peuvent être calculés à partir de chacune de ces dimensions. On obtient donc trois classes différentes.

Descripteurs obtenus à partir de la formule brute (1D) Ce sont les descripteurs basés uniquement sur les atomes et leur nombre. Ce sont donc souvent des compteurs d'atomes simples (nombre de carbones, nombre d'hydrogènes...). On peut également trouver des binaires codant l'absence ou la présence de certains éléments spécifiques. Enfin, parmi ces descripteurs on trouve des valeurs telles que le poids moléculaire qui peuvent être calculées à partir de cette formule brute.

Descripteurs obtenus à partir de la formule développée (2D) Ce sont des descripteurs basés sur la topologie de la molécule, les groupes d'atomes et les fonctions chimiques (comme le coefficient de partage octanol/eau ou LogP, la présence de certains fragments). On trouve également les descripteurs liés à la connectivité tels que les matrices de connectivité vues précédemment.

Descripteurs obtenus à partir de la conformation dans l'espace (3D) Il s'agit ici de descripteurs basés sur la stéréochimie et la géométrie de la molécule. On y trouve notamment les descripteurs de surface (par exemple la surface de distribution du potentiel électrostatique), les conformations possibles que peut prendre une molécule dans l'espace, les énergies (HOMO, LUMO). On peut également y classer les descripteurs appelés pharmacophores. Ceux-ci sont représentés sous forme de graphes (connexes ou non). Ils indiquent des fonctions chimiques ainsi que leur enchaînement qui pourraient être caractéristiques d'une activité biologique.

1.2.2.3 Classement selon le type de propriétés décrites

Enfin il est possible de classer les descripteurs selon les propriétés qu'ils décrivent d'une molécule.

Propriétés physico-chimiques macroscopiques Il s'agit par exemple du poids moléculaire et du LogP.

Propriétés structurales Ces propriétés concernent les constituants de la molécule, il s'agit des compteurs d'atomes et de liaisons.

Propriétés topologiques Cet ensemble rassemble les descripteurs liés à la connectivité de la molécule (les matrices de connectivité par exemple et divers indices dérivés de ces matrices).

Ainsi ces classements permettent à l'utilisateur de choisir la ou les classes les plus adaptées à son projet et/ou aux informations dont il dispose sur la molécule. Par exemple si on ne dispose pas des informations 3D sur le composé, certains descripteurs ne pourront pas être calculés. Ou encore, une recherche basée sur les sous-structures ne requiert pas de descripteurs sur les propriétés physico-chimiques. En revanche, lors de la recherche de médicaments, un filtrage (cf. section 1.3) doit être utilisé pour, entre autres, éliminer des molécules potentiellement toxiques. Dans ce cas, les propriétés physico-chimiques sont indispensables.

1.2.3 Réduction du nombre de descripteurs

Nous avons vu qu'il existe plusieurs milliers de descripteurs. Or il n'est pas possible de tous les utiliser pour comparer des molécules. Cela ne serait d'ailleurs pas pertinent puisque plusieurs d'entre eux sont redondants et d'autres ne sont pas forcément pertinents selon l'objectif que l'on souhaite atteindre.

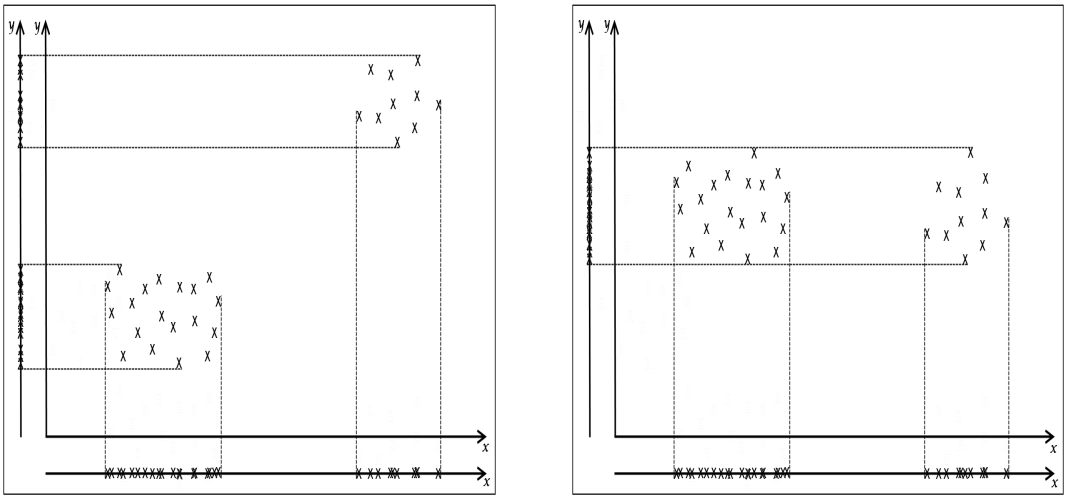
1.2.3.1 Redondance

On appelle descripteurs redondants, des variables qui apportent la même information. En statistique il s'agit de deux variables très corrélées entre elles. Pour Whitley [50] par exemple, la redondance est "une dépendance linéaire exacte entre plusieurs colonnes (descripteurs) d'une matrice de données. Ce qui signifie qu'au moins une de ces colonnes contribue à une information non unique". Nous pouvons illustrer la redondance par le schéma 1.6(a). Sur celui-ci un ensemble de points est décrit par deux variables x et y, et est donc projeté selon deux axes. On remarque que cet ensemble comporte deux groupes distincts. Or si on supprime la variable y, l'ensemble de points est alors projeté sur l'axe x. Dans

ce cas, la variable x permet toujours de discriminer deux groupes au sein de l'ensemble. De même si on supprime la variable x, les points sont projetés sur l'axe y ; on obtient le même type de résultats. Ces deux variables sont donc considérées comme redondantes puisqu'elles apportent toutes deux la même information sur l'ensemble de points. Si ces deux variables sont conservées pour la suite de l'analyse, cela risque de donner un poids artificiellement plus élevé à l'information qu'elles contiennent.

1.2.3.2 Pertinence

On définit la pertinence d'une variable, par sa capacité à délivrer une information "intéressante" sur les observations qu'elle décrit au regard de l'objectif de l'analyse. Cette pertinence dépend donc du projet dans lequel on se place. Dans le cadre de l'apprentissage supervisé et toujours selon Whitley [50], les descripteurs pertinents ont "une corrélation statistique significative avec la variable réponse (d'une analyse) et ne doivent pas avoir une variance faible". Dans le cadre de l'apprentissage non supervisé, nous pouvons illustrer le concept de pertinence avec le schéma 1.6(b). Sur celui-ci un ensemble de points est décrit par deux variables x et y, et est donc projeté selon deux axes. On peut voir que cet ensemble comporte deux groupes distincts. Lorsqu'on retire la variable y, on obtient la projection de cet ensemble sur l'axe x. Dans ce cas on observe toujours deux groupes. En revanche lorsqu'on supprime la variable x et que l'on effectue alors la projection des points sur l'axe y, on remarque qu'il n'existe plus qu'un seul groupe de points. Dans le cadre de ce schéma où l'on souhaite discriminer deux groupes au sein d'un ensemble, la variable y n'est donc pas pertinente.



(a) Illustration de la redondance

(b) Illustration de la pertinence

FIGURE 1.6: Illustration de la redondance et de la pertinence de deux variables x et y

Il est donc important de réduire le nombre de descripteurs pour éliminer la redondance et ne conserver que les descripteurs pertinents. Mais cette réduction permet également à l'utilisateur d'observer moins de variables et donc de mieux comprendre les phénomènes expliquant les données et les résultats de son analyse. Il existe plusieurs techniques de réduction de dimensions : la compression, la création de nouveaux descripteurs rassemblant plusieurs informations, la sélection.

1.2.3.3 La compression

Nous avons vu que les fingerprints codent pour la présence ou l'absence de certains fragments structuraux dans une molécule. Or, étant donné le nombre considérable d'informations que l'on peut résumer dans un fingerprint, sa taille peut vite devenir trop grande pour réaliser les calculs de comparaison de molécules en un temps raisonnable. Ils sont donc souvent compressés. L'efficacité de cette compression est controversée et a été étudiée par Baldi et al. [51] qui présente plusieurs algorithmes de compression.

Pour mieux comprendre ce qu'est la compression d'un fingerprint nous pouvons expliquer une méthode simple citée dans cet article. Prenons un fingerprint de 10 bits : (1,0,0,1,0,0,0,0,1,0). On définit l'index de représentation par un vecteur d'entiers indexant la position dans le fingerprint des bits valant 1. Pour l'exemple, l'index de représentation est donc (0,3,8). Ensuite la représentation dite "run-length" est un vecteur indexant la longueur des séries de bits valant 0 suivies d'un bit valant 1. Ici, ce vecteur serait le suivant : (0,2,4). Ces deux représentations du fingerprint produisent un vecteur d'entiers qui est une forme de compression.

1.2.3.4 La création de nouveaux descripteurs (permettant la réduction de dimensions)

Tout d'abord, plusieurs travaux proposent des indices qui résument des ensembles de descripteurs. Chacun de ces indices peut être considéré comme un descripteur à part entière. On peut trouver plusieurs exemples de ces indices notamment dans les travaux de Balban et al. [52] et de Gregori-Puigjané et al. [53] mais également dans deux livres [54, 55]. Ensuite, il est également possible d'utiliser des méthodes telles que l'ACP (Analyse en Composante Principale). Celle-ci a pour principe de réduire un ensemble de descripteurs en un nombre réduit de composantes correspondant chacune à une combinaison linéaire de l'ensemble des descripteurs. Elle élimine dans le même temps la redondance.

L'inconvénient de telles combinaisons est que, lors d'analyses et de comparaisons de molécules, elles ne permettent pas une compréhension aisée des propriétés régissant les phénomènes existants entre molécules. Une alternative à la compression et à la combinaison des données, est la sélection de descripteurs.

1.2.3.5 Sélection de descripteurs

Il s'agit de choisir parmi les milliers existants, les descripteurs qui décrivent le mieux une molécule en termes de structure, de propriétés. Il est donc question d'identifier la pertinence de certains descripteurs au regard d'un but précis (dans notre cas : favoriser une bonne sélection par diversité). Mais également, on souhaite éliminer la redondance (définie précédemment), c'est à dire ne pas conserver deux descripteurs codant le même type d'information.

Cadre de la sélection Dans notre cas, nous souhaitons effectuer une sélection sur un jeu de descripteurs associés à des molécules dont on n'a aucune connaissance a priori. Par exemple, dans le cas de notre étude nous travaillons avec des molécules dont ne connaît pas l'activité biologique éventuelle. Plus généralement, on ne possède aucune information sur une quelconque organisation des molécules les unes par rapport aux autres. En apprentissage artificiel (ou Machine Learning), on se situe donc dans le cadre de l'apprentissage non supervisé. Par opposition l'apprentissage supervisé est le cas où l'on possède des connaissances sur le jeu de données (une certaine classification par exemple).

Or en apprentissage, de nombreux travaux traitent de la sélection de descripteurs dans le cadre supervisé. En revanche pour l'apprentissage non supervisé, peu d'études et de méthodes existent. Nous donnerons ici les différents types possibles de sélection de descripteurs pour l'apprentissage non supervisé.

Les différentes approches de sélection de variables Il existe 2 approches différentes pour la sélection de variables en apprentissage non supervisé :

- les approches "filtre"
- les approches "enveloppe"

Avec les approches "filtre", la sélection de descripteurs s'effectue avant toute opération d'apprentissage. Il s'agit la plupart du temps de méthodes dites de ranking ou classement. Les descripteurs sont classés selon un critère comme l'entropie, la variance, la corrélation etc... Ces approches sont indépendantes de l'application d'apprentissage utilisée par la suite pour traiter les données. Par exemple, Dash et Liu [56] proposent une méthode de ranking basée sur l'entropie du jeu de données. D'autres méthodes positionnées dans différents contextes d'apprentissage sont notamment citées dans la review de Guyon et Elisseeff [57] et concernent principalement l'apprentissage supervisé.

Les approches "enveloppe" sont constituées de trois étapes : la sélection d'un sous-ensemble de descripteurs, l'apprentissage à partir de ce sous-ensemble et l'évaluation du résultat de l'apprentissage pour savoir si le sous-ensemble est optimal ou non. Si le résultat de l'apprentissage n'est pas satisfaisant, un autre sous-ensemble est testé jusqu'à obtention du résultat souhaité. Pour cela, il convient d'effectuer trois choix :

- le choix de la stratégie de recherche des sous-ensembles possibles dans l'espace des variables
- le choix du prédicteur ou classifieur utilisé en fonction de l'opération d'apprentissage désirée
- le choix de la méthode d'évaluation de la performance

Tout d'abord, une recherche exhaustive de tous les sous-ensembles possibles de descripteurs (au nombre de 2^d où d est le nombre de descripteurs total) n'est pas envisageable [58]. C'est pourquoi plusieurs stratégies de recherche de ces sous-ensembles ont été développées comme nous pouvons le voir dans la review de Kohavi et John [59]. Parmi ces stratégies, les plus connues sont les stratégies gloutonnes "forward selection" et "backward selection". La stratégie "forward" consiste à commencer avec un ensemble vide auquel on ajoute séquentiellement le descripteur qui, en combinaison avec les descripteurs déjà sélectionnés, optimise un critère choisi. Ce critère peut être le coefficient de corrélation par exemple. Dans ce cas, on cherchera à ajouter itérativement le descripteur le moins corrélé à ceux déjà sélectionnés. Un exemple d'une telle stratégie est développé par Withley et al. en 2000 [50] pour le domaine de la chemoinformatique.
La stratégie "backward" commence avec la totalité des descripteurs et les élimine un par un itérativement selon le même type de critère à optimiser. Pour ces deux stratégies, l'ajout ou la suppression s'arrête lorsque le critère n'est plus optimisé par les descripteurs restants. Ces stratégies gloutonnes ne sont pas les meilleures et ne garantissent pas une solution optimale mais il est admis qu'elles sont rapides et produisent une solution raisonnable.

Ensuite, il s'agit de choisir l'algorithme d'apprentissage utilisé pour traiter les données avec le sous-ensemble de descripteurs proposé par la stratégie de recherche. Dans le cadre de l'apprentissage supervisé, il s'agira de choisir un prédicteur tel que les arbres de décision ou les Support Vector Machine (SVM). Dans le cadre de l'apprentissage non supervisé, il

s'agit souvent de clustering.

Enfin, pour évaluer la performance de l'apprentissage avec un sous-ensemble de descripteurs donné, il existe plusieurs possibilités. Celles-ci permettent de guider la recherche et de la stopper. Dans le cadre de l'apprentissage supervisé, étant donné que les données sont étiquetées, il suffit de mesurer la qualité de la prédiction. Si la qualité n'est pas acceptable, les trois étapes de l'approche enveloppe sont répétées jusqu'à obtention d'une bonne prédiction.
En revanche dans le cadre de l'apprentissage non supervisé, on ne connaît pas a priori les classes des objets. Il faut donc déterminer un critère à optimiser. En général il s'agit de critères liés à la qualité du clustering tels que la séparation des clusters.

On trouve un exemple de sélection de descripteurs via l'approche enveloppe dans les travaux de Dy et Brodley [60]. Tout comme cette thèse, leur travaux se situent dans le cadre non supervisé.

1.3 Druggabilité - Les filtres

Les descripteurs que nous venons de définir permettent d'évaluer les propriétés d'une molécule. Or cette étape est cruciale dans le développement d'un médicament. En effet, pour devenir un médicament, un composé chimique doit certes posséder une bonne activité biologique, mais il doit également avoir des propriétés qui lui confèrent une bonne biodisponibilité. On entend par ce terme : des propriétés d'Absorption, de Distribution, de Métabolisme, d'Elimination (propriétés ADME) acceptables pour le corps humain. A ces propriétés, s'ajoute souvent l'évaluation de la Toxicité (ADMET). Un composé chimique qui possède de bonnes propriétés ADMET est un bon candidat pour être un médicament, il est alors appelé "drug-like".

Par le passé, les tests de biodisponibilité apparaissaient très tard dans le processus de développement d'un médicament. Cela avait pour conséquence d'éliminer des produits dans des phases finales, faisant ainsi perdre le bénéfice de plusieurs années de recherche. Afin de réduire les coûts de développement il est devenu indispensable d'éliminer les mauvais candidats le plus tôt possible. Aujourd'hui, cette élimination peut intervenir avant même la création de sous-librairies de produits à tester, par l'application de filtres que nous allons décrire.

1.3.1 Définitions

1.3.1.1 Composés drug-like

Pour créer des filtres, il a fallu donner une définition précise de ce que l'on attend d'un composé drug-like. Voici quelques définitions qui ont conduit à la mise en place de nouveaux filtres :

- Lipinski [61, 62] décrit les molécules drug-like comme "des composés qui possèdent des propriétés ADMET suffisamment acceptables pour une bonne absorption orale".
- Walters et Murcko [63] en 2002 définissent un composé drug-like comme "une molécule qui contient des groupes fonctionnels et/ou qui possède des propriétés physico-chimiques cohérentes avec la majorité des médicaments connus". Ils passent en revue plusieurs techniques informatiques pour identifier les molécules drug-like.

- Muegge [64] propose que "le critère drug-like soit un descripteur général qui définit le potentiel d'une petite molécule à devenir un médicament".
- Vieth et al. [65] disent que "les propriétés drug-like sont vues comme celles qui confèrent à la molécule des propriétés pharmacocinétiques et pharmacodynamiques désirables, indépendamment de la cible biologique".

Mais comme le souligne Lipinski, la notion drug-like n'est pas seulement liée à des propriétés ADMET. Elle est également liée au mode d'administration que l'on souhaite pour le médicament [66] ; il faut prendre en compte les cibles (exemple : système nerveux central ou CNS) ou les barrières à franchir pour atteindre la cible (exemple : Blood-Brain Barrier ou BBB). C'est pourquoi les filtres qui vont être décrits sont à utiliser avec précautions et à adapter en fonction du projet que l'on mène.

1.3.1.2 Composés lead-like

Il existe également des composés dits "lead-like". Le concept de molécule lead-like est basé sur le même principe que le concept de molécule drug-like, mais il est plus restrictif. Lors d'un criblage à haut débit, lorsqu'un ou plusieurs hits (molécules réagissant avec la cible) sont découverts, on cherche alors un lead, c'est à dire une molécule dont la structure est optimisée à partir de celle(s) des hits. Ce composé lead servira ensuite à trouver un composé dont l'activité sur la cible biologique est optimale. Cette étape se situant après la découverte d'activité, nous nous penchons uniquement sur le concept drug-like qui peut être appliqué avant les criblages à haut débit.

1.3.2 Filtres sur les propriétés

Il existe plusieurs filtres pour obtenir des composés drug-like et lead-like. On trouve un résumé des différents filtres lead-like dans les travaux de Charifson et al. [67]. Nous présentons ci-dessous les filtres drug-like.

La "règle des 5" de Lipinski (pour administration par voie orale) [66] est la plus souvent utilisée. Cette règle contient 4 paramètres concernant les propriétés qu'un médicament peut avoir :

- Poids moléculaire ≤ 500 Da
- LogP ≤ 5
- Nombre d'atomes donneurs de liaisons hydrogènes ≤ 5
- Nombre d'atomes accepteurs de liaisons hydrogènes ≤ 10

Lipinski indique [62, 61] que si au moins deux de ces conditions ne sont pas vérifiées, la molécule pourra avoir une faible absorption ou une faible perméabilité.

En 2002, Veber [68] a complété cette règle avec deux nouveaux paramètres :

- surface polaire de la molécule ≤ 140 Å
- Nombre de liaisons tournantes ≤ 10

Ces règles ne doivent pas être considérées comme discriminatrices de bons/mauvais candidats médicaments [69]. Elles permettent de délimiter un périmètre de propriétés. Certaines molécules sortant de ce périmètre pourraient être des médicaments mais avec peu de chance pour une administration par voie orale.

D'autres auteurs se sont penchés sur l'importance de certaines propriétés moléculaires pour obtenir des filtres drug-like :

- Bergstrom et al. [70] en 2003 montrent que "les propriétés de surface moléculaire d'un composé peuvent prédire la solubilité et la perméabilité d'un médicament avec une confiance suffisante pour conduire à une classification des molécules médicaments".
- Vieth et al. [65] en 2004 ont étudié la différence entre des médicaments oraux et non-oraux par des analyses statistiques. Ils ont ainsi montré que plusieurs facteurs influencent la biodisponibilité. Par exemple, "les médicaments administrés par voie orale possèdent moins de donneurs et d'accepteurs de liaisons H et moins de liaisons tournantes que les médicaments administrés par d'autres voies".
- Vistoli et al. [71] en 2008 ont présenté la flexibilité moléculaire comme un autre descripteur important pour évaluer le critère drug-like d'un composé.

L'utilisation de ces filtres classiques est maintenant très répandue dans la plupart des logiciels commerciaux comme : Absolv [72], Cerius 2 [36], Idea [73], OraSpotter [74], QikProp [75], QMPRPlus [76], Volsurf [37].

Enfin les règles décrites dans ces filtres sont résumées plus complètement dans les propriétés ADMET (Absorption, Distribution, Métabolisme, Elimination, Toxicité). Gola et al. [34] passe en revue les méthodes de prédiction de ces propriétés. Récemment, Gleeson [77] a généré un jeu de règles ADMET simples et interprétables. De nos jours aucun projet de découverte de médicaments ne peut se faire sans la prédiction des propriétés ADMET et une sélection associée des molécules.

1.3.3 Scores

Plusieurs travaux présentent des approches statistiques pour définir la biodisponibilité des molécules.

Martin [78] en 2005 présente un score de biodisponibilité. Pour ce faire, les différents descripteurs utilisés pour définir la règle des 5, ainsi que le PSA (Polar Surface Area) et d'autres décrivant la perméabilité sont calculés. Grâce à des molécules dont la biodisponibilité et la perméabilité ont déjà été testés in vivo, elle a identifié des règles propres à chaque type de molécules (anions, neutres...). De ces règles un score ABS (A Bioavaibility Score) a été construit. Celui-ci correspond à la probabilité qu'un composé ait une perméabilité ou une biodisponibilité supérieure à 10%. Ce score a été validé par sa capacité à identifier les composés connus comme bien ou peu absorbés par l'homme.

Monge et al. [79] ont implémenté un score drug-like progressif. Pour chaque propriété définie dans les filtres classiques (8 critères au total), une pénalité variant de 0 à 1 est calculée. Elle est calculée à partir d'une fonction empirique basée sur les valeurs seuils des critères drug-like. Par exemple pour le critère "nombre d'accepteurs de liaisons hydrogène" (HBA), la limite supérieure pour une molécule drug-like est 10. Si une molécule possède moins de 7 HBA (soit 10 - 30%), la pénalité est de 0. Si elle possède plus de 13 HBA (soit 10 + 30%), la pénalité maximum de 1 est appliquée. Des valeurs intermédiaires sont appliquées pour les valeurs comprises entre 7 et 13. Toutes les fonctions résultent de la distribution des propriétés pour des médicaments connus ou des limites proposées par d'autres auteurs. Ce score permet non seulement de discriminer les composés drug-like, des non drug-like, mais il permet également de classer les composés selon leur druggabilité croissante. De plus, contrairement aux méthodes classiques de sommes de critères satisfaits, cette méthode élimine les effets de bords. Ceci est le cas par exemple pour le logP qui peut avoir des valeurs variables pour un même composé selon la méthode de calcul utilisée. Un composé pouvait donc être classé comme drug-like ou non drug-like selon la méthode utilisée. Ce score évite donc cet écueil.

Hutter [80] en 2007 a étudié la distribution des types d'atomes et leurs combinaisons par paires dans les molécules médicaments connues et non-médicaments. Il fait une étude statistique des probabilités d'apparition de ces combinaisons pour dériver un score drug-like sur une échelle logarithmique. Les médicaments ont un score supérieur à 0.3 et les autres produits doivent avoir un score proche de 0. L'efficacité de ce score est confirmée par un taux de prédiction des médicaments existants de 71%.

1.3.4 Filtres structuraux

Il existe d'autres types de filtres : les fonctions réactives et les warheads décrits par Rishton [81]. Ces filtres ne sont pas exactement destinés à sélectionner des composés drug-like, mais ils sont utilisés pour la sélection de composés destinés au criblage à haut débit.

Les fonctions réactives sont des fonctions se situant sur les composés chimiques qui induisent des faux-positifs lors des criblages à haut débit. En effet ces composés semblent réagir avec les cibles biologiques, mais elles présentent en réalité une réactivité chimique avec ces cibles plutôt qu'une activité biologique. Elle se lient de manière covalente aux cibles biologiques. Il est donc important d'éliminer les composés présentant de telles fonctions avant tout criblage.

Les warheads sont des molécules qui entraînent également des faux-positifs dans les criblages mais en formant des liaisons non covalentes.Ces composés se distingue lors des criblages par une forte relation structure-réactivité[81]. De tels composés doivent donc être supprimés avant un criblage pour ne pas fausser les résultats.

D'autres composés à éliminer, les promiscuous aggregating inhibitors, possèdent des fonctionnalités non-compétitives, une faible relation structure-activité et une mauvaise sélectivité. Leur mécanisme d'action a été décrit par McGoven et al. [82] et leur identification in silico [6] a été présentée par Seidler [83].

1.4 Espace chimique/Espace de visualisation

Les molécules chimiques appartiennent à ce que l'on appelle l'espace chimique. Cet espace est défini par Lipinski et Hopkins [84] comme suit : "L'espace chimique peut être vu comme étant analogue à l'univers cosmique dans son immensité, avec les composés chimiques peuplant l'espace en lieu et place des étoiles".

Or parmi les étapes du processus de développement d'un médicament, il en existe plusieurs où l'on a besoin de comparer les molécules de cet espace entre elles (effectuer une sélection par diversité, trouver des molécules similaires à un lead...). Il est donc utile de les représenter dans un espace mathématique compréhensible par l'homme et traduisant au mieux cet espace chimique absolu. En effet on peut représenter les composés chimiques dans un espace multi-dimensionnel qui permet de matérialiser les similarités, et la diversité dans un ensemble de produits. L'intérêt d'une telle représentation est bien expliqué par Raghavendra et Maggiora [85] : "Les espaces chimiques produisent une base intuitive et conceptuelle pour la compréhension des nombreuses relations existantes parmi les collections diverses de composés".

En apprentissage, on appelle cet espace multi-dimensionnel : l'espace des descriptions, défini par les descripteurs décrivant les objets. En chémoinformatique, cet espace est carac-

6. In silico signifie "dans l'ordinateur" par opposition à in vitro (en tube à essai) ou à in vivo (au sein d'un organisme). Le terme in silico désigne une recherche ou un test effectué à l'aide de l'outil informatique

térisé par les descripteurs moléculaires. Dobson [86] précise d'ailleurs que le terme "espace chimique" est souvent employé à la place "d'espace des descripteurs multi-dimensionnel". De tels espaces sont en réalité des régions de l'espace général défini par Lipinski précédemment. Ces régions étant définies par le choix des descripteurs les caractérisant et de leurs limites, elles ne produisent pas les mêmes espaces et ne peuvent être considérées comme espace absolu.

Plusieurs techniques existent pour construire de tels espaces multi-dimensionnels. Maggiora et al. [87] proposent de classer les techniques d'extraction de molécules en 2 types, en fonction de l'espace dans lequel se placent les molécules concernées. Nous proposons donc de classer les méthodes de construction d'espaces chimiques et d'espaces de visualisation selon ces deux mêmes types :

- "Single molecule coding (espace basé sur les coordonnées). Chaque molécule a des coordonnées précises dans un espace multi-dimensionnel. Ces coordonnées sont déduites de plusieurs critères connus des molécules. Chaque composé possède une position absolue dans cet espace.
- "Pairwise molecule coding (espace libre de coordonnées). Les molécules sont ici caractérisées par leurs distances aux autres. Leur position dans l'espace n'est donc plus absolue puisqu'elle variera avec le jeu de molécules qui l'entourent.

1.4.1 Espaces "single molecule coding"

La façon la plus courante de représenter les molécules dans l'espace est d'utiliser les valeurs des descripteurs comme coordonnées des molécules. Chaque descripteur correspond alors à une dimension de l'espace. Cette technique est souvent utilisée non seulement pour caractériser la position des molécules dans l'espace mais aussi et surtout pour permettre aux utilisateurs de visualiser les molécules et ainsi comprendre les relations qui les lient. Or étant donné le grand nombre de descripteurs, la dimensionnalité de l'espace devient très vite inconcevable par un cerveau humain. C'est pourquoi il est nécessaire d'utiliser des méthodes de réduction de la dimensionnalité. Maniyar et al. [88] comparent les principaux algorithmes et méthodes statistiques de visualisation :

- l'analyse en composantes principales [89]
- le mapping de Sammon [90] et les self-organizing maps (SOM) [91]
- le Generative Topographic Mapping (GTM) [92] et son pendant hiérarchique (HGTM)

Les plus couramment utilisées en chémoinformatique sont l'ACP et les cartes de Kohonen (ou SOM). Nous décrivons donc rapidement le principe de ces deux méthodes.

L'ACP est une méthode statistique qui opère une transformation linéaire sur les descripteurs. Celle-ci aboutit à la définition de plusieurs composantes ou encore axes représentant au mieux les données dans l'espace. Lorsque les données sont décrites en n dimensions, cette méthode va permettre de découvrir au plus n axes de coordonnées classés en fonction de la variance du nuage de points (molécules) qu'ils représentent.

Les cartes de Kohonen sont, eux, des réseaux de neurones non supervisés. Lorsqu'on cherche à représenter un espace chimique, la couche de neurones en entrée compte autant de neurones que de descripteurs. En général, la couche de neurones de sortie se compose de neurones organisés en surface de tore. Ceux-ci correspondent alors à une représentation en deux dimensions de l'espace chimique.

Différentes applications implémentent les différentes techniques de visualisation et d'analyse des ensembles de molécules précédemment citées :

- Neuroscale [93] est une technique utilisant les réseaux de neurones s'apparentant au mapping de Sammon et au MDS (MultiDimensional Scaling défini dans le paragraphe 1.4.2). Neuroscale offre une meilleure flexibilité que ces deux techniques pour effectuer la transformation des données.
- L'application commerciale SpotFire [94] utilise l'ACP pour produire un espace de visualisation.
- Maniyar et al. [88] proposent un logiciel incluant la technique des GTM et interfaçable avec d'autres logiciels commerciaux comme Pipeline Pilot.
- En 2001, Oprea et al. [95] ont développé ChemGPS qui positionnent les nouvelles structures dans un espace-médicament. Celui-ci est borné par un nombre restreint de molécules et un score de prédiction basé sur les résultats d'une ACP permet le positionnement dans cet espace.

Grâce à ces techniques graphiques, qui permettent d'obtenir un espace à 2 ou 3 dimensions dans lequel chaque molécule peut être représentée par un point à une position donnée, la diversité des librairies peut être aisément visualisée. Ainsi la couverture de l'espace par ces points donne une information sur la diversité de la collection projetée dans cet espace. Cependant il ne faut pas oublier que réduction de dimension implique également perte d'information. Dans ce genre d'espace toutes les relations entre les molécules ne sont pas représentées. Il se peut même que deux molécules paraissant proches dans cet espace, ne le soient pas du tout en réalité. Enfin l'inconvénient de ces techniques réside dans le fait que les molécules sont représentées par des variables composites issues de combinaisons de descripteurs. Et nous avons vu que de telles combinaisons ne permettent pas une compréhension aisée des phénomènes régissant l'organisation des molécules dans l'espace.

D'autres techniques existent pour représenter les molécules dans l'espace. Celles-ci utilisent plutôt les distances inter-moléculaires pour créer de nouveaux espaces, nous en donnons quelques exemples dans le paragraphe suivant.

1.4.2 Espaces "pairwise molecule coding"

Tout d'abord, le MultiDimensional Scaling (MDS) a été présenté par Torgerson [96] en 1958 pour la première fois. Cette méthode se base sur une matrice de distances entre les objets à représenter dans l'espace. Etant donné cette matrice, le but du MDS est de trouver les vecteurs de coordonnées des objets tels que les distances entre eux soient préservées dans le nouvel espace défini par cet ensemble de coordonnées et dont on a choisi le nombre de dimensions.

Ce problème est souvent formulé comme un problème d'optimisation où les vecteurs de coordonnées sont considérés comme des minimiseurs d'une fonction de coût. Le MDS est utilisé dans de nombreux domaines tels que la physique, la biologie ou encore la psychologie.

Godden et Bajorath en 2006 [97], ont développé une fonction de distance : la fonction "activity-centered". Pour chaque activité connue, leur méthode centre les espaces chimiques à haute dimension sur un sous-espace peuplé par un jeu de composés dont les activités connues sont similaires. Puis la distance euclidienne entre le centre de ce sous-espace et tous les composés de la base est calculée. Cette distance est considérée comme une mesure de dissimilarité entre chaque composé et les molécules actives connues. On peut ainsi l'utiliser pour produire un classement des produits de la base en fonction de leur similarité à des composés actifs. Ils définissent ensuite le "rayon d'activité" qui correspond à la distance moyenne de tous les composés actifs d'un sous-espace à leur centre.

Cette nouvelle représentation simplifiée de l'espace multidimensionnel leur permet

d'identifier des composés actifs connus parmi de très grandes bases de molécules, et ce en un temps raisonnable.

Raghavendra et Maggiora [85] utilisent un ensemble de molécules p restreint pour définir un espace chimique. Chacune de ces molécules est alors considérée comme un vecteur moléculaire abstrait de base. Une matrice de dimension $p \times p$ est créée à partir du produit scalaire de chacun de ces vecteurs avec lui-même et les autres. Ils utilisent ensuite la similarité entre ces molécules de base pour effectuer une transformation de cette matrice. Ils en déduisent alors un jeu de vecteurs de base qui permettent de décrire un nouveau repère pour l'espace chimique.

Ce repère peut alors être utilisé pour décrire d'autres molécules dans l'espace correspondant.

Comme nous l'avons dit ces techniques permettent de visualiser des similarités et la diversité de jeux de molécules. Mais lorsque des librairies sont comparées, les résultats peuvent être différents d'un espace à l'autre. C'est pourquoi nous décrirons des mesures mathématiques dans la section suivante qui permettent de comparer les composés chimiques de manière quantitative et non plus de manière visuelle.

1.5 Similarité/ dissimilarité

La notion de similarité est souvent reliée au principe selon lequel "deux molécules avec une structure similaire possèdent des propriétés similaires" [98]. Ce principe a été très discuté quant à son universalité [99, 100, 101]. Mais c'est bien sur celui-ci que reposent certaines étapes de recherche d'un médicament. Par exemple, lorsque l'on cherche à sélectionner un sous-ensemble de produits divers pour les tester sur une cible inconnue, l'idée sous-jacente est de tester des molécules de structures différentes. Ou encore lorsqu'une molécule réagit avec une cible, on cherche les molécules de structures similaires pour tenter d'en découvrir une plus active.

Il nous faut donc des mesures pour quantifier cette similarité ou à l'inverse cette distance entre molécules. Dans un premier temps nous présenterons les principales métriques existantes ainsi que leurs propriétés. Puis nous exposerons une autre façon de comparer les molécules via leur structure.

1.5.1 Métriques

Pour comparer les objets entre eux, on utilise donc une notion de proximité. Celle-ci peut être exprimée par une mesure de similarité, dissimilarité ou par une distance. La construction et le choix de cette mesure sont déterminants pour le résultat d'une sélection de composés. Il convient donc d'adapter la mesure de comparaison aux données et au domaine concerné. Par exemple, les mesures seront différentes selon que les données sont quantitatives ou qualitatives.

Dans un premier temps nous présentons les propriétés d'une mesure de similarité puis nous exposerons quelques mesures de similarité adaptées aux différents types de variables.

1.5.1.1 Propriétés

Une mesure de similarité est une application réelle positive symétrique s de $\mathcal{M} \times \mathcal{M} \longrightarrow \mathbb{R}^+$ telle que la similarité entre un objet et lui même $s(m_i, m_i)$ est maximale. Plus deux objets m_i et m_j sont similaires et plus cette mesure $s(m_i, m_j)$ est élevée. De la même

manière, on peut définir la mesure de dissimilarité entre deux objets m_i et m_j par $d(m_i, m_j)$ avec les propriétés opposées à la mesure de similarité présentée avant. Nous présentons les propriétés de minimalité, de symétrie, d'identité et d'inégalité triangulaire (citées de la thèse [102]) pour définir les notions d'indice de dissimilarité ou de distance.

Propriété 1.1 ***Minimalité*** *: Une mesure de dissimilarité* $d : \mathcal{M} \times \mathcal{M} \longrightarrow \mathbb{R}^+$ *vérifie la propriété de minimalité si et seulement si :*

$$\forall m_i \in \mathcal{M},\ d(m_i, m_i) = 0$$

Propriété 1.2 ***Symétrie*** *: Une mesure de dissimilarité* $d : \mathcal{M} \times \mathcal{M} \longrightarrow \mathbb{R}^+$ *est symétrique si et seulement si :*

$$\forall m_i, m_j \in \mathcal{M},\ d(m_i, m_j) = d(m_j, m_i)$$

Propriété 1.3 ***Identité*** *: Une mesure de dissimilarité* $d : \mathcal{M} \times \mathcal{M} \longrightarrow \mathbb{R}^+$ *vérifie la propriété d'identité si et seulement si :*

$$\forall m_i, m_j \in \mathcal{M},\ d(m_i, m_j) = 0 \Rightarrow m_i = m_j$$

Propriété 1.4 ***Inégalité triangulaire*** *: Une mesure de dissimilarité* $d : \mathcal{M} \times \mathcal{M} \longrightarrow \mathbb{R}^+$ *vérifie l'inégalité triangulaire si et seulement si :*

$$\forall m_i, m_j, m_k \in \mathcal{M},\ d(m_i, m_j) \leq d(m_i, m_k) + d(m_k, m_j)$$

On appelle distance, une mesure qui vérifie les quatres propriétés précédemment citées (cf. Tableau 1.2) alors qu'un indice de dissimilarité ne vérifie que les propriétés de minimalité et de symétrie.

Type de mesure	Minimalité	Symétrie	Identité	Inégalité triangulaire
Indice de dissimilarité	X	X		
Distance	X	X	X	X

TABLE 1.2: Propriétés mathématiques des indices de dissimilarités et des distances

1.5.1.2 Les différentes mesures de similarité

Nous citons deux types de mesures adaptées respectivement aux données numériques et aux données dites symboliques ou qualitatives. Pour les besoins des définitions, nous donnons quelques notations :

- Soit $\mathcal{M}$ un ensemble de n molécules noté $\mathcal{M} = \{m_i\}_{i=1...n}$
- et $\mathcal{V}$ un ensemble de p variables noté $\mathcal{V} = \{v_j\}_{j=1...p}$ où $v_j(m_i)$ indique la valeur que prend la variable v_j pour la molécule m_i

Mesures pour les variables numériques La distance la plus connue et la plus utilisée est la distance euclidienne, cas particulier de la distance de Minkowski que l'on définit comme suit :

Définition 1.1 ***Distance de Minkowski***

$$d(m_i, m_j) = (\sum_{k=1...p} |v_k(m_i) - v_k(m_j)|^l)^{1/l}$$

Selon les valeurs que prend le paramètre l, on distingue :
- la distance euclidienne avec $l = 2$
- la distance de Manhattan avec $l = 1$
- la distance de Chebychev avec $l = \infty$

Une autre distance est également très utilisée : celle du cosinus. La distance du cosinus correspond au cosinus de l'angle θ formé par les deux vecteurs m_i et m_j (ce sont en effet des vecteurs de variables) :

Définition 1.2 ***Distance du cosinus***

$$d(m_i, m_j) = cos(\theta) = \frac{m_i \cdot m_j}{\parallel m_i \parallel \parallel m_j \parallel}$$

où "$m_i \cdot m_j$" désigne le produit scalaire $\sum_{k=1...p} v_k(m_i)\ v_k(mj)$ et $\parallel m_i \parallel$ la norme de m_i soit $\sqrt{\sum_k v_k(m_i)^2}$.

Mesures pour les variables symboliques Quand les variables sont qualitatives, les distances citées plus haut n'ont pas de sens. Les indices de similarité les plus couramment utilisés dans le cadre des valeurs qualitatives sont les indices de Rand[103] et de Jaccard [104]. Ils permettent de comparer deux objets m_i et m_j, dont les vecteurs d'attributs respectifs sont notés A et B, en effectuant un comptage des propriétés communes. L'indice de Rand permet une comparaison de vecteurs symétriques [7], alors que l'indice de Jaccard permet une comparaison de vecteurs asymétriques [8]. Pour notre étude nous ne détaillons que le coefficient de Jaccard, sa forme générale est la suivante :

Définition 1.3 ***Indice de Jaccard***

$$J(m_i, m_j) = \frac{|A \cap B|}{|A \cup B|}$$

En règle générale, l'espace des descriptions constitué de variables qualitatives est redéfini avec des attributs binaires. On appelle cela un codage disjonctif complet. Par exemple, la variable "couleur des yeux" prenant 3 modalités (bleu, vert, marron) est redéfinie par 3 variables bi-modales : "bleu" : oui ou non, "vert" : oui ou non, "marron" : oui ou non (les modalités oui et non peuvent prendre respectivement les valeurs 1 et 0 pour obtenir des vecteurs d'attributs binaires). Le comptage de propriétés communes devient alors plus aisé. Nous définissons 4 compteurs avec lesquels on peut construire le tableau de contingence (cf. Tableau 1.3) :
- a : le nombre de bits valant 1 dans le vecteur A
- b : le nombre de bits valant 1 dans le vecteur B
- c : le nombre de bits valant 1 partagés par les vecteurs A et B
- p : le nombre de bits total

L'indice de Jaccard devient donc pour la comparaison de vecteurs de bits asymétriques :

$$J(m_i, m_j) = \frac{c}{(a-c) + (b-c) + c} = \frac{c}{a + b - c}$$

7. toutes les modalités d'une variable sont prises en compte de la même façon

8. Certaines modalités ne sont pas prises en compte car elles ne comportent que peu d'intérêt dans la comparaison. Par exemple en chémoinformatique, lorsque l'on compare des vecteurs de bits codant pour l'absence ou la présence de certains fragments. La présence de fragment a une signification alors que l'absence de fragment ne donne pas d'information intéressante sur le composé. Dans ce cas, le fait que deux molécules aient la valeur 0 pour un fragment n'aura pas d'influence dans la comparaison.

m_j / m_i	0	1
0	$p-(a+b-c))$	$a-c$
1	$b-c$	c

TABLE 1.3: Tableau de contingence pour des vecteurs binaires

Toutefois, cet indice est plus adapté dans le cas de descripteurs initialement binaires. En effet, une redéfinition de l'espace des descriptions telle que nous l'avons exposée risque de donner trop de poids à certaines variables comportant beaucoup de modalités. En effet, si on a 2 descripteurs : l'un X prenant deux modalités et l'autre Y prenant dix modalités ; le descripteur X étant déjà binaire, il ne changera pas, en revanche le descripteur Y sera redéfini en 10 descripteurs binaires. Lorsqu'on effectuera un calcul de similarité entre deux objets décrits par ces variables, le descripteur Y aura dix fois plus de poids dans la similarité que le descripteur X.

Indice de Jaccard et coefficient de Tanimoto En chémoinformatique l'un des coefficients de similarité le plus utilisé est le coefficient de Tanimoto. Défini pour la première fois en 1957 [105], ce coefficient peut être écrit sous sa forme générale de la même façon que l'indice de Jaccard (cf. définition 1.3). Ces deux mesures de similarité sont donc souvent considérées comme équivalentes. Cependant, Cha et al. [106] montrent qu'ils diffèrent sous leur forme de vecteur numérique et par la façon dont ils sont dérivés. Néanmoins, pour des vecteurs binaires, non seulement les équations de l'indice de Jaccard et du coefficient de Tanimoto sont équivalentes, mais en plus l'inégalité triangulaire de la distance de Tanimoto a été prouvée dans ce cadre [107]. En chémoinformatique, le coefficient de Tanimoto est utilisé pour des binaires, cela nous permet donc de considérer la distance de Tanimoto (D_T) obtenue par la formule : $D_T = 1 - T$ (avec T le coefficient de Tanimoto) comme une distance à part entière. Cette notion est importante car certains algorithmes de clustering, que nous détaillerons par la suite, sont conçus pour obtenir des résultats optimaux avec de vraies distances.

Résumé des mesures de similarité et distances D'autres distances et mesures de similarité comme la distance de Soergel et le coefficient de Dice sont souvent utilisées en chémoinformatique. Nous résumons dans le tableau 1.4 les principales mesures de similarité, dissimilarité et de distances citées notamment par [102, 108, 109, 110].

Mesures	Formule variables continues	Formule variables binaires
Distance de Hamming ou de Manhattan	$D_{m_i,m_j} = \sum_{k=1...p} \lvert v_k(m_i) - v_k(m_j) \rvert$	$D_{m_i,m_j} = a + b - 2c$
Distance Euclidienne	$D_{m_i,m_j} = \sqrt{\sum_{k=1...p} (v_k(m_i) - v_k(m_j))^2}$	$D_{m_i,m_j} = \sqrt{a + b - 2c}$
Distance de Soergel	$D_{m_i,m_j} = \frac{\sum_{k=1...p} \lvert v_k(m_i) - v_k(m_j) \rvert}{\sum_{k=1...p} max(v_k(m_i), v_k(m_j))}$	$D_{m_i,m_j} = 1 - \frac{c}{a+b-c} = \frac{a+b-2c}{a+b-c}$
Coefficient du cosinus	$S_{m_i,m_j} = \frac{\sum_{k=1...p} v_k(m_i) v_k(m_j)}{\sqrt{\sum_{k=1...p} v_k(m_i)^2 \sum_{k=1...p} v_k(m_j)^2}}$	$S_{m_i,m_j} = \frac{c}{\sqrt{ab}}$
Coefficient de Tanimoto ou de Jaccard	$S_{m_i,m_j} = \frac{\sum_{k=1...p} v_k(m_i) v_k(m_j)}{\sum_{k=1...p} v_k(m_i)^2 + \sum_{k=1...p} v_k(m_j)^2 - \sum_{k=1...p} v_k(m_i) v_k(m_j)}$	$S_{m_i,m_j} = \frac{c}{a+b-c}$
Distance de Tanimoto		$D_{m_i,m_j} = 1 - \frac{c}{a+b-c}$
Coefficient de Dice	$S_{m_i,m_j} = \frac{2 \sum_{k=1...p} v_k(m_i) v_k(m_j)}{\sum_{k=1...p} v_k(m_i)^2 + \sum_{k=1...p} v_k(m_j)}$	$S_{m_i,m_j} = \frac{2c}{a+b}$

TABLE 1.4: Tableau récapitulatif des mesures les plus utilisées en apprentissage et en chémoinformatique, a = variables à 1 pour m_i, b = variables à 1 pour m_j et c = variables à 1 pour m_i et m_j

Nous remarquons que dans le cas des vecteurs de bits, la distance de Tanimoto est équivalente à la distance de Soergel.

1.5.2 Les sous-structures

Une autre façon de comparer les molécules réside dans la comparaison de sous-structures. Il existe plusieurs types de sous-structures tels que les scaffolds, ou les frameworks définis par Murcko [111, 112]. La figure 1.7 montre une façon d'obtenir ces sous-structures à partir d'une molécule représentée en deux dimensions. L'étape 1 consiste à supprimer les atomes d'hydrogène, puis dans l'étape 2 on supprime successivement les atomes avec une seule liaison. On obtient ainsi un scaffold (étape 3). Pour obtenir un framework (étape 4), il suffit de définir tous les types d'atomes comme étant des carbones et de définir tous les types de liaisons comme des liaisons simples. Dans cette figure, on a fait le choix de conserver les liaisons dites "aromatiques" comme telles.

Les scaffolds et les frameworks sont en quelque sorte le squelette d'un composé et peuvent constituer des familles dans lesquelles regrouper les produits chimiques. On peut d'ailleurs remarquer que la plupart des médicaments se situent dans seulement quelques familles. Avec ces sous-structures, on considère que deux molécules sont similaires lorsqu'elles possèdent le même squelette. Des techniques récentes ont été développées pour prendre en compte cette notion de structure commune

- Fitzgerald et al. [113] en 2006 introduisent une méthode pour comparer les librairies en terme de similarité. Cette méthode est basée sur les scaffolds et permet à un utilisateur de choisir parmi plusieurs librairies celle qui lui convient structuralement.
- Batista et al. [114] ont proposé la méthode MolBlaster pour évaluer la similarité. Cette technique dépend des profils de fragments des molécules. Elle présente l'avantage d'être indépendante des descripteurs.
- 2007 : Rupp et al. [115] utilisent des graphes annotés pour mesurer la similarité moléculaire. Les molécules sont représentées par des graphes dans lesquels les atomes sont des noeuds et les liaisons des arêtes. Couramment les graphes sont compressés sous forme de fingerprints qui sont comparés pour évaluer la similarité entre les molécules. Mais la compression entraîne une perte d'information. Pour éviter cette perte, ils effectuent une comparaison directe des graphes. Leurs résultats montrent que leur méthode est plus efficace qu'une méthode basée sur les descripteurs.

1.5.3 Les limites de ces mesures

Ces mesures de similarité/dissimilarité ne prennent souvent en compte qu'une partie de l'information concernant la molécule (la structure, des fragments pour les fingerprints, les descripteurs non binaires pour la distance euclidienne...). A ma connaissance seul le score EDI (Explicit Diversity Index) proposé par Papp et al. [116] combine propriétés physico-chimiques et scaffolds pour évaluer la diversité d'une base. Mais ce score donne une trop forte importance aux structures, ce qui a pour effet d'annuler la prise en compte des propriétés physico-chimiques. Ainsi ces mesures peuvent donner des résultats très différents voire contradictoires quant à la similarité entre deux molécules. Un exemple est présenté dans la figure 1.8. En effet, on peut voir en haut de la figure que deux composés ayant le même framework peuvent avoir un score de similarité (par exemple : le coefficient de Tanimoto présenté dans la section précédente) indiquant qu'ils ne sont pas similaires. Et à l'inverse, dans le bas de la figure on peut voir deux composés considérés comme similaires par le coefficient de Tanimoto et ccomme non similaires par leurs frameworks respectifs.

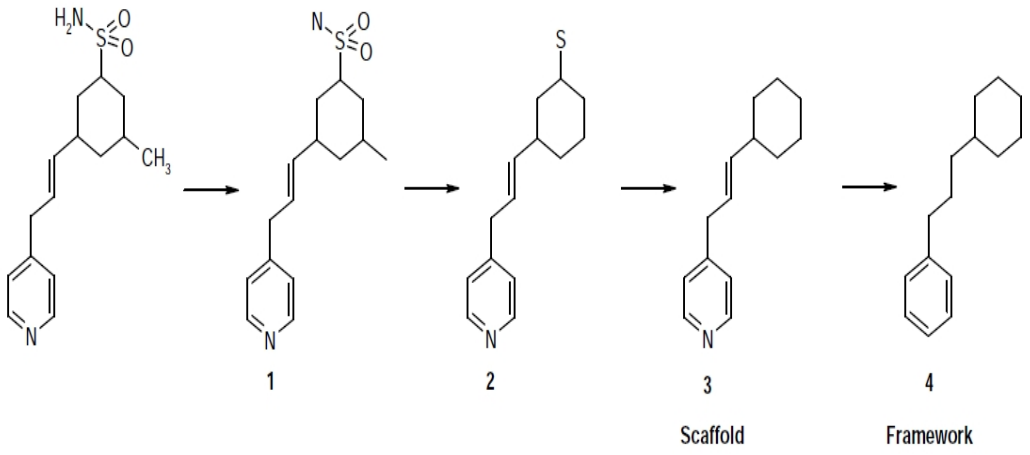


FIGURE 1.7: Un exemple d'obtention de scaffold et de framework à partir d'une structure, d'après [108]

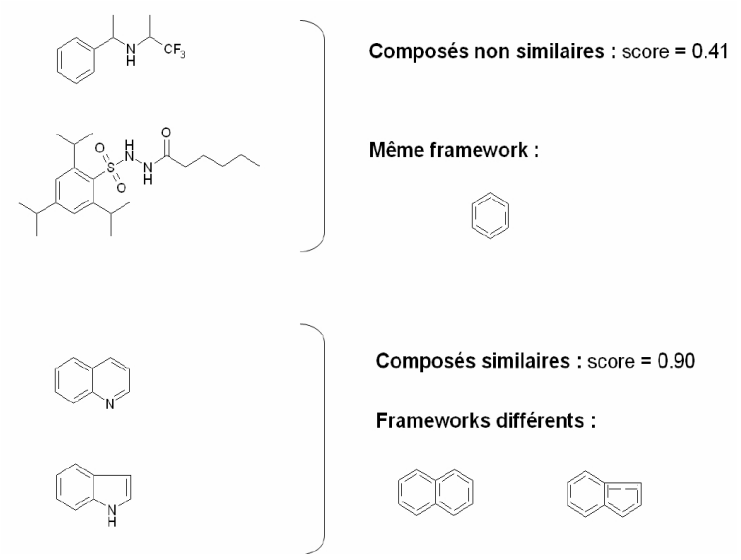


FIGURE 1.8: Différences entre la similarité de frameworks et la similarité de score, d'après [108]

1.6 Diversité chimique

Quand on n'a aucune connaissance sur les cibles biologiques, on doit tester un maximum de molécules chimiques pour tenter d'en trouver une active. Or comme il existe plusieurs millions de molécules différentes, il est impossible de les tester toutes pour des raisons de coût et de temps. Il faut donc sélectionner un sous-ensemble de molécules à tester. Afin de ne pas effectuer de tests redondants, les méthodes de sélection se basent sur le principe de similarité, déjà évoqué, selon lequel les molécules similaires doivent avoir une activité biologique similaire [98]. De ce fait, tester une seule molécule d'un groupe similaire donne une estimation raisonnable du potentiel d'activité des autres membres du même groupe. C'est pourquoi, de nombreuses méthodes ont été développées pour effectuer des sélections de sous-sensembles divers.

Nous tenterons tout d'abord de donner une définition de la diversité et les différentes définitions qui ont été données dans la littérature. Ensuite nous présenterons les différents types de sélection par diversité ainsi que quelques exemples de méthodes. Puis nous donnerons le cadre formel du clustering qui sera le cadre principal de notre étude. Enfin nous exposerons un état de l'art de l'évaluation des sélections par diversité.

1.6.1 Diversité et Représentativité

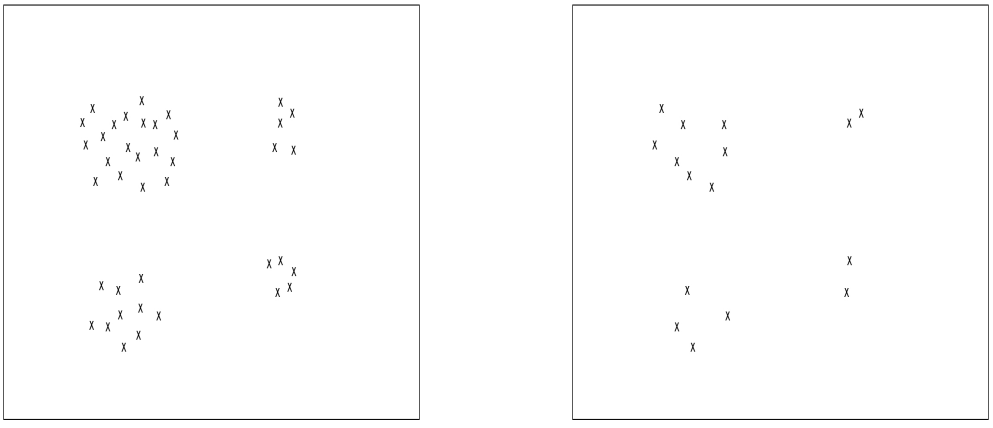
Pour construire les méthodes de sélection, deux notions sont évoquées dans le domaine de la chémoinformatique : la représentativité et la diversité. Ces deux notions peuvent avoir plusieurs définitions que nous exposons ici.

1.6.1.1 La représentativité

La représentativité d'un échantillon peut être définie de différentes façons. Dans un premier temps nous donnons la définition statistique de ce terme, puis nous donnerons les définitions utilisées en chémoinformatique qui, nous allons le voir, ne sont pas forcément équivalentes à la première.

Définition statistique En statistique, un échantillon est dit représentatif lorsqu'il possède les mêmes caractéristiques que la population dont il est extrait. Un échantillonnage aléatoire (tirage aléatoire uniforme sans remise) est l'un des meilleurs moyens de parvenir à la représentativité : chaque individu de la population initiale (ou population mère) a une même probabilité de figurer dans l'échantillon.

Dans le cadre de cette définition statistique, un sous-ensemble de molécules représentatif serait un échantillon couvrant les différentes zones de l'espace de description de la même manière que la population initiale. Le schéma 1.9(a) montre une population initiale couvrant un espace en deux dimensions de façon non homogène. Un tirage aléatoire dans cette population pourra donner l'échantillon du schéma 1.9(b). Les zones les plus densément couvertes par la population initiale le sont également par l'échantillon.



(a) Population initiale

(b) Echantillon représentatif

FIGURE 1.9: Illustration de la représentativité d'un échantillon obtenu par tirage aléatoire

Par la suite, pour éviter les confusions nous appelerons cette représentativité : la représentativité statistique.

Représentativité en chémoinformatique En chémoinformatique, lorsque l'on parle de la représentativité d'un échantillon, ce n'est pas toujours la définition statistique qui est utilisée. Clark et Langton [117] en 1998 l'utilisent dans le sens statistique et la définissent ainsi : "Un sous-ensemble représentatif est un ensemble dans lequel la distribution de ses membres reflète la distribution de la population entière".

Mais par exemple, Bayada et al. en 1999 [118] ne considèrent pas la représentativité de la même façon : "La représentativité est souvent considérée comme la sélection de composés qui sont typiques d'un jeu de données particulier, c'est à dire qu'ils sont sélectionnés à partir d'un cluster (ou groupe) de données, un composé pour chaque groupe". Cette définition est différente car elle amène la notion de groupes de molécules. Nous verrons, avec les définitions de diversité données dans le paragraphe suivant, que ces groupes sont souvent apparentés à des activités biologiques potentielles. Cette notion de représentativité est très proche de la notion de diversité qui nous intéresse et que nous détaillons par la suite.

1.6.1.2 La diversité

Il n'existe pas de définition unique de la diversité. Certaines se basent sur la distance, d'autres sur des notions d'espaces. Toutes se basent plus ou moins directement sur le principe de similarité cité précédemment.

D'ailleurs Hassan et al. en 1996 [119] nous rappellent bien ce principe sous-jacent au souhait de maximiser la diversité des librairies testées : "Le raisonnement de cette approche est que, en général, une libraire chimiquement diverse aura de meilleures chances de trouver des composés actifs dès les premiers tests de criblage qu'une librairie contenant des molécules semblables". Ils opposent donc la notion de diversité d'un ensemble à celle de similarité. Ils décrivent alors les sous-ensembles de molécules hautement divers comme étant : "Les sous-ensembles de molécules qui couvrent l'espace des propriétés du jeu entier".

Cette notion de couverture d'un espace est également utilisée par Ferguson et al. [120] qui expliquent que les algorithmes de sélection par diversité sont basés sur la supposition selon laquelle : "un jeu de composés divers qui couvre l'espace structural (défini en termes de paramètres physico-chimiques, de fragments sous-structuraux ou d'indices topologiques) couvrira l'espace de l'activité biologique".

En 2006, Maldonado et al. [109] proposent que "l'analyse de la diversité moléculaire explore la façon dont les molécules couvrent un espace structural déterminé".

Ensuite d'autres définitions sont plutôt basées sur la distance entre molécules. Notamment Clark et Langton [117] définissent un sous-ensemble divers : "Les membres d'un sous-ensemble divers sont facilement distingables les uns des autres, ce qui signifie qu'ils sont relativement dissimilaires de chaque autre".

Bayada [118] propose également une définition de la diversité basée sur la distance inter-moléculaire en opposition à sa définition de la représentativité : "La sélection par diversité est indépendante de groupes de données et [...] le critère de sélection est basé sur la distance aux autres composés [...]". Il propose alors que la sélection par diversité reviennent à "sélectionner le plus petit jeu possible de composés couvrant autant d'activités biologiques que représentées par le jeu entier". En 2004, Stockwell [121] définit la diversité comme une description quantitative : "La diversité d'une librairie est une description quantitative de combien les composés sont différents les uns des autres [...] Les descripteurs conduisent à une description quantitative de la diversité chimique". Mais il précise que : "La diversité de structures chimiques n'implique par forcément la diversité

d'activités biologiques".

Il existe plusieurs notions sous-jacentes à la notion de diversité : la distance entre molécules mais également la couverture d'un espace de propriétés indirectement lié à un espace d'activités biologiques potentielles. Et alors que certains auteurs mettent en opposition les notions de diversité et de représentativité [117, 118], d'autres proposent une notion de diversité chevauchante avec la définition de représentativité (non statistique) citée plus haut.

Notamment avec Reynolds et al. [122] qui proposent de "sélectionner un sous-ensemble divers de composés qui soient représentatifs d'une large libraire structurale". Ou encore Papp et al. [116] qui définissent un sous-ensemble divers comme "un jeu de composés le plus représentatif possible qui couvre l'espace chimique pertinent par rapport à la cible [biologique] appropriée qui est visée".

Dans un premier temps, le problème soulevé par les méthodes basées uniquement sur les distances inter-moléculaires est qu'elles ont tendance à produire beaucoup d'outliers. Dans ce cas, l'espace des descriptions est bien couvert aux extrémités mais mal couvert au centre.

Ensuite, les méthodes uniquement basées sur la représentativité (telles que le clustering que nous expliquons dans la section suivante) sélectionnent des composés à partir de groupes formés dans l'ensemble de molécules. Cependant ces groupes ne sont pas toujours distribués homogènement dans l'espace (comme nous le démontrerons dans le chapitre 4 pour les k-means par exemple).

Or nous souhaitons définir un sous-ensemble à la fois tel que la distance entre les molécules sélectionnées soit la plus grande possible et tel que l'espace occupé par le jeu initial soit couvert de la façon la plus homogène possible par cet échantillon. En effet, pour que les tests biologiques ne soient pas redondants, on souhaite que l'échantillon soit divers, mais pour que chaque activité potentielle soit représentée, nous souhaitons également que chaque molécule initiale soit représentée dans l'échantillon. C'est pourquoi d'une part nous développerons plusieurs indices d'évaluation des échantillons (cf. chapitre 4) qui nous permettront d'étudier et d'optimiser ces deux critères. Et d'autres part nous formaliserons un nouveau critère de diversité prenant en compte ces deux contraintes. Nous comparerons ensuite la méthode que nous avons développée à partir de ce critère à d'autres méthodes existantes dont nous citons des exemples dans la section suivante.

1.6.2 Types de sélection

Dans le domaine de la chémoinformatique, on classe souvent les méthodes en 4 catégories de sélection :

- Sélection basée sur les clusters
- Sélection basée sur le partitionnement
- Sélection basée sur la dissimilarité
- Sélection basée sur l'optimisation

On retrouve ce classement notamment dans Willett [123].

Ces méthodes font appel à des techniques d'apprentissage artificiel que nous avons déjà évoqué dans la section 1.2.3.5. L'apprentissage artificiel ou automatique, aussi appelé Machine Learning, est un domaine de recherche très vaste qui regroupe plusieurs approches et techniques. On distingue souvent deux types d'apprentissage : l'apprentissage supervisé et l'apprentissage non supervisé. Pour tout type d'apprentissage on définit les concepts

suivants :

- les entrées (ou observations) : ce sont les données sur lesquelles on effectue l'apprentissage
- les sorties : ce sont les étiquettes attribuées à chaque entrée. Ces étiquettes sont une connaissance que l'on possède sur les observations (par exemple : on peut étiqueter des molécules avec leur activité biologique).

Dans le cas de l'apprentissage supervisé, il s'agit d'apprendre une fonction (f) à partir de couples d'entrées (X) - sorties (Y) telle que f(X)= Y.

Dans le cas de l'apprentissage non supervisé, il s'agit de discerner des motifs dans l'entrée alors qu'aucune valeur de sortie spécifique n'est connue. Plus précisément, il s'agit de découvrir une organisation ou classification des entrées sans connaissance a priori sur ces données.

Dans notre cas, on se situe dans le cadre de l'apprentissage non supervisé car nous avons des observations (les molécules et leur description) et nous ne possédons aucune connaissance a priori sur un quelconque classement de celles-ci.

Enfin dans le domaine de l'apprentissage artificiel, les méthodes dites de partitionnement sont considérées comme des méthodes de clustering. Le partitionnement tel qu'il est mentionné en chemoinformatique correspond au clustering par grille en apprentissage. Bien que le partitionnement soit donc conceptuellement relié à des méthodes de clustering conventionnelles, l'algorithme est assez différent. Pour le clustering, les molécules regroupées forment un cluster alors que pour les techniques de partitionnement l'espace chimique est prédécoupé en cellules. Nous présenterons donc les méthodes basées sur le clustering et le partitionnement séparément. Nous définissons tout d'abord le cadre formel du clustering. Ces définitions permettront de mieux décrire les méthodes que nous comparons dans le chapitre 4.

1.6.2.1 Cadre formel du clustering

Nous présentons ici quelques définitions et concepts utiles au clustering tirés de la thèse de [102]. Les notations utilisées sont adaptées à notre problème qui concerne un ensemble de molécules. On considère un ensemble de n objets (molécules) noté $\mathcal{M} = \{m_1, ..., m_n\}$. La dissimilarité entre deux objets m_i et m_j est notée $d(m_i, m_j)$. L'ensemble des dissimilarités entre les objets deux à deux peut être stockée dans une matrice de taille $n \times n$ notée D. Chaque objet est souvent décrit par un ensemble de p variables noté $\mathcal{V} = \{v_1, ..., v_p\}$. Ces variables sont telles que $v_j(m_i)$ désigne la valeur de l'objet $m_i \in \mathcal{M}$ pour la variable $v_j \in \mathcal{V}$

Le clustering génère un ensemble de k clusters (ou groupes), noté $\mathcal{C} = \{C_1, ...C_k\}$, tel que chaque cluster C_l est un sous-ensemble de $\mathcal{M}$ ($C_l \subset \mathcal{M}$) et l'union des clusters contient l'ensemble des objets de départ ($\bigcup_{l=1...k} \ C_l = \mathcal{M}$).

Définition 1.4 *$\mathcal{C}$ est une **Partition** de $\mathcal{M}$ si et seulement si $\mathcal{C}$ vérifie les propriétés suivantes :*

- *$C_l \subset \mathcal{M}$ pour tout $C_l \in \mathcal{C}$ et $C_l \neq \emptyset$*
- *$\bigcup_{l=1...k} \ C_l = \mathcal{M}$*
- *$C_i \bigcap C_j = \emptyset$ pour(i, j) tel que $i \neq j$*

La dernière propriété signifie qu'aucun objet ne peut se trouver dans deux clusters à la fois.

Certaines méthodes de clustering fournissent un arbre hiérarchique en sortie, appelé également dendogramme. Nous définissons donc les propriétés de la hiérarchie comme suit :

Définition 1.5 *Soit P un ensemble de parties non vides sur $\mathcal{M}$, P est une* ***hiérarchie*** *si les propriétés suivantes sont vérifiées :*

- $\mathcal{M} \in P$
- $\forall m_i \in \mathcal{M}$, $\{m_i\} \in P$
- $\forall h, h' \in P, h \bigcap h' \in \{\emptyset, h, h'\}$
- $\forall h \in P, \bigcup \{h' \in P : h' \subset h\} \in \{h, \emptyset\}$

La racine de l'arbre est constituée de l'ensemble $\mathcal{M}$ selon la première propriété, et les feuilles de l'arbre sont les singletons $\{m_1\}, ..., \{m_n\}$ selon la deuxième propriété. Les deux dernières propriétés permettent que deux clusters ne puissent s'intersecter que si l'un est inclus dans l'autre. Chaque cluster doit contenir tous ses successeurs aussi appelés clusters fils et doit être contenu dans son unique cluster prédécesseur appelé aussi cluster père.

Les méthodes de clustering associent souvent chaque groupe de la partition à un seul point pour différentes raisons (notamment pour des raisons de complexité). Ce point est considéré comme "représentatif" du cluster et peut être un centroïde ou un médoïde.

Définition 1.6 *Le* ***centroïde*** *noté m^* d'un cluster C_l est le point défini dans l'ensemble $\mathcal{V}$ par :*

$$\forall j = 1, ..., p, \quad v_j(m^*) = \frac{1}{|C_l|} \sum_{m_i \in C_l} v_j(m_i)$$

Dans le cas où toutes les variables dans $\mathcal{V}$ sont quantitatives, ce centroïde est décrit sur chaque dimension par la valeur moyenne de l'ensemble des objets du groupe qu'il représente. De ce fait, il arrive souvent que ce centroïde n'appartienne pas aux objets dans $\mathcal{M}$. On parle alors d'objet virtuel. Enfin le centroïde constitue le centre de gravité du cluster.

Cependant lorsque les variables sont de types qualitatives, le centroïde n'a pas de sens puisqu'on ne peut appliquer les opérateurs classiques comme l'addition ou la division sur un tel type de variable. Ou encore on peut souhaiter que l'objet représentant le cluster soit un objet réel, c'est à dire appartenant aux objets du groupe. Dans ces deux cas on cherche donc l'objet appartenant au cluster le plus représentatif de celui-ci. On l'appelle alors médoïde (il peut également être nommé prototype).

Définition 1.7 *Le* ***médoïde*** *noté m^* d'un cluster C_l est alors le point défini dans l'ensemble $\mathcal{V}$ par :*

$$m^* = \underset{m_i \in C_l}{ArgMin} \frac{1}{|C_l|} \sum_{m_j \in C_l} d(m_i, m_j)$$

Cette définition signifie que le médoïde est l'objet du cluster dont la moyenne des distances entre lui et les objets du cluster est minimum. En d'autres termes, c'est le plus proche de tous les objets du groupe.

Connaissant le centre d'un cluster, qu'il soit médoïde ou centroïde, on peut alors définir le rayon et le diamètre d'un cluster comme suit :

Définition 1.8 *Le* ***rayon*** *ρ d'un cluster C_l de représentant m^* est la plus grande distance entre le représentant (parfois appelé centre) et un autre objet appartenant au cluster :*

$$\rho(C_l) = \underset{m_i \in C_l}{Max}\ d(m_i, m^*)$$

Définition 1.9 *Le **diamètre** d'un cluster C_l centré sur m^* est la plus grande distance entre deux objets du cluster :*

$$diam(C_l) = \underset{m_i, m_j \in C_l}{Max}\ d(m_i, m_j)$$

Dans ce cas, le diamètre n'est donc pas le double d'un rayon en général.

Enfin pour les besoins de l'évaluation du clustering il convient de définir l'inertie intra-clusters, inter-clusters et la variance d'un cluster.

Définition 1.10 *L'**inertie intra-cluster** d'un cluster C_l est la somme des carrés des distances au centre m^* de C_l, soit :*

$$I_{intra}(C_l) = \sum_{m_i \in C_l} d(m_i, m^*)^2$$

Définition 1.11 *La **variance** d'un cluster C_l est la moyenne des carrés des distances au centre du cluster m^* :*

$$var(C_l) = \frac{1}{|C_l|} \sum_{m_i \in C_l} d(m_i, m^*)^2$$

Définition 1.12 *L'**inertie inter-clusters** d'une partition $\mathcal{C} = \{C_1, ..., C_k\}$ est la somme des carrés des distances entre les centres des clusters $\{m_1^*, ..., m_k^*\}$:*

$$I_{inter}(\mathcal{C}) = \sum_{i=2,...,k} \sum_{j<i} d(m_i^*, m_j^*)^2$$

1.6.2.2 Sélection basée sur les clusters

Pour cette sélection, l'ensemble initial de molécules est divisé en groupes ou clusters en utilisant un algorithme de clustering. Ces groupes ont un haut degré de similarité intra-cluster et de dissimilarité inter-clusters (ces notions sont détaillées dans la section 1.6.2.1). Puis dans chaque cluster, une molécule est sélectionnée pour former le sous-ensemble divers. Il existe plusieurs types de clustering :

- clustering hiérarchique (par exemple [124] présente le clustering hiérarchique agglomératif)
- clustering par partitionnement dont nous utiliserons une méthode (k-medoïd, k-means [125]) dans notre étude (cf. chapitre 3)
- clustering par grilles dont nous détaillons le principe dans le paragraphe suivant intitulé sélection basée sur le partitionnement
- clustering par densité (par exemple : DBSCAN [126], OPTICS [127])

De nombreux travaux utilisent cette technique de sélection basée sur le clustering, notamment [117, 118]

1.6.2.3 Sélection basée sur le partitionnement

Cette sélection recquiert le partitionnement de l'espace en cellules. Pour cela, chaque descripteur (ou dimension décrivant les molécules, au nombre de p) est divisé en plusieurs sections. Des cellules à p dimensions sont ainsi créées dans l'espace des descriptions du jeu initial. Ensuite une molécule est extraite de chaque cellule. Les travaux [128] donnent un exemple de l'utilisation de telles méthodes. L'inconvénient de telles techniques est qu'elles ne peuvent être utilisées qu'en faible dimensionnalité, c'est à dire avec peu de descripteurs (une dizaine au maximum). Or nous travaillons avec plusieurs centaines de descripteurs, ces méthodes ne sont donc pas adaptées à notre étude.

1.6.2.4 Sélection basée sur la dissimilarité

Comme dit précédemment, les méthodes de clustering et de partitionnement sont basées sur le même principe. En effet, il s'agit de former des groupes pour ensuite sélectionner une molécule dans chacun. Pour la sélection par dissimilarité, le principe est de sélectionner directement les molécules de façon itérative en cherchant celles qui apportent le plus de dissimilarité possible au sous-ensemble déjà sélectionné. Trouver le sous-ensemble le plus dissimilaire parmi tous les sous-ensembles possibles par cette technique et de manière exhaustive est considéré comme un problème NP-Complet [9]. Mais il existe plusieurs algorithmes d'approximation dont Maximum-Dissimilarity et Sphere-Exclusion [129] que nous présenterons par la suite (cf. Chapitre 3.2.3) ou encore les travaux de [130, 131]

1.6.2.5 Sélection basée sur l'optimisation

Pour ce type de sélection, des indices de diversité sont définis [132, 133], puis la recherche d'un sous-ensemble divers est formulé comme un problème d'optimisation de cet indice. Typiquement, des techniques telles que les algorithmes génétiques, ou les recuits simulés ont été utilisées comme problème d'optimisation.

1.7 Outliers

Nous avons plusieurs fois fait référence aux outliers.

Il n'existe pas de définition universelle d'un outlier et cette définition peut différer selon le domaine d'application. Comme pour la notion de diversité, nous donnerons la définition d'outliers selon le domaine statistique et selon le domaine de la chémoinformatique.

1.7.1 Cadre statistique

1.7.1.1 Définitions

En statistique, un outlier est une observation qui est numériquement distante du reste des données. Par exemple Grubbs [134] définit un outlier comme une observation qui apparaît dérivant des autres membres de l'échantillon dans lequel il se trouve.

Un outlier est souvent considéré comme une donnée anormale ou abérrante, voire comme une erreur à supprimer. Par exemple, Hawkins [135] définit un outlier comme une donnée qui dévie tellement des autres observations qu'elle laisse soupçonner que l'outlier a été produit par un mécanisme différent de celui qui a généré les autres données.

1.7.1.2 Méthodes de détection

Il existe plusieurs critères et tests statistiques pour déterminer des outliers dont nous citons quelques exemples ici :

- Le Q test de Dixon [136]
- Le critère de Peirce [137]
- Le test de Grubbs [134]

9. La classe des problèmes vérifiant une solution efficacement (en temps polynomial) est appelée NP (non deterministic polynomial). Bien qu'une solution proposée pour un problème NP-complet soit vérifiable efficacement, on ne peut pas en trouver une en temps polynomial. En général, les problèmes NP-Complet sont résolus par des algorithmes qui ont des temps d'exécution exponentiels.

Le problème de ces tests est qu'ils sont univariés (applicables sur une seule variable à la fois) alors que nous possédons souvent des données multi-variées. Il n'existe pas de méthode statistique multi-variée pour la détection d'outliers.

1.7.2 Cadre de la chémoinformatique

1.7.2.1 Définitions

En chémoinformatique, plusieurs définitions de la notion d'outliers sont possibles. La première correspond à la définition statistique où un outlier est généralement un composé avec une ou plusieurs propriétés (ou atttributs) qui diffèrent substantiellement des autres.

On trouve également des définitions basées sur les caractéristiques structurales, la définition la plus répandue repose sur l'espace chimique des descriptions. Dans ce cas, rechercher des outliers revient à identifier les composés qui sont structurellement différents des autres molécules au regard du jeu de descripteurs utilisé pour définir l'espace chimique. En 1998 par exemple, Menard et al. [138] définissent les outliers comme étant "des structures ayant des valeurs extrêmes pour leurs propriétés calculées et donc pouvant représenter des molécules inhabituelles ayant peu d'intérêt comme candidat médicament". Ils définissent ainsi des espaces chimiques de référence, à partir de jeux de molécules, en découpant chaque descripteur en intervalles égaux pour obtenir un espace quadrillé. Chaque cellule ainsi obtenue doit avoir un pourcentage seuil de remplissage en molécules. Les outliers sont alors les 5 à 10% de molécules ne permettant pas d'obtenir les intervalles de valeurs des descripteurs conférant un tel pourcentage de remplissage des cellules de l'espace. Récemment, Casalegno et al. [139] ont proposé que les outliers soient les molécules positionnées en dehors d'un espace chimique donné.

1.7.2.2 Méthodes de détection

La plupart du temps, la détection d'outliers en chémoinformatique s'effectue à l'aide de valeurs seuils pour chaque descripteur. Ce type de méthode est peu applicable lorsque l'on se trouve en haute dimensionnalité car elles supposent une connaissance expérimentale des abérrations possibles pour chaque propriété mesurée. De plus, encore une fois le traitement s'effectue de manière univariée.

Une autre manière de détecter les outliers est d'utiliser les techniques d'analyse de la diversité des composés. Certaines méthodes telles que celle proposée par Guha et al. [140] permettent une quantification de la diversité d'ensembles de composés ainsi que la détection d'outliers.

Menard et al. [141] propose une méthode de sélection par diversité basée sur le clustering. Leur méthode produit des molécules dites singletons (molécules seules dans leur groupe ou cluster). Ils les considèrent comme les molécules outliers du jeu de donné utilisé pour la sélection.

Enfin, Casalegno et al. [139] proposent une méthode de clustering récursif basé sur les fragments structuraux des molécules. Dans leur approche, les molécules sont découpées en fragments structuraux. Ils effectuent ensuite un clustering de ces molécules dont la similarité est basée sur le nombre de fragments communs. Seuls les fragments contribuant le plus à la formation des clusters constituent les descripteurs structuraux d'un sous-espace. Celui-ci est contenu dans l'espace chimique défini par l'ensemble des fragments. Ainsi le sous-espace n'étant pas défini par tous les descripteurs initiaux, certaines molécules peuvent se situer en dehors si elles possèdent des fragments ne décrivant pas ce sous-espace. Elles sont alors considérées comme des outliers.

1.7.3 Conclusion

On ne peut définir de manière absolue ce qu'est un outlier. Parfois il s'agit d'erreurs de saisie ou de mesures, parfois de composés seuls dans une partie de l'espace chimique (aux extrémités de cet espace généralement) et souvent structuralement différents des autres composés étudiés. Dans ce dernier cas, un outlier peut être intéressant à étudier. Certains utilisateurs, lors de criblages, souhaitent tester tous les mécanismes d'action possibles (sous-entendu, un maximum de structures différentes), tandis que d'autres préfèrent ne pas tester des produits qui ne possèdent pas assez de produits similaires connus dans la base ou dans une autre. En effet, la découverte d'une activité avec un tel produit signifierait qu'il faudrait soit synthétiser les produits similaires s'ils sont connus mais non possédés par l'utilisateur, soit qu'aucun autre produit ne pourrait le remplacer en cas de trop forte toxicité pour l'homme. Dans ce cas, il vaut mieux retirer les outliers pour ne pas les tester inutilement.

1.8 Publication

Ce chapitre a fait l'objet, en partie, de la publication d'une review en début de thèse sous la référence suivante :
Dubois, J. ; Bourg, S. ; Vrain, C. ; Morin-Allory, L. Collections of Compounds - How to Deal with them ? *Current Computer-Aided Drug Design.* **2008**, 156-168.
Cette review, ci-après, portait sur la gestion des chimiothèques et couvrait un sujet un peu plus général que celui de la thèse.

Collections of compounds – How to deal with them ?

J. Dubois[1,3*], *S. Bourg*[2], *C. Vrain*[3] *and L. Morin-Allory*[1]

1)Institut de Chimie Organique et Analytique, ICOA, Laboratoire de Chemoinformatique, UMR CNRS 6005, Université d'Orléans, BP 6759, F-45067 Orléans Cedex2, France
2)Fédération de recherche "Physique et Chimie du Vivant", FR CNRS 2708, Rue Charles Sadron, F-45071 Orléans Cedex2, France
3)Laboratoire d'Informatique Fondamentale d'Orléans, LIFO, Université d'Orléans, BP 6759, F-45067 Orléans Cedex2, France

Keywords : Chemical libraries, chemical collections of compounds, molecular descriptors, chemical spaces, drug-like, subset selection, molecular diversity, molecular similarity.

Abstract : Chemical libraries or databases are collections of compounds which can be screened (virtually or experimentally) in order to discover drug candidates. These libraries are very variable in their content (description of structures, molecular descriptors, literature links...) and their size (number of compounds). Over the last decade, a large number of papers have been published on the subject. In this review, we summarize these studies by introducing different types of compound collections and reviewing the main kinds of software used to manipulate them. We present the descriptors which have a fundamental role in the characterisation of the molecules, and describe how they are used to define the molecular filters applied before screening, in order to obtain both a representation of chemical spaces and selections of subsets by diversity or similarity.

Introduction

Over the last decade, the use of High Throughput Screening (HTS) has dramatically increased, but its capacity to screen a large number of compounds is still limited by comparison with the number of existing or potential chemical compounds. Furthermore, HTS is both time-consuming and costly. To solve this problem, several in-silico solutions exist, notably Virtual Screening (VS), which uses a computational approach to discover drug candidates. In spite of the constant increase in computer performance, VS is still greedy in computational time. Preprocessing of the data is required before that any complex computations and/or biological assays can be considered. The number of available chemical compounds increases quickly every year. As their descriptions become more precise, the number of values which define each molecule also increases. In order to facilitate access to these data, databases or collections of compounds are used. However, the information content of these chemical libraries is not necessarily structured in such a way that computational treatment can be applied :

each molecule is described with many values or descriptors (i.e. values which characterize the product such as lipophilicity, molecular weight, electronic surface...) ; in addition, they have a multi-dimensional structure and can contain millions of compounds which cannot all be tested in a drug discovery process. In order to reduce the size of these libraries, many methods (such as drug-like filters, diversity selection...) have therefore been developed. Diversity selection, for example, gives smaller-sized sets in which the molecules are as representative as possible of the structural diversity of the total chemical space. A good realization of these different steps (representation of molecules, filters, selection of subsets) in the drug discovery process is very important in order to prepare Virtual Screening. Several papers have addressed these problems in manipulation of compound collections. In 1998 and 1999, some studies dealt with the topic of diversity in combinatorial chemistry [1-3]. They were followed by work presenting techniques to design combinatorial libraries, focusing on diversity selection and drug-like properties [4-6]. In 2005, Lumley [7] discussed these types of selection and filtering in the design of all kinds of libraries (not only combinatorial ones). In 2006, Gorse [8] presents several molecular representations leading to the definition of "chemical space, drug space and activity space". He studies the "strategies for compound selection in such spaces" in term of diversity. In this review, we will describe how to deal with collections of compounds. In the following sections, we first present different types of compound collections and some specialized software to deal with them. The different ways to define the molecules are then compared. Finally, we will discuss methods which enable entire collections and the selection techniques of subsets to be represented.

Large compound collections : definitions and listing

In the fields of chemoinformatics and drug discovery, there is a plethora of designations, types and definitions for compound collections. We will try to give a precise definition for every type of collection and for all the terms used throughout this review. The search for a drug is related to a known target or activity and hence requires biological knowledge. While HTS and VS can produce such knowledge, they must be driven with reduced data sets because of obvious problems of cost and computational time. The scope of this review is upstream to the search for a targeted drug, and upstream to the preparation of small sets of candidate compounds. Large chemical collections can contain both chemical and biological information on compounds. In this paper, we will focus only on the chemical information. One of the most important notions in this field is that of "chemical space". But what is the exact meaning of this expression ? Two partially embedded definitions can be given. In some works, chemical space corresponds to the ensemble of possible chemical products [9]. In 2007, Ogata et al. [10] developed a quantitative approach to estimate the size of such a space. Other studies on the size of various chemical spaces are detailed below. In other work, a chemical space is a multidimensional space in which a frame of references (i.e. a coordinate system) is defined. In this multidimensional space, each compound is represented by a point. Each axis corresponds to a numerical value characterizing the product, called a descriptor. Sometimes this space is named "Multidimensional descriptor space". Each new frame of references defines a new chemical space. Such a definition can be found in [11] and its use is further expanded in section 5. In order to classify the different types of chemical collections, Hann and Oprea [12] describe four types of chemical spaces (using the first definition) :

– Virtual : it groups the ensemble of che-

mical compounds which can be made. The size of the virtual space has been estimated at more than 1060 molecules [13].
- Tangible : it corresponds to molecules which could be easily synthesized or have already been synthesized. Their number has been evaluated between 1020 and 1024 [14].
- Global : it contains all real and pure compounds (synthesized or natural) from organisms worldwide. Here, natural products are only pure molecules and not extracts [15]. It is impossible to know the exact size of this group, because many molecules are confidential.
- Real : one such space is a set of molecules that a company really holds. This set is available for VS and HTS.

Within these groups, one can find different types of descriptions of the products. In this paper, we will distinguish three types of collections : databases with bibliographic data, libraries with structural information, and sub-libraries (i.e. subsets of libraries). The frontiers between these three types are unclear. For example the Cambridge Structural Database (CSD) [16] can be considered either as a database or as a library. Each type is now defined in more details.

Databases

In the literature of the field, "database" is used to define different concepts of collections. In this review, only those compound collections which belong to global chemical space will be named "databases". In the databases one can obtain a reference number, the brut formula, 2D representation, literature links and, in certain cases, some experimental data. They allow the user to search for compounds from formulae or part of formulae, but they can hardly be used to search for several molecules which respond to a given property for example, as the information is not in a convenient format for the analysis. Among the most well-known databases, there are "CAS Registry" (Chemical Abstract Services [17]) which contains 35 million organic and inorganic substances and Beilstein [18] which describes 10 million products associated with more than 320 million of experimental data. The review "Basic overview of chemoinformatics" [19] gives an exhaustive list of available databases.

Libraries

Simple Libraries are often an aggregation of compounds coming from multiple sources such as commercial vendor catalogs or available academic repositories. The difference between libraries and databases lies in the fact that compounds in libraries appear in multiple representations (2D graphics, SMILES [20], fingerprints...). Furthermore, most of the libraries contain a large number of values (descriptors) which characterize each molecule. Several computations and predictions (e.g. energy minimisation from the structure, similarity from the descriptors...) are possible from these data. As many such collections exist, we will only give a few examples. The largest library is "iResearch" from ChemNavigator [21] with more than 15 million unique structures from 179 suppliers. Other libraries that can be mentioned are ZINC [22], ChemDB [23], ChemBank [24], and PubChem [25]. The website of ChemDB offers a comparison [26] between these 4 libraries. There is also the French National Chemical Library [27], ChemIDplus [28], ChemMine [29], and ChEBI [30]. Some libraries, such as NCI [31], contain biological activity information on compounds and others, like Prestwick chemical library[32], have been created from a very specific set of compounds. They can all be used however in the same way as the other libraries, i.e. by using only the chemical information. Finally, there are commercial libraries that can be obtained from suppliers (Table 1) and databases of suppliers catalogs such as ACD [33], CHEMCATS [34], or ChemSource [35]. Each study that is undertaken has a specific

goal and framework. For example, we may want to use a library to explore the most diverse space possible in order to find an innovative compound (diversity search) ; or on the contrary the aim may be to find a compound among a restricted set which presents specific characteristics (similarity search). Consequently, it may be difficult to choose the most appropriate library. Verheij [36] analyzed 45 commercial libraries and revealed that significant differences exist between them such as the diversity of structure type presented by compounds. His analysis provides the initial guidelines for selecting the most suitable libraries for a project which will require the use of a specific tool. Among these libraries, it exist particular libraries : the combinatorial and the focused libraries.

Combinatorial libraries may be virtual or real. Virtual libraries are based on an ensemble of simple molecular fragments and a set of possible reactions. The compounds in these libraries result from a large number of possible and synthesizable combinations of fragments. As the manufacturing of combinatorial libraries is not the topic of this review, the reader is referred for more information to [6,37,38]. Recently Fink and Reymond [39] constructed a very large combinatorial library, GDB. This collection contains 26.4 million molecules enumerating all possible organic compounds up to 11 heavy atoms. All the compounds in GDB "obey Lipinski's bioavailability rule". Real combinatorial libraries derive from combinatorial chemistry ; they are considered as "real libraries". The compounds of focused libraries have specific characteristics in relationship with one or several targets. For example, recent work [40] describes a new technique to design a focused library enriched in bioactivity. The specificities of focused libraries are beyond the scope of this review. As suggested earlier, libraries are often too large to search for a targeted drug. Indeed, HTS cannot test the millions of compounds contained in libraries because of the time and cost involved in these experiments. For virtual screening, the size of libraries (i.e. the number of compounds) and especially the large number of values for each molecule (i.e. the number of descriptors) entail a high computational cost and represent a considerable strain on current resources. It is therefore important to reduce the number of compounds for tests. This is achieved by designing a sub-library.

Sub-Libraries or Core-Libraries

In designing a sub-library, the goal is to prepare a set of compounds ready to screen. This set can come either from one large library or from several available libraries (reference libraries), but must be smaller than its source libraries. Generally, it is desirable that the reduction in size do not significantly lower the chemical diversity. The selection of compounds in sub-libraries should therefore respect a good "representation of chemical spaces" (see section 5). The user has the possibility to specify that the sub-library contain only compounds with specific properties (such as the "drug-like criteria" defined in section 4). The following section explains how to manage reference libraries and how to handle them in order to obtain sub-libraries.

Software

For the creation, management and treatment of libraries, we need computer resources and specific software. Many computational tools have been developed to meet this need. They allow data export/import, search, analysis, visualization, filtering... However, to manage large sets of compounds, it is necessary to code the chemical information. We will first present some codes to identify the molecules and some development environments and then give examples of software and web applications for library management.

Tools for chemical information manipulation

Each molecule must be named uniquely, as this name or code differentiates it from other products. For example SMILES [20] (Simplified Molecular Input Line Entry Specification) and InChI [41] (International Chemical Identifier) characterize a molecule by a text line. A comparison of InChI and SMILES available in [42] shows that InChI is more effective in identifying each molecule uniquely. InChI has more functionalities than the SMILES code (differentiation of tautomers for example). On the other hand, SMILES is more user-friendly. The different computational chemistry tools use many different file formats. OpenBabel [43] can convert more than 70 file formats ! Various tools have been developed to manipulate chemical codes and information. A novel language codes chemical information in a unique format : the Chemical Markup Language (CML) [44]. Multiple software libraries exist to develop chemical applications and software, such as the Chemistry Development Kit (CDK) [45], JOELib [46], PerlMol [47], Molecular Handling Template Library (MHTL) [48], Molecular Query Language (MQL) [49] and VEGA [50]. These libraries are sets of modules coded in a particular language (Java, C++,. . .) which allow the user to develop his own tools. For most users, it is easier to work with an application which possesses a graphical interface and some ready-to-use computational tools. The next paragraph presents some software and web applications.

End user software

In Drug Discovery, software can be open-source or proprietary and distribution can be either free or commercial. The distinction between these categories is often fuzzy (for example the software may be free for academics but commercial for companies). Some works list some open-source libraries and software [51], "free web tools supporting medicinal chemistry" [52] and some "tools for drug discovery" [53]. Here we will mention applications related to the management of chemical libraries. First of all, many "pipelines" have been developed these last years. A pipeline allows an easy use of several separate computation modules which are combined to help chemists in their studies from the dataset to the screening analysis. Table 2 gives a list of such pipelines. A second resource is applications which permit the construction of a database, the computation of physical properties for each compound and a graphical representation of molecules. Some examples are : ChemOffice Ultra [54] with drawing and graphical functionalities, Database-Centric Virtual Chemistry System [55] can merge different sources of data. No statistics on the database content are available. Certain tools are delivered with a data set to which the user can add his own data. Finally, there are software packages that have been specifically developed for the management of chemical libraries. They enable the design of libraries, statistics and their visualization, filtering drug-like compounds and in certain cases the creation of sub-libraries. For example :

- Commercial software : ADEPT (A Daylight Enumeration and Profiling Tool) [56], Cyclops from Novartis [57], Isis/Base and Isis/Host [58], ActivityBase [59], Origin [60].
- Free software : Screening Assistant [61,62], ISIDA [63], ADAAPT [64].
- Each of these has its own specialization.

Descriptors

As mentioned in the previous section, molecules must be characterized by a unique code or name to ensure their unique identification. To compare two or more molecules, these codes are insufficient as they do not encode molecular properties. We could say that the unique code is the name of the molecule and its properties are the definition. So, each compound is defined by a set of "descriptors".

Definitions

Descriptors encode molecular properties by numerical values. There are several thousands of descriptors. The "Handbook of Molecular Descriptors" [65] by Todeschini is a reference in this field. Descriptors can be sorted either according to the dimensionality of the chemical representation of the structure, to the nature of the property that they describe or to their numerical representation. If sorted according to the dimensionality of the molecular representation, descriptors can be separated into three groups :

- 1D : based on the formula, i.e. solely on the existing atoms. This group covers, for example, the molecular weight, the presence or absence of a specific element, or the number of atoms.
- 2D : based on the topology, the types of atoms and the connectivity between them. This covers, for example, the calculation of octanol/water partition coefficient (Clog P), the presence of a given fragment or the numerous topological descriptors.
- 3D : based on stereochemistry and geometry. This concerns, for example, surfaces, conformations, pharmacophore representation, energies of HOMO, LUMO.

The nature of the properties described by the descriptors [66] can be stored into three categories :

- Macroscopic physicochemical properties : Clog P, molecular weight, HOMO ...
- Derived properties : surface distributions of electrostatic potential.
- Abstract concepts : substructural fingerprints.

Descriptors can be represented with bit strings (fingerprints), numerical values (molecular weight,...) or with vectors and matrices. A large number of papers deal with the computation or the prediction of descriptors. For example Liao et al. [67] present methods to compute Clog P, while Gola et al. [68] review methods to predict descriptors related to ADMET properties. Some software has been developed to compute many descriptors in as short a computational time as possible : QSARIS [69], Cerius2 [70], Volsurf [71], and Dragon [72]. A source code library, the Chemical Descriptors Library [73], also exists, to develop several algorithms for the computation of descriptors.

Coding and compression of descriptors

Due to the great number of descriptors, it is impossible to use all of them to characterize a set of molecules. Many studies therefore describe indices which summarize a set of molecular properties under a single value [74-76]. Each index can then be viewed as a single descriptor, thus reducing computational time. Other types of information can be coded by fingerprints. Fingerprints summarize molecular information in a vector of bits : a succession of 1 and 0 codes the presence or the absence of fragments or properties in a molecule. An example of fingerprint design can be found in [77]. Since many possible fragments or properties can be included in fingerprints, their size may become too large for computation. To address this problem, different compression methods can be used. The effectiveness of this compression has been studied by Baldi et al. [78] and the interest of various types of fingerprints discussed in [79,80]. Recently Li et al. [81] studied the classification of drug and nondrug compounds with a novel fingerprint ECFP-4, and compared it with a circular fingerprint MOLPRINT2D [82]. An alternative to fingerprints was proposed by Varneck et al. with the substructural fragments [83]. This code is "a universal language to encode reactions, molecular and supramolecular structures". A new set of molecular descriptors based on Shannon Entropy was proposed by Gregori-Puigjané and Mestres [84] while Zyrianov [85] gave another type of descriptors of molecular shape.

Table 1. List of Library's Suppliers			
Suppliers (49)	Web Address	Date	Number of Compounds
ACB Blocks	http ://www.acbblocks.com	Q2/08	2945
AKOS	http ://www.akosgmbh.de	Q2/08	690018
AnalytiConDiscovery	http ://www.ac-discovery.com	Q2/08	19080
ASDI	http ://www.asdi.net	Q2/08	9069
Asinex	http ://www.asinex.com	Q2/08	360170
Aurora Fine Chemicals	http ://www.aurorafinechemicals.com	Q2/08	>3700000
BioFocus	http ://www.biofocus.com	Q2/05	20280
Bionet	http ://www.keyorganics.ltd.uk	Q2/08	46384
Biotech corp of America	http ://www.biotech-us.com	Q2/08	>120000
Cerep	http ://www.cerep.fr	Q2/08	>16500
ChemBridge	http ://chembridge.com	Q2/08	>435000
ChemDiv	http ://www.chemdiv.com	Q2/08	>500000
ChemTI	http ://www.chemti.com	Q2/08	500000
Chemical block	http ://www.chemical-block.com	Q2/08	3870
Chim. Nat.	http ://chimiotheque-nationale.enscm.fr	Q2/08	36473
Combi-Blocks	http ://www.combi-blocks.com	Q2/08	1064
CombiPure	http ://www.combipure.com	Q3/05	910
Comgenex	http ://www.comgenex.com	Q2/08	>300000
EMC Microcollections	http ://www.microcollections.de	Q2/08	>30000
Enamine	http ://www.enamine.net	Q2/08	542329
Exclusive chemistry	http ://www.exchemistry.com	Q2/08	>2000
FCHC	http ://www.ark.chem.ufl.edu	Q2/08	>30340
Frontier Scientific	http ://www.frontiersci.com	Q2/08	874
Greenpharma	http ://www.greenpharma.com	Q2/08	240
InterBioScreen	http ://www.ibscreen.com	Q2/08	>410000
Labotest	http ://www.labotest.com	Q2/08	>1500000
Life chemicals	http ://www.lifechemicals.com	Q2/08	645000
Matrix Scientific	http ://www.matrixscientific.com	Q2/08	>18500
MayBridge	http ://www.maybridge.com	Q2/08	70400
MDPI	http ://www.mdpi.org	Q2/08	10655
Molecular design discovery	http ://www.worldmolecules.com	Q2/08	33320
Nanosyn	http ://www.nanosyn.com	Q2/08	>40000
NCI	http ://dtp.nci.nih.gov	Q2/08	>40000
Otava	http ://www.otava.com.ua	Q2/08	120000
Peakdale	http ://www.peakdale.com	Q2/08	8548
Pharmeks	http ://www.pharmeks.com	Q2/08	>185000
Prestwick	http ://www.prestwickchemical.com	Q2/08	1120
Princeton biomolecular	http ://www.princetonbio.com	Q2/08	>500000
Pyxis discovery	http ://www.pyxis-discovery.com	Q2/08	2950
SALOR	http ://www.sigmaaldrich.com	Q2/08	>100000
Scientific exchange	http ://www.htscompounds.com	Q2/08	964619
Specs	http ://www.specs.net	Q2/08	240000
Spectrum Info	http ://www.spectrum.kiev.ua	Q2/08	8678
SynChem	http ://www.synchem.com	Q2/08	858
SynphaBase	http ://www.synphabase.com	Q2/08	>350
TimTec	http ://www.timtec.net	Q2/08	>100000
TosLab	http ://www.toslab.com	Q2/08	16419
Tripos	http ://leadquest.tripos.com	Q2/08	73380
VitasM Laboratory	http ://www.vitasmlab.com	Q2/08	>200000

Descriptors selection

However, fingerprints and compression methods are not sufficient to reduce the computational time and the complexity of comparisons between products. To achieve these aims, it is therefore important to select a lower number of descriptors. It has been shown in [86-89] that the quality of models (such as similarity comparison, diversity construction, QSAR studies) depends on the selected descriptors. For example, Charton demonstrated in 2003 [90] that topological parameters (e.g. number of atoms, bonds,...) are not fundamental parameters for the modeling of molecular properties (melting point, critical temperature, solubility of gases...). In the case of pharmacophoric search, in contrast, topological parameters are of prime importance [91]. It is therefore the application that drives the selection of descriptors. Glen and Adams [66] reviewed the methods to make a good choice of descriptors in combination with a similarity method adapted for different types of studies. The importance of selection lies in the fact that, since for a given set of products, many descriptors are highly correlated, their simultaneous use does not provide any additional information. Godden and Bajorath [92] have constructed a scheme to classify descriptors according to their information content and their database dependence. They showed that differences found between databases or libraries in comparison studies depend on the distribution of descriptors within each database. Given the importance of the descriptors selection, other avenues have been explored in Machine Learning. For example Mazzatorta et al. [93] based their study on a genetic algorithm and artificial neural networks, and Frölich et al. [94] on SVM (Support Vector Machine).

Table 2. List of Pipelines

Pipeline	Type of Availability
ABCD [176]	Commercial
Blue Obelisk [177]	Free
Inforsense [178]	Commercial
Knime [179]	Free
Pipeline-Pilot [180]	Commercial
VCCLAB [181]	Free
Web Service Infrastructure [182]	Free

Drug-likeness and similar filters

To become a drug, a chemical compound must have a good biological activity. It also needs certain properties which ensure good bioavailability. This bioavailability is often summarized in the ADMET properties and many criteria, scores and classification have been developed to characterize and evaluate the capacity of a compound to be a good drug. A similar approach is necessary in HTS in order to avoid too many false-positive responses. In the past, tests of the bioavailability of compounds appeared late in the drug discovery process. As a result, many drug candidates were eliminated during the final phases of the process. In order to reduce the costs of drug research, it is essential to eliminate, as soon as possible, the compounds which do not have requisite properties. This elimination can intervene before the design of the sublibraries by means of filters which we will now describe. The compounds which have the right properties to become a drug are called "drug-like" compounds. Various definitions of "drug-like" compounds or "drug-likeness" have been proposed :

- Lipinski [95,96] described drug-like compounds as "those compounds that have sufficiently acceptable ADME properties and sufficiently acceptable toxicity properties to survive through the completion of human Phase I clinical trials".
- In 2002, Walters and Murcko [97] defined drug-like compounds as "molecules which contain functional

groups and/or have physical properties consistent with the majority of known drugs". They reviewed several computational techniques for identifying drug-like molecules.
- In 2003, Muegge [98] wrote that "Drug-likeness is a general descriptor of the potential of a small molecule to become a drug".
- In 2004, Vieth et al. [99] wrote that "drug-like properties are viewed as those that convey desirable pharmacokinetic and pharmacodynamic properties, independent of pharmacological target or indication".

Furthermore, several parameters are used to define the filters which characterize the drug-likeness of a compound. Lipinski [100] states that "the meaning of drug-like is dependent on mode of administration"; his rule (described in detail below) is associated with acceptable aqueous solubility and intestinal permeability for orally administrated compounds. Although Lipinski's rule is the most famous in the area of drug-like selection, it should be applied with caution because of its restriction to a kind of absorption mode. The biological process targeted by a drug is also an important parameter in the drug-like definition. For example, drugs which target the central nervous system (CNS) must exhibit characteristics that enable them to cross the blood-brain barrier (BBB), whereas such properties may lead to a toxic activity in other areas; a topical intestinal antibiotic must present a low intestinal permeability. Muegge [98] warns that drug-likeness is based on "specific" studies, in which drug-like properties are deduced from a small number of databases, often created from commercial drugs. He argues that the filters proposed by these studies are not sufficient, and recommends the use of machine learning methods. He proposes studying drug-likeness from many sets of compounds, not just from one, and reviews computational techniques to address the drug-likeness of compound selections. Recently, Zheng and al. [101] presented a 3D filter using two descriptors : molecular saturation and the proportion of hetero-atoms in a molecule. This filter discriminates a "drug-like" database from a "non drug-like" one, enabling the user to choose a database which could be interesting for a specific screening. In summary, the term drug-likeness includes many properties : oral absorption, aqueous solubility, permeability, blood brain barrier penetration and stability. Given these different points of view and the existing reviews, we will present two main approaches for the selection of drug-like compounds : the use of rules based on filters and some new techniques used in machine learning.

Rules based on filters

These rules are based on the study of compounds' properties. This section describes the most common filters used, such as Lipinski rules and the ADMET properties, and then focuses on recent studies which use physicochemical properties of compounds as filters. Filters are used to select drug-like compounds for developing orally administrated drugs. Charifson et al. in 2002 [102] reviewed filters for the selection of lead-like compounds, which are more restrictive than drug-like filters. Here we will deal only with drug-like compounds.

The most widely-used filter is Lipinski's "rule of five" [100]. This rule contains four parameters concerning the molecular properties that a drug must have :
- Molecular Weight $\leq$ 500 Da
- LogP $\leq$ 5
- Hydrogen bond donors $\leq$ 5
- Hydrogen bond acceptors $\leq$ 10

In his papers, Lipinski [95,96] indicates that "if two parameters are out of range, a poor absorption or permeability is possible". So, at least 3 validated parameters for a compound are necessary for it to have a chance of possessing drug-like properties. In 2002, Veber [103] completed this rule with two new parameters :
- Polar Surface Area $\leq$ 140 $Å^2$

– Rotatable bonds ≤ 10

These rules must not be interpreted as drug/non drug discriminators [104]. They only delimit an area of properties. If compounds are out of this area, they have little chance of becoming good drugs regarding oral administration. Other statistical work showed the importance of molecular properties as drug-like filters :

– 2003 : Bergstrom et al. [105] showed that "molecular surface properties of compounds can predict drug solubility and permeability with sufficient accuracy to allow theoretical absorption classification of drug molecules". This study is based on multivariate data analysis.
– 2004 : Vieth et al. [99] studied the difference between oral and non-oral marketed drugs by statistical analysis to capture the differences between routes of administration. They found several factors influencing oral bioavailability. For example, "oral drugs have fewer H-bonds donors, acceptors and rotatable bonds than drugs with other routes of administration".
– 2008 : Vistoli et al. [106] presented molecular flexibility as an important descriptor to evaluate the drug-likeness of compounds.

The use of these classic filters is so widespread that several easy-to-use commercial programs now exist : Cerius2 [70], Idea [107], Absolv [108], QikProp [109], QMPR-Plus [110], Volsurf [71], OraSpotter [111]. The rules of absorption described by filters are summarized more completely in an ensemble of drug-like properties : the ADMET properties (absorption, distribution, metabolism, excretion and toxicity) of compounds. Gola et al. in 2006 reviewed how to predict such properties [112]. Recently, Gleeson [113] generated a set of simple and interpretable ADMET rules. Nowadays, no drug discovery project can be conducted without an ADMET prediction and an associated selection of compounds. Recent work presents statistical approaches in order to evaluate the bioavailability of compounds :

– 2005 : Martin [114] presented a bioavailability score. She defined F as the permeability and bioavailability properties of compounds. The score "assigns the probability that a compound will have F $>$ 10 % in the rat". For example she showed that 55 % of compounds which pass the rule of five have F $>$ 10%.
– 2006 : Monge et al. [42] implemented a progressive drug-like score. For each criterion (described in previous classic filters) a penalty is computed. This method is more restrictive than classic rules, and selects good drug candidates.
– 2007 : Hutter [115] studied the distribution of atom species and their pairwise combinations in drugs and non-drugs. A drug-likeness score can be derived from the statistical analysis of occurrence probabilities. With this score, drugs can be predicted with an accuracy of at least 71%.

Another type of filters is the reactive functions and warheads described by Rishton [116,117]. These filters are not exactly used for the discrimination of drug-like compounds, but mainly for the selection of compounds for HTS. The reactive functions are parts of ligands allowing covalent bonds with its target or false-positive reactions in HTS. The warheads are molecules which entail false positives in screening because of strong structure-reactivity relationships. Such compounds must be eliminated before screening, because it is difficult to use their results in HTS. Some compounds, the promiscuous aggregating inhibitors, have a non-competitive functionality, a low structure-activity relationship and a bad selectivity. Their mechanism was described by McGovern et al [118] and their "in-silico" identification was presented by Seilder in 2003 [119].

Machine Learning

Finally, methods based on Machine Learning have recently been developed to classify drug-like and non drug-like compounds. These techniques are based on the learning of rules from the observation of known compounds. For example, Support Vector Machine (SVM) was used to classify ADMET properties of compounds [120] and to filter the drug-likeness [81].

In 2003, Byvatov et al. [121] compared the efficiency of the SVM method with another method often used in drug/non-drug classification, the Artificial Neural Network (ANN). They showed that SVM is more robust than ANN, whatever the descriptors and the size of training set. However the difference between the standard error of SVM (20%) and that of ANN (20.75%) was not significant. So in 2005, Müller et al. [122] drastically improved these results, reducing the error of SVM to 7%.

In 2003, Mattioni and al. [123] predicted the genotoxicity of compounds using a genetic algorithm approach.

Truchon and Bayly [124] approached the problem of viable reagents selection in combinatorial libraries with a deterministic algorithm : GLARE.

Amini et al. [125] proposed a novel approach based on the logic in order to predict the toxicology of compounds. Furthermore a website [126] propose several tools for the toxicology prediction. In 2008, Schneider et al. [127] compared decision trees with SVM for the classification of drugs/non-drugs. These two techniques presented similar prediction accuracy. Then the decision trees generated by authors were "used to derive guidelines for the design of drug-like substances".

Finally, there are some hybrid methods which use machine learning and statistics. In 2006, Biswas et al. [128] learned (with ANN) the pattern of the distribution of molecular descriptors which are representative of drug molecules. They deduced from it a drug-likeness score between 0.0 and 1.0, and showed that more than 70% of the drugs considered have a score above 0.5.

Representation of chemical spaces

In the drug discovery process, molecules need to be compared. To do this, one can represent molecules in a multi-dimensional space which enables compounds to be visualized in regard to the others, and similarity or diversity to be materialized in a collection. Raghavendra and Maggiora [129] stress the interest of the representation of chemical spaces as follows : "Chemical spaces provide an intuitive and conceptual basis for understanding many relationships among diverse sets of compounds." Several techniques exist to construct chemical spaces. In "Recent Advances in Chemoinformatics" [130], Agrafiotis et al. describe two ways to represent molecules in spaces :

- "Single molecule coding (coordinate-based space)". Each molecule has precise coordinates in a multidimensional space. These coordinates are deduced according to some known criteria of molecules. Each compound has an absolute position in this space.
- "Pairwise molecule coding (coordinate-free space)". Here, the molecules are characterized by their distances with each other.

Coordinate-based spaces

The most usual way to represent compounds in a space is by using descriptors as coordinates. Many studies have been carried out in this domain due to the extensive knowledge of descriptors. Given the great number of descriptors, however, the dimensionality of the space can become very high. To enhance visualization and understanding by human users, it is therefore necessary to reduce the dimensionality. Maniyar et al. [131] compare the main visualization algorithms : Principal Component Analysis (PCA) [132], Sammon's mapping [133] and Self-Organizing Maps (SOM) [134], Gene-

rative Topographic Mapping (GTM) [135] and their own hierarchical GTM. Their results indicate that Generative Topographic Mapping and Hierarchical GTM seem to be better than the others to cluster active compounds. Several applications implement these different algorithms of visualization and analysis of compound sets :

- Neuroscale [136] is a neuronal network implementation of Sammon's mapping.
- SpotFire, a commercial software [137] used in various fields of science, improves the design of chemical libraries.
- Maniyar et al. [131] present a software which implements the GTM algorithm and the other main visualization algorithms. This software can integrate the output of these algorithms in other commercial software such as Pipeline Pilot.
- In 2001, Oprea et al. [138] developed ChemGPS (Chemical Global Positioning System) which positions novel structures in drug space via PCA-score prediction. It provides a new mapping device for the drug-like chemical space.

Thanks to such graphical techniques the structural diversity of compound libraries can be easily visualized. Indeed, in a reduced space in 2 or 3 dimensions, each molecule is represented by a dot and positioned in this space. The coverage of the space and its homogeneity provide information about the diversity of the plotted library. For example, if a library covers a large area in the space, and if this coverage is homogeneous, then the majority of molecular properties are represented by compounds in this library ; it can be described as "diverse". On the other hand, if only a small part of the space is covered then only some properties are represented by compounds, corresponding to a specific library. As an example, one can mention the database created by Fink and Reymond [39] which contains "All molecules of up to 11 atoms of C, N, O, and F possible under consideration of simple valency, chemical stability, and synthetic feasibility rules". In their paper some projections of this set of compounds on a 2D space clearly show a good coverage of this space. The drawback with the reduction of dimensionality is that it results in a selection of a small number of descriptors or of new composite variables to represent the molecules. We saw previously that the selection of descriptors cannot be ideal. A reduction can never represent correctly all the information embedded in a set of compounds. Other techniques have therefore been developed, called coordinate-free space methods. Instead of using absolute coordinates, they represent the molecules by the distances between them.

Coordinate-free spaces

First of all, in 2006, Godden and Bajorath [139] developed a distance function : the "activity-centered" function, used in a high-dimensional space based on transformed descriptors. This function can detect molecular similarity relationships in high-dimensional chemical spaces. In this approach, descriptors can be used to represent chemical space, but a reduction of dimensionality is not necessary. In 2007, Raghavendra and Maggiora [129] described a new method based on generalized Fourier analysis. It uses "the concept of molecular basis sets to represent chemical space within an abstract vector space". They associated molecular similarities between pairs of molecules with their corresponding inner products of vectors. The method is independent of descriptors and uses another way to compare molecules. Chemical spaces display the similarity between compounds and the diversity of collections. While enabling visualization by library users, each set of descriptors and each computational method give a specific chemical space. When the libraries are compared, the results in a given chemical space can be different from those obtained in another one. The final section of this review therefore addresses the issue of the computation of similarity and diversity and

introduces techniques to guide the selection of a diverse sub-library from one or several compound collections.

Diversity, Similarity and Selection

The notions of similarity and diversity are among the most important in drug discovery. For example, such notions are used in virtual screening to search for compounds which have similar properties to a hit. Similarity relies on a principle which says "two compounds with similar structures have similar biological activities"[140] . This assumption and the efficiency of selection by diversity have been extensively discussed [141-143]. Similarity is a criterion to compare molecules by establishing a distance between them. Many metrics exist to characterize similarity or dissimilarity and hence diversity. Depending on the goal of the screening, two different types of compound collections are required : In hits search, it is necessary to have a sub-library that is as diverse as possible, as this increases the possibility of finding innovative compounds and returns a set of hits that are structurally diverse and cover a large area in the chemical space. In searching for leads from hits, the molecules must be close to hits and so must be similar to them. We first present several of the metrics used to compute similarity, and then discuss the methods used to design diverse sets of compounds. Note that in a recent paper, Maldonado [144] reviews similarity and diversity computations.

Metrics

Most of the time, similarity and diversity of compounds are evaluated as distances or coefficients. The use of these coefficients is described in [144] and their relative performance for similarity searches are compared in [145]. Among the most widely-used, one can find the Tanimoto coefficient, the Cosine coefficient, Squared Euclidean, Variances... They compute the difference in one or several molecular properties (descriptors) between two compared compounds. A more mathematical and general approach is given by Tang et al. [146] who studied six diversity metrics. Recently novel methods have appeared for the computation of these metrics :

Similarity metrics

In 2006, Fitzgerald et al. [147] introduced a method to compare libraries, based on similarity. This method is scaffold-oriented (in this work, the scaffolds are defined as "a significant component of the contact surface") and enables the user to choose a library among others. This choice is guided by the user's goal (for example : the search for particular structures).

Batista et al. [148] proposed a method to evaluate molecular similarity : MolBlaster. This technique depends on the use of "fragment profiles" of molecules. It presents the advantage of being independent of molecular descriptors. Indeed, we saw in sections 3 and 5 that a selection of descriptors is never ideal.

In 2007, Rupp et al. [149] used annotated graphs to measure molecular similarity. Molecules are represented by a graph in which nodes are atoms and edges are bonds. Usually, the graphs are compressed on fingerprints which are compared to evaluate the similarity between corresponding molecules. Since compression entails a loss of information, the authors propose to solve this problem by a direct comparison of graphs. Their results show that this method seems to be more efficient than other techniques based on descriptors.

Swamidass et al. [150] pointed out that similarity measures based on the comparison of fingerprints contain a systematic error. "As the length of the fingerprints decreases, their typical density and overlap tend to increase, and so does any similarity measure based on overlap, such as the widely used Tanimoto similarity." To address this problem, the authors present a mathematical correction.

Gleb [151] used a two-dimensional representation of molecular similarity to predict similar biological activities. The results validate the similarity principle [140].

Brewer [152] presented a method to determine the degree of intermolecular similarity by spectral clustering. This method allows a good classification of molecules in different clusters.

Finally in 2008, Theertham et al. [153] report novel algorithms which can reflect similarity or diversity and apparent bioactivity in a compound collection. These algorithms are based on a selection of descriptors and a divergence score. The results show that these metrics present a reasonable bioactive enrichment and outperform other techniques based on the variance criterion.

Wang and Bajorath [154] discuss the molecular complexity which "biases the evaluation of fingerprint similarity". They propose "a variant of the Tversky coefficient" to modulate this molecular complexity effect. They study the optimization of hit rates for a search of active molecules with varying complexity.

These metrics are used with algorithms to perform a similarity search for compounds, a topic dealt with in several recent studies [155-158]. Indeed, the evaluation of similarity can facilitate the analysis of virtual screening results, by allowing for example the molecules that are obtained to be grouped in different families.

Diversity metrics

In 2006, Papp et al. [159] provided a new technique to evaluate the diversity of compound collections, the Explicit Diversity Index (EDI). This index combines structural and synthesis-related dissimilarity values in a single value, enabling the rapid computation of diversity.

Schuffenhauer et al. [160] linked structural diversity with complexity and biological activity of compounds. They used feature counts as a complexity measure and found that this value increases with the increase in ligands activity. They address the problem of complexity by a diversity selection based on sphere-exclusion and the Tanimoto coefficient.

In 2007, Rabal et al. [161] improved a cell-based method which presented some drawbacks. "Cell-based methods divide the descriptor space into hypercubic cells or bins and assign each molecule to the cell that matches the set of binned properties of that molecule". One of its defects is the edge effects (cluster artifacts) produced by the arbitrary cell boundaries. Indeed, molecules near this frontier but in two different cells have a higher diversity score than molecules in the same cell but less similar. The authors presented a "diversity metric based on quantifying the distance to the center of the bins resulting from partitioning the descriptor space". This method solves the problem of edge effects.

Diversity metrics can be combined with a Machine Learning method in order to select a diverse or a representative set of compounds. This technique intervenes upstream of virtual screening. Indeed, in the drug discovery process, it can be useful to search for molecules in a structurally diverse area to find compounds with different action modes or with different structures. Diversity and representativity are two different concepts [162]. Representativity is related to the selection of compounds from a "reference" library, it is a sampling which respects the distribution of this reference library. The representativity is a relative notion, while diversity is a much more absolute notion with comparison of a frame of references of descriptors - a diverse set may (and ought to) come from many libraries. When the compounds in a sub-library are as representative as possible of the entire chemical space, this sub-library can be described as duverse. The concluding paragraph presents methods which can be used to select representative compounds, and techniques to design sublibraries with respect to diversity.

Selection of compounds

Maldonado et al. [144] cited the main Machine Learning methods used in representativity/diversity selection. They classed them into two parts :

- Classification methods : groups of compounds are separated into classes in which the compounds have a high degree of similarity. Compounds can be classified by :
  - Supervised methods : Neural Networks, Maximum Likelihood Classification, Support Vector Machines and k-Nearest Neighbors classifiers.
  - Unsupervised methods : Clustering, Partitioning, SOM and PCA.
- Selection methods : Cluster-based selection, partition-based selection, dissimilarity-based selection, optimization-based selection and Genetic Algorithms.

We will focus on recent advances in this domain.

Selection for representative sub-library construction

In 2006, Landon and Shaus [163] presented a "joint entropy diversity analysis" program, JEDA. They used a set of chemical descriptors to partition the chemical space of a combinatorial library. A subset of compounds is selected via a Shannon-entropy based scoring function. The originality of the method lies in this function because, usually, compounds are selected in each partition of the space. Furthermore JEDA determines a representative subset of compounds from a library with the possibility for the user to define a size for this subset. The results of this method are equivalent, in terms of diversity of the subset, to those of other partition-based methods, and provide an enhanced representativity of the reference library.

Selzer et al. [164] used a combination of an Artificial Neural Network with "radial distribution function molecular descriptors" to extract representative subsets from combinatorial libraries. The neural network is constructed with the same number of neurons as the number of expected compounds in the final subset. At the end of the learning process, the compound centroïds of each neuron are selected and become the compounds of the representative subset. This method provides "quick and intuitive feedback about the results of the cheminformatics experiment".

In 2008, Xie and Chen [165] proposed a strategy to select a diverse-structure and representative subset from NIH PubChem. This subset "was selected by whole-molecule chemistry-space matrix calculation using the cell-based partition algorithm".

Selection for diverse sub-library construction

In 2005, Muresan and Sadowski [166] developed a scoring scheme for the acquisition of compounds. They combined Neural Network and Ghose-Crippen fingerprints [167] to create this scheme, which can discriminate between data sets (with overlap) in order to select compounds to design a new diverse library.

In 2006, Li [168] studied the redundancy of large libraries with a novel fast clustering method. The motivation for the study was that "Many available clustering methods focus on accurate classification of compounds ; they are slow and are not suitable for very large compound libraries". He developed an incremental clustering algorithm which is faster than classical methods.

Engels et al. [169] developed a novel method to characterize the overlap between two compound collections. They used "a novel hierarchical clustering algorithm called divisive k-means". The diversity of different libraries is then evaluated by means of their clustering. This method can detect duplicates and can be used to study the diversity of many libraries.

Skold et al. [170] applied these methods and presented a structurally diverse and commercially available drug data set which can be used for benchmarking studies. They

used multivariate data analysis to construct this data set.

In 2007, Maldonado et al. [171] presented a molecular diversity analysis tool, MolDiA. This tool compares several diverse data sets. It is based on fingerprints which contain structural information and physicochemical properties. The intermolecular comparison relies on Ullman's algorithm [172] which is optimized with a system of chemical graph analysis. Furthermore, this tool is flexible in the choice of similarity/diversity measures, in the fuzziness of comparison... It allows "the user to adapt the tool to particular needs".

To conclude, many methods can be used to design a diverse or representative sublibrary. Some studies have investigated the relevance of such classification and selection methods with the biological aim of drug discovery processes. One of these studies [173] showed that "no classification method is overall superior to all other studied methods, but there is a general trend that rule-based, scaffold-oriented methods are the better choice if classes with homogeneous biological activity are required". "Clustering based on chemical fingerprints is superior if fewer and larger classes are required, and some loss of homogeneity in biological activity can be accepted". Finally, one can cite Yeap et al. [174] : " The choice of compound file, rational subset method, and ratio of the subset size to the compound file size are key factors in the relative performance of random and rational selection, and statistical simulation is a viable way to identify the selection approach appropriate for a subset".

Conclusion

In this paper we have presented different steps to prepare the collections of compounds for Virtual Screening or High Throughput Screening. Each year, more and more scientific papers deal with one of these steps. The increase in computing power and in the flow of the HTS systems causes an increase in the size of the libraries used. The methods must therefore progress. Automatic learning approaches, used in other fields of science, may be adapted to our field. These developments ensure that the domain of chemoinformatics will be very active in coming years.

Bibliography

[1] Agrafiotis, D.K.; Myslik, J.C.; Salemme, F.R. Mol. Divers., 1998, 4, 1-22.
[2] Bures, M.G.; Martin, Y.C. Curr. Opin. Chem. Biol., 1998, 2, 376-380.
[3] Spellmeyer, D.C.; Grootenhuis, P.D.J. Annu. Rep. Med. Chem., 1999, 34, 287-296.
[4] Drewry, D.H.; Young, S.S. Chemomet. Intell. Lab. Sys., 1999, 48, 1-20.
[5] Leach, A.R.; Hann, M.M. Drug Discov. Today, 2000, 5, 326-336.
[6] Lutz, W. QSAR Comb. Sci., 2005, 24, 809-823.
[7] Lumley, J.A. QSAR Comb. Sci., 2005, 24, 1066-1075.
[8] Gorse, A.D. Curr. Topics Med. Chem., 2006, 6, 3-18.
[9] Lipinski, C.A.; Hopkins, A. Nature, 2004, 432, 855-861.
[10] Ogata, K.; Isomura, T.; Yamashita, H.; Kubodera, H. QSAR Comb. Sci., 2007, 26, 596-607.
[11] Dobson, C.M. Nature, 2004, 432, 824-828.
[12] Hann, M.M.; Oprea, T.I. Curr. Opin. Chem. Biol., 2004, 8, 255-263.
[13] Bohacek, R.; Martin, C.; Guida, W. Med. Res. Rev., 1996, 16, 3-50.
[14] Ertl, P. J. Chem. Inf. Comput. Sci., 2003, 43, 374-380.
[15] Scior, T.; Bernard, P.; Medina-Franco, J.L.; Maggiora, G.M. Mini Rev. Med. Chem., 2007, 7, 851-860.
[16] http ://www.ccdc.cam.ac.uk/products/csd/
[17] http ://www.cas.org/casdb.html
[18] http ://www.Beilstein.com/
[19] Engel, T. J. Chem. Inf. Model., 2006, 46, 2267-2277.
[20] Weininger, D. J. Chem. Inf. Comput. Sci., 1988, 28, 31-36.

[21] http ://www.chemnavigator.com/
[22] Irwin, J.J. ; Shoichet, B.K. J. Chem. Inf. Model., 2005, 45, 177-182.
[23] Chen, J. ; Swamidass, S.J. ; Dou, Y. ; Bruand, J. ; Baldi, P. Bioinformatics, 2005, 21, 4133-4139.
[24] Seiler, K.P. ; George, G.A. ; Happ, M.P. ; Bodycombe, N.E. ; Carrinski, H.A. ; Norton, S. ; Brudz, S. ; Sullivan, J.P. ; Muhlich, J. ; Serrano, M. ; Ferraiolo, P. ; Tolliday, N.J. ; Schreiber, S.L. ; Clemons, P.A. Nucl. Acids Res., 2007, 1-9.
[25] Austin, C.P. ; Brady, L.S. ; Insel, T.R. ; Collins, F.S. Science, 2004, 306, 1138.
[26] http ://cdb.ics.uci.edu/CHEM/Web/ cgibin/supplement/Comparison.py
[27] http ://chimiotheque-nationale.enscm.fr/
[28] http ://chem.sis.nlm.nih.gov/chemidplus/
[29] Girke, T. ; Cheng, L.C. ; Raikhel, N. Plant Physiol., 2005, 138, 573.
[30] Brooksbank, C. ; Cameron, G. ; Thornton, J. Nucl. Acids Res., 2005, 33.
[31] Tetko, I.V. Mini Rev. Med. Chem., 2003, 3, 809.
[32] http ://www.prestwickchemical.fr/ index.php ?pa=26
[33] http ://www.daylight.com/products/ acd.html
[34] http ://www.cas.org/expertise/ cascontent/chemcats.html
[35] http ://www.chemsources.com/ chemonline.html
[36] Verheij, H. Mol. Divers., 2006, 10, 377-388.
[37] Agrafiotis, D.K. ; Martin, E.J. J. Mol. Graph. Model., 2000, 18.
[38] Sharma, P. ; Salapaka, S. ; Beck, C. J. Chem. Inf. Model., 2008, 48, 27-41.
[39] Fink, T. ; Reymond, J. L. J. Chem. Inf. Model. 2007, 47, 342-353.
[40] Lars Arve, T.V.H. W. QSAR Comb. Sci., 2006, 25, 449-456.
[41] http ://iupac.org/projects/2004/ 2004-039-1-800.html
[42] Monge, A. ; Arrault, A. ; Marot, C. ; Morin-Allory, L. Mol. Divers., 2006, 10, 389-403.
[43] http ://openbabel.sourceforge.net
[44] Holliday, G.L. ; Murray-Rust, P. ; Rzepa, H.S. J. Chem. Inf. Model., 2006, 46, 145-157.
[45] Steinbeck, C. ; Han, Y. ; Kuhn, S. ; Horlacher, O. ; Luttmann, E. ; Willighagen, E. J. Chem. Inf. Comput. Sci., 2003, 43, 493-500.
[46] http ://www-ra.informatik.uni-tuebingen.de/ software/joelib/
[47] http ://www.perlmol.org
[48] Zhang, W. ; Hou, T. ; Qiao, X. ; Xu, X. J. Chem. Inf. Comput. Sci., 2004, 44, 1571-1575.
[49] Proschak, E. ; Wegner, J.K. ; Schuller, A. ; Schneider, G. ; Fechner, U. J. Chem. Inf. Model., 2007, 47, 295-301.
[50] Pedretti, A. ; Villa, L. ; Vistoli, G. J. Comp.-Aid. Mol. Des., 2004, 18, 167-173.
[51] Geldenhuys, W.J. ; Gaasch, K.E. ; Waston, M. ; Allen, D. ; Van der Schyf, C.J. Drug Discov. Today, 2006, 11, 127-132.
[52] Ertl, P. ; Jelfs, S. Curr. Topics in Med. Chem., 2007, 7, 1491-1501.
[53] Villoutreix, B.O. ; Renault, N. ; Lagorce, D. ; Sperandio, O. ; Montes, M. ; Miteva, M. A. Curr. Protein Peptide Sci., 2007, 8, 381-411.
[54] Smith C. Nature, 2007, 446, 223-224.
[55] Buntrock, R.E. J. Chem. Inf. Comput. Sci., 2002, 42, 1505-1506.
[56] Lind, P. ; Alm, M. J. Chem. Inf. Model., 2006, 46, 1034-1039.
[57] Leach, A.R. ; Al, A. J. Chem. Inf. Comput. Sci., 1999, 39, 1161-1172.
[58] Gobbi, A. ; Al, A. Perspect. Drug Des. Discov., 1997, 131-58.
[59] www.mdl.com
[60] www.id-bs.com/activitybase/
[61] Edwards, P.M. J. Chem. Inf. Comput. Sci., 2002, 42, 1270-1271.
[62] http ://screenassistant.sourceforge.net/
[63] Monge, A. http ://tel.archives-ouvertes.fr/ tel-00122995/fr/
[64] http ://infochem.u-strasbg.fr/recherche/ ISIDA/
[65] Cho, S. ; Sun, Y. ; Harte, W. J. Comp.-Aid. Mol. Des., 2006, 20, 249-261.
[66] Todeschini, R. ; Consonni, V. Handbook of Molecular Descriptors ; Wiley-VCH : Weinheim, 2000.

[67] Robert, C. ; Glen-Samuel, E.A. QSAR Comb. Sci., 2006, 25, 1133-1142.
[68] Liao, Q. ; Yao, J. ; Yuan, S. Mol. Divers., 2006, 10, 301-309.
[69] Joelle Gola, O.O. E.C.M.S. QSAR Comb. Sci., 2006, 25, 1172-1180.
[70] QSARIS http ://www.scivision.com/ QSARIS.html
[71] Cerius2 http ://www.accelrys.com/ products/cerius2/
[72] Volsurf http ://www.moldiscovery.com/ soft_volsurf.php
[73] DRAGON http ://www.talete.mi.it/ products/dragon_description.htm
[74] Library, C.D. http ://sourceforge.net/ projects/cdelib
[75] Devillers, J. ; Balaban, A.T. Topological Indices and Related Descriptors in QSAR and QSPR ; Gordon and Breach : The Netherlands, 1999.
[76] Bonchev, D. Information Theoritic Indices for Characterization of Chemical Structures ; Wiley _ Research Studies Press : New York, 1983.
[77] Balaban, A.T. ; Beteringhe, A. ; Constantinescu, T. ; Filip, P.A. ; Ivanciuc, O. J. Chem. Inf. Model., 2007, 47, 716-731.
[78] Haigh, J.A. ; Pickup, B.T. ; Grant, J.A. ; Nicholls, A. J. Chem. Inf. Model., 2005, 45, 673-684.
[79] Baldi, P. ; Benz, R.W. ; Hirschberg, D.S. ; Swamidass, S. J. J. Chem. Inf. Model., 2007, 47, 2098-2109.
[80] Uli Fechner, J.P.G.S. QSAR Comb. Sci., 2005, 24, 961-967.
[81] Godden, J.W. ; Stahura, F.L. ; Bajorath, J. J. Chem. Inf. Model., 2005, 45, 1812-1819.
[82] Li, Q. ; Bender, A. ; Pei, J. ; Lai, L. J. Chem. Inf. Model., 2007, 47, 1776-1786.
[83] MOLPRINT2D
http ://www.molprint.com/
[84] Varnek, A. ; Fourches, D. ; Hoonakker, F. ; Solov'ev, V. J. Comp. Aid. Mol. Des., 2005, 19, 693-703.
[85] Gregori-Puigjane, E. ; Mestres, J. J. Chem. Inf. Model., 2006, 46, 1615-1622.
[86] Zyrianov, Y. J. Chem. Inf. Model., 2005, 45, 657-672.
[87] Blum, A.L. ; Langley, P. Artif. Intellig., 1997, 97, 245-271.
[88] Guyon, I. ; Elisseeff, A. J. Machine Learn. Res., 2003, 3, 1157-1182.
[89] John, G. ; Kohavi, R. ; Pleger, K. Machine Learning : Proc. of the 11th Intern. Conf., 1994, 121-129.
[90] Nikolova, N. ; Jaworska, J. QSAR Comb. Sci., 2003, 22, 9-10.
[91] Charton, M. J. Comp.-Aid. Mol. Des., 2003, 17, 197-209.
[92] Hert, J. ; Willett, P. ; Wilton, D. ; Acklin, P. ; Azzaoui, K. ; Jacoby, E. ; Schuffenhauer, A. Organ. Biomol. Chem., 2004, 2, 3256-3266.
[93] Godden, J.W. ; Bajorath, J. QSAR Comb. Sci., 2003, 22, 487-497.
[94] Mazzatorta, P. ; Vracko, M. ; Benfenati, E. J. Comp.-Aid. Mol. Des., 2003, 17, 335-346.
[95] Holger Fröhlich, Jörg, K.W.A.Z. QSAR Comb. Sci., 2004, 23, 311-318.
[96] Lipinski, C.A. J. Pharmacol. Toxicol. Methods, 2000, 44, 235-249.
[97] Lipinski, C.A. ; Lombardo, F. ; Dominy, B.W. ; Feeney, P.J. Adv. Drug Deliv. Rev., 1997, 23, 3-25.
[98] Walters, W.P. ; Murcko, M.A. Adv. Drug Deliv. Rev., 2002, 54, 255-271.
[99] Muegge, I. Med. Res. Rev., 2003, 23, 302-321.
[100] Vieth, M. ; Siegel, M.G. ; Higgs, R.E. ; Watson, I.A. ; Robertson, D.H. ; Savin, K.A. ; Durst, G.L. ; Hipskind, P.A. J. Med. Chem., 2004, 47, 224-232.
[101] Lipinski, C.A. Drug Discov. Today : Technol., 2004, 1, 337-341.
[102] Zheng, S. ; Luo, X. ; Chen, G. ; Zhu, W. ; Shen, J. ; Chen, K. ; Jiang, H. J. Chem. Inf. Model., 2005, 45, 856-862.
[103] Charifson, P. ; Walters, W. J. Comput. Aided Mol. Des., 2002, 16, 311-323.
[104] Veber, D.F. ; Johnson, S.R. ; Cheng, H.Y. ; Smith, B.R. ; Ward, K.W. ; Kopple, K.D. J. Med. Chem., 2002, 45, 2615-2623.
[105] Frimurer, T.M. ; Bywater, R. ; Narum, L. ; Lauritsen, L.N. ; Brunak, S. J. Chem. Inf. Comput. Sci., 2000, 40, 1315-1324.
[106] Bergstrom, C.A.S. ; Strafford, M. ; Lazorova, L. ; Avdeef, A. ; Luthman, K. ; Ar-

tursson, P. J. Med. Chem., 2003, 46, 558-570.
[107] Vistoli, G.; Pedretti, A.; Testa, B. Drug Discov. Today, 2008, 13, 285-294.
[108] iDEA pk Express; Lion Bioscience AG : Heidelberg, Germany.
[109] Absolv; Sirius Analytical Instrumetns Ltd : East Sussex UK.
[110] QikProp; Schrödinger, Inc.
[111] QMPRPLUS; Simulations plus, Inc : Lancaster, CA 93534-2902.
[112] Oraspotter; zyxbio : Cleveland, OH.
[113] Gola, J.; Obrezanova, O.; Champness, E.; Segall, M. QSAR Comb. Sci., 2006, 25, 1172-1180.
[114] Gleeson, M.P. J. Med. Chem., 2008, 51, 817-834.
[115] Martin, Y.C. J. Med. Chem., 2005, 48, 3164-3170.
[116] Hutter, M.C. J. Chem. Inf. Model., 2007, 47, 186-194.
[117] Rishton, G.M. Drug Discov. Today, 2003, 8, 86-96.
[118] Rishton, G. M. Drug Discov. Today, 1997, 2, 382-384.
[119] McGovern, S.L.; Caselli, E.; Grigorieff, N.; Shoichet, B.K. J. Med. Chem., 2002, 45, 1712-1722.
[120] Seidler, J.; McGovern, S.L.; Doman, T.N.; Shoichet, B. K. J. Med. Chem., 2003, 46, 4477-4486.
[121] Matthew W.B. Trotter, Sean B.H. QSAR Comb. Sci., 2003, 22, 533-548.
[122] Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. J. Chem. Inf. Comput. Sci., 2003, 43, 1882-1889.
[123] Muller, K.R.; Ratsch, G.; Sonnenburg, S.; Mika, S.; Grimm, M.; Heinrich, N. J. Chem. Inf. Model., 2005, 45, 249-253.
[124] Mattioni, B.E.; Kauffman, G.W.; Jurs, P.C.; Custer, L.L.; Durham, S.K.; Pearl, G.M. J. Chem. Inf. Comput. Sci., 2003, 43, 949-963.
[125] Truchon, J.F.; Bayly, C.I. J. Chem. Inf. Model., 2006, 46, 1536-1548.
[126] Amini, A.; Muggleton, S.H.; Lodhi, H.; Sternberg, M.J. J. Chem. Inf. Model., 2007, 47, 998-1006.
[127] http ://www.predictive-toxicology.org/
[128] Schneider, N.; Jackels, C.; Andres, C.; Hutter, M.C. J. Chem. Inf. Model., 2008, 48, 613-628.
[129] Biswas, D.; Roy, S.; Sen, S. J. Chem. Inf. Model., 2006, 46, 1394-1401.
[130] Raghavendra, A.S.; Maggiora, G.M. J. Chem. Inf. Model., 2007, 47, 1328-1340.
[131] Agrafiotis, D.K.; Bandyopadhyay, D.; Wegner, J.K.; vanVlijmen, H. J. Chem. Inf. Model., 2007, 47, 1279-1293.
[132] Maniyar, D.M.; Nabney, I.T.; Williams, B.S.; Sewing, A. J. Chem. Inf. Model., 2006, 46, 1806-1818.
[133] Bishop, C.M. Neural Networks for Pattern Recognition, 1st Ed.; Oxford University Press : New York, 1995.
[134] Sammon, J.W. IEEE Trans. Comput., 1969, C-18, 401-409.
[135] Kohonen, T. Self-Organizing Maps; Springer-Verlag : Berlin, 1995.
[136] Bishop, C.M.; Svensen, M.; Williams, C.K.I. Neural Comput., 1998, 10, 215-234.
[137] Lowe, D.; Tipping, M.E. AdV. Neural Inf. Proc. Syst., 1997, 9, 543-549.
[138] Spotfire www.spotfire.com
[139] Oprea, T.I.; Gottfries, J. J. Comb. Chem., 2001, 3, 157-166.
[140] Godden, J.W.; Bajorath, J. J. Chem. Inf. Model., 2006, 46, 1094-1097.
[141] Maggiora, G.M.; Johnson, M.A. Concepts and Applications of Molecular Similarity; John Wiley Sons : New York, 1990.
[142] Spencer, R.W. J. Biomol. Screen., 1997, 2, 69-70.
[143] Potter, T.; Matter, H. J. Med. Chem., 1998, 41, 478-488.
[144] Martin, Y.C.; Kofron, J.L.; Traphagen, L.M. J. Med. Chem., 2002, 45, 4350-4358.
[145] Maldonado, A.; Doucet, J.; Petitjean, M.; Fan, B.-T. Mol. Divers., 2006, 10, 39-79.
[146] Haranczyk, M.; Holliday, J. J. Chem. Inf. Model., 2008, 48, 498-508.
[147] Tang, E.; Suganthan, P.; Yao, X. Machine Learn., 2006, 65, 247-271.
[148] Fitzgerald, S.H.; Sabat, M.; Geysen, H.M. J. Chem. Inf. Model., 2006, 46, 1588-1597.
[149] Batista, J.; Godden, J.W.; Bajorath,

J. J. Chem. Inf. Model., 2006, 46, 1937-1944.
[150] Rupp, M.; Proschak, E.; Schneider, G. J. Chem. Inf. Model., 2007, 47, 2280-2286.
[151] Swamidass, S.J.; Baldi, P. J. Chem. Inf. Model., 2007, 47, 952-964.
[152] Gleb, D.P. QSAR Comb. Sci., 2007, 26, 346-351.
[153] Brewer, M. L. J. Chem. Inf. Model., 2007, 47, 1727-1733.
[154] Theertham, B.; Wang, J.L.; Fang, J.; Lushington, G.H. Curr. Comput. Aided Drug Des., 2008, 4, 23-24.
[155] Wang, Y.; Bajorath, J. J. Chem. Inf. Model., 2008, 48, 75-84.
[156] Rhodes, N.; Clark, D.E.; Willett, P. J. Chem. Inf. Model., 2006, 46, 615-619.
[157] Grant, J.A.; Haigh, J.A.; Pickup, B.T.; Nicholls, A.; Sayle, R.A. J. Chem. Inf. Model., 2006, 46, 1912-1918.
[158] Melville, J.L.; Riley, J.F.; Hirst, J.D. J. Chem. Inf. Model., 2007, 47, 25-33.
[159] Vogt, M.; Godden, J.W.; Bajorath, J. J. Chem. Inf. Model., 2007, 47, 39-46.
[160] Papp, A.; Gulyas-Forro, A.; Gulyas, Z.; Dorman, G.; Urge, L.; Darvas, F. J. Chem. Inf. Model., 2006, 46, 1898-1904.
[161] Schuffenhauer, A.; Brown, N.; Selzer, P.; Ertl, P.; Jacoby, E. J. Chem. Inf. Model., 2006, 46, 525-535.
[162] Rabal, O.; Pascual, R.; Borrell, J.I.; Teixido, J. J. Chem. Inf. Model., 2007, 47, 1886-1896.
[163] Bayada, D.M.; Hamersma, H.; Van Geerestein, V.J. J. Chem. Inf. Comput. Sci., 1999, 39, 1-10.
[164] Landon, M.; Schaus, S. Mol. Divers., 2006, 10, 333-339.
[165] Selzer, P.; Ertl, P. J. Chem. Inf. Model., 2006, 46, 2319-2323.
[166] Xie, X. Q.; Chen, J. Z. J. Chem. Inf. Model., 2008, 48, 465-475.
[167] Muresan, S.; Sadowski, J. J. Chem. Inf. Model., 2005, 45, 888-893.
[168] Viswanadhan, V.N.; Ghose, A.K.; Revankar, G.R.; Robins, R.K. J. Chem. Inf. Comput. Sci., 1989, 29, 163-172.
[169] Li, W. J. Chem. Inf. Model., 2006, 46, 1919-1923.
[170] Engels, M.F.M.; Gibbs, A.C.; Jaeger, E.P.; Verbinnen, D.; Lobanov, V.S.; Agrafiotis, D.K. J. Chem. Inf. Model., 2006, 46, 2651-2660.
[171] Skold, C.; Winiwarter, S.; Wernevik, J.; Bergstrom, F.; Engstrom, L.; Allen, R.; Box, K.; Comer, J.; Mole, J.; Hallberg, A.; Lennernas, H.; Lundstedt, T.; Ungell, A.L.; Karlen, A. J. Med. Chem., 2006, 49, 6660-6671.
[172] Maldonado, A.G.; Doucet, J.P.; Petitjean, M.; Fan, B.T. J. Chem. Inf. Model., 2007, 47, 2197-2207.
[173] Sethi, R.; Ullman, J.D. J. ACM, 1970, 17, 715-728.
[174] Schuffenhauer, A.; Brown, N.; Ertl, P.; Jenkins, J.L.; Selzer, P.; Hamon, J. J. Chem. Inf. Model., 2007, 47, 325-336.
[175] Yeap, S.K.; Walley, R.J.; Snarey, M.; vanHoorn, W.P.; Mason, J.S., J. Chem. Inf. Model., 2007, 47, 2149-2158.
[176] Agrafiotis, D.K.; Alex, S.; Dai, H.; Derkinderen, A.; Farnum, M.; Gates, P.; Izrailev, S.; Jaeger, E.P.; Konstant, P.; Leung, A.; Lobanov, V.S.; Marichal, P.; Martin, D.; Rassokhin, D.N.; Shemanarev, M.; Skalkin, A.; Stong, J.; Tabruyn, T.; Vermeiren, M.; Wan, J.; Xu, X.Y.; Yao, X. J. Chem. Inf. Model., 2007, 47, 1999-2014.
[177] Guha, R.; Howard, M.T.; Hutchison, G.R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E.L. J. Chem. Inf. Model., 2006, 46, 991-998.
[178] http ://www.inforsense.com/products/inforsense_platform/
[179] http ://www.knime.org/
[180] Hassan, M.; Brown, R.; Varma-O'Brien, S.; Rogers, D. Mol. Divers., 2006, 10, 283-299.
[181] Tetko, I.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V.; Radchenko, E.; Zefirov, N.; Makarenko, A.; Tanchuk, V.; Prokopenko, V. J. Comp.-Aid. Mol. Des., 2005, 19, 453-463.
[182] Dong, X.; Gilbert, K.E.; Guha, R.; Heiland, R.; Kim, J.; Pierce, M.E.; Fox, G.C.; Wild, D.J. J. Chem. Inf. Model., 2007, 47, 1303-1307.

Chapitre 2

Présentation des données

Dans ce chapitre nous présentons les données utilisées pour l'étude de la sélection par diversité : leur origine, les traitements appliqués, leur description. Puis nous présentons la création des jeux de tests.

2.1 Origine des données

Nous souhaitons que nos jeux de tests soient représentatifs de l'espace chimique réel qui correspond aux produits réellement disponibles selon Hann et Oprea [4]. Pour cela nous avons établi une base (à l'aide de Screening Assistant que nous présentons dans la section suivante) contenant un maximum de composés commerciaux et académiques proposés par un grand nombre de fournisseurs. Les chimiothèques collectées sont principalement constituées de produits synthétisés à l'aide de la chimie classique et de la chimie combinatoire. Il en existe quelques unes constituées de produits d'origine naturelle, notamment celle de GreenPharma qui contient exclusivement des composés naturels.

La base ainsi constituée contient plus de 6 millions d'enregistrements provenant de 53 fournisseurs différents qui correspondent en réalité à 4 237 949 composés uniques (nous définissons dans la section suivante ce que nous considérons comme composé unique). La liste des fournisseurs et leurs coordonnées est donnée dans les tableaux 2.1 et 2.2. Les dates exprimées dans ce tableau sont sous la forme Trimestre/Année (Q4/09 correspond par exemple au trimestre 4 de l'année 2009) et indique la dernière mise à jour disponible des données collectées avant la création des jeux de test. Ensuite le tableau donne le nombre de composés mais également le pourcentage de composés drug-like [1] et le pourcentage de composés exclusifs [2] de chaque chimiothèque. Les proportions et l'apport de chaque fournisseur sont mis en lumière par le graphique 2.1. Pour une meilleure lisibilité, les trois plus grosses bases ont été représentées séparément dans le graphique 2.2. Parmi les chimiothèques académiques utilisées, on retrouve celle du laboratoire (ICOA) et la chimiothèque nationale qui réunit une partie des composés académiques provenant des laboratoires français. Ensuite certains fournisseurs importants comme Biotrend, Princeton et le NCI ne sont pas présents dans cette collection. Cette absence est dûe à des erreurs survenant lors de la construction de la base avec Screening Assistant. Ces erreurs ont été corrigées par la suite et la base totale a été enrichie plus récemment ; elle contient près de 17 millions d'enregistrements pour plus de 6 millions de composés uniques provenant de 73 fournisseurs différents. Nos jeux de test ont été construits avant ces modifications.

1. Un composé est considéré comme drug-like lorsque son score PDL [79] est inférieur ou égal à 1.
2. Un composé est dit exclusif lorsqu'il n'est présent que dans la chimiothèque concernée.

Enfin on remarque que les bases comportant le plus de composés sont celles d'Enamine avec plus d'un million de produits, de Chemdiv et VitasLab qui comportent toutes deux plus de 600 000 produits. Cependant ces bases ne sont pas forcément celles qui donnent le plus de molécules exclusives. Enamine en comporte certes plus de 90% mais les deux autres bases ne donnent respectivement que 59.95% et 12.79% de molécules exclusives. Ceci s'explique par le fait que certains fournisseurs proposent des compilations de plusieurs bases existantes. Pour évaluer le critère drug-like des structures, on utilise le score établi par Monge et al. [79]. On considère alors les composés comme drug-like lorsque leur score PDL est inférieur ou égal à 1. Le pourcentage de composés drug-like est supérieur à 80% pour toutes les chimiothèques quelque soit leur origine. Les bases de ACB Blocks, Azasynth et SynChem sont entièrement drug-like. Le pourcentage de composés exclusifs peut aller de 4.98% à 100%. Seul Azasynth propose une base totalement exclusive.

Fournisseurs	Adresse internet	Date	Composés	% Drug-like	% Exclusifs
ACBBlocks	http ://www.acbblocks.com	Q4/09	3033	100	50.61
AF ChemPharm	http ://www.afchempharm.co.uk	Q1/10	635	92.91	77.48
Alinda Chemical Ltd	http ://www.alinda.ru/about_en.html	Q1/10	254111	97.93	5.85
AMRI	http ://www.comgenex.com	Q1/10	255176	91.12	99.06
Analyticon	http ://www.ac-discovery.com	Q4/09	26499	83.04	98.78
Apollo Scientific	http ://www.apolloscientific.co.uk	Q1/10	39630	96.2	41.76
Aronis	http ://www.aronis.ru	Q1/10	22866	97.9	10.53
ARVI	http ://www.ar-vi.com	Q1/10	9987	97.58	10.99
Asinex	http ://www.asinex.com	Q4/09	389797	96.47	49.91
Azasynth	http ://www.azasynth.com	Q1/10	67	100	100
Bionet	http ://www.keyorganics.ltd.uk	Q4/09	42601	98.69	84.04
Chembridge	http ://chembridge.com	Q4/09	421722	99.26	25.94
ChemDiv	http ://www.chemdiv.com	Q4/09	663476	96.08	59.95
Chemivate	http ://www.chemivate.com	Q1/10	1108	99.64	99.37
ChemTI	http ://www.chemti.com	Q4/09	170884	97.24	13.43
Chess	http ://www.chess-chem.com	Q1/10	562	99.47	4.98
Chimiotheque Nationale	http ://chimiotheque-nationale.enscm.fr	Q4/09	25396	93.6	85.84
EMC	http ://www.microcollections.de	Q4/09	27042	93.06	99.12
Enamine	http ://www.enamine.net	Q4/09	1350064	98.29	90.88
Endeavour	http ://www.endeavourchem.co.uk	Q1/10	808	99.01	38.99
Endotherm	http ://www.endotherm-lsm.com	Q1/10	675	90.37	80.15
Exclusive Chemistry	http ://www.exchemistry.com	Q4/09	2272	98.5	72.1

TABLE 2.1: Liste des fournisseurs utilisés pour la construction de la base (première partie)

Fournisseurs	Adresse internet	Date	Composés	% Drug-like	% Exclusifs
FCHC	http ://ark.chem.ufl.edu	Q4/09	30896	92.81	77.33
FluoroChem	http ://www.fluorochem.net	Q1/10	28157	98.34	32.76
FocusSynthesis	http ://www.focussynthesis.com/	Q1/10	2367	97.93	63.67
Frontier Scientific	http ://www.frontiersci.com	Q4/09	1709	94.27	39.03
GreenPharma	http ://www.greenpharma.com	Q1/10	651	83.87	51.31
ICOA		Q1/10	4927	88.8	70.59
Infarmatik	http ://www.infarmatik.com	Q4/09	1369	99.93	56.9
InterBioScreen	http ://www.ibscreen.com	Q4/09	455282	96.53	20.99
KaironKem	http ://www.kaironkem.com	Q1/10	651	99.54	51.31
Labotest	http ://www.labotest.com	Q4/09	105937	95.23	48.67
LifeChemicals	http ://www.lifechemicals.com	Q4/09	316135	98.23	60.87
Matrix Scientific	http ://www.matrixscientific.com	Q4/09	38397	98.67	30.91
Maybridge	http ://www.maybridge.com	Q4/09	56764	98.14	76.68
MDPI	http ://www.mdpi.org	Q4/09	10188	91.7	76.24
Nanosyn	http ://www.nanosyn.com	Q4/09	65118	95.75	24.69
Otava	http ://www.otavachemicals.com	Q4/09	120455	98.3	29.62
Peakdale	http ://www.peakdale.com	Q4/09	14619	99.23	98.56
Pharmeks	http ://www.pharmeks.com	Q4/09	227718	93.11	12.64
Prestwick	http ://www.prestwickchemical.com	Q4/09	1111	86.86	31.86
Scientific Exchange	http ://www.htscompounds.com	Q4/09	47218	97.2	23.37
Sequoia	http ://www.seqchem.com	Q1/10	2192	77.78	44.34
Specs	http ://www.specs.net	Q4/09	209538	96.07	27.74
Spectrum	http ://www.spectrum.kiev.ua	Q4/09	8675	96.41	9.41
SynChem	http ://www.synchem.com	Q4/09	1556	100	41.65
SynphaBase	http ://www.synphabase.com	Q4/09	485	88.45	74.64
SynthonLab	http ://www.synthon-lab.com	Q1/10	47250	97.39	20.76
Szintekon	http ://szintekon.hu	Q1/10	2676	99.14	68.68
TimeTec	http ://www.timtec.net	Q4/09	178163	96.07	9.28
TosLab	http ://www.toslab.com	Q4/09	16578	89.61	53.52
VitasLab	http ://www.vitasmlab.com	Q4/09	770109	96.08	12.79

TABLE 2.2: Liste des fournisseurs utilisés pour la construction de la base (dernière partie)

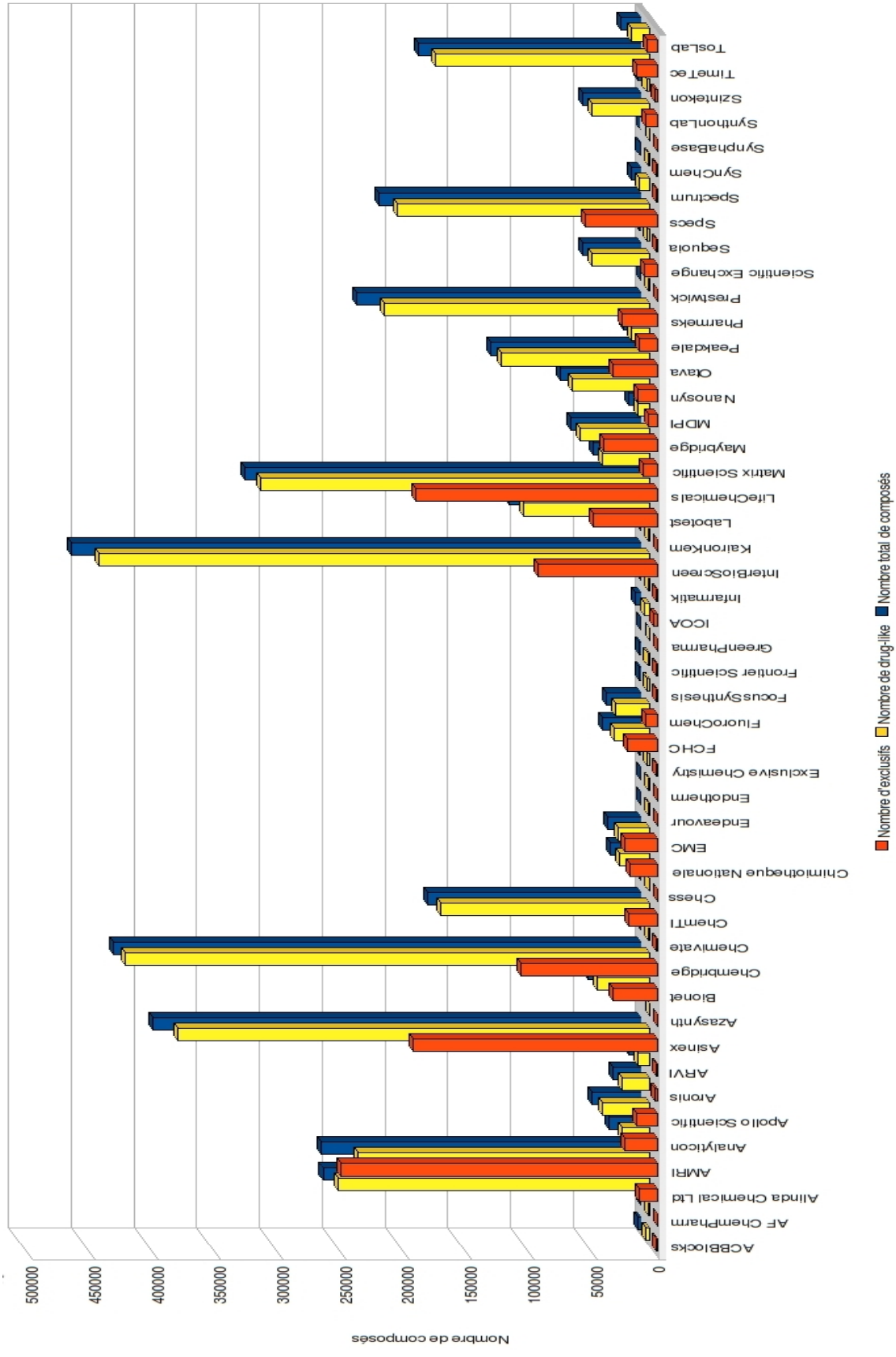


FIGURE 2.1: Nombre de composés total, composés exclusifs et composés drug-like par chimiothèque

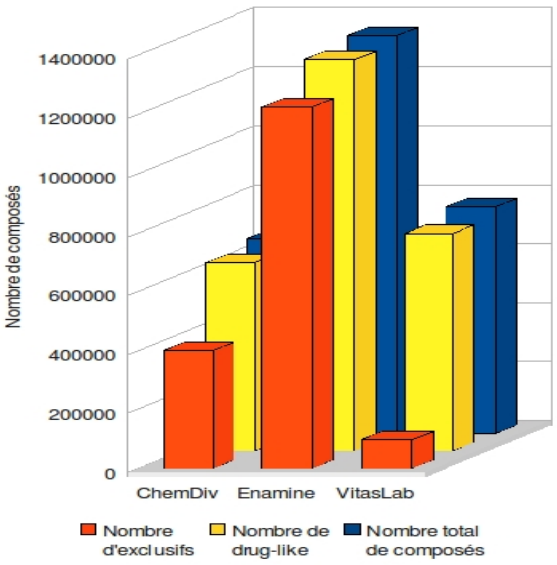


FIGURE 2.2: Nombre de composés total, composés exclusifs et composés drug-like pour les 3 plus grandes chimiothèques

2.2 Traitement de la collection

Le logiciel Screening Assistant (développé par l'équipe de chémoinformatique) a été utilisé pour réunir les chimiothèques en une seule base et pour appliquer des filtres sur les molécules ainsi que traiter les doublons. Screening Assistant (SA) utilise une plateforme Java et la librairie JOELib pour les opérations sur les structures chimiques interfacées avec une base de données relationnelle en SQL. Depuis la création de nos jeux de tests, le logiciel a été modifié, ici nous ne faisons état que des étapes existantes lors de la construction de la base d'où ont été extraits les jeux tests.

Tout d'abord lorsque les fichiers de structures provenant des fournisseurs sont lus par SA, les contre-ions [3] sont supprimés. De ce fait deux structures identiques avec des contre-ions différents seront considérées comme doublon. Ces structures sont ensuite protonées au pH physiologique [4] puis ajoutées à la base. Pour cela, chaque molécule est caractérisée par un code InChI. Pour savoir si une structure existe déjà dans la base, le hashcode InChI de cette molécule est comparé au hashcode des molécules déjà présentes dans la base. Si deux structures ont le même hashcode, alors leur InChI complet est comparé. On détermine ainsi les doublons et on obtient une base composée de structures uniques (celle-ci pouvant appartenir à plusieurs fournisseurs référencés pour chacune).

3. Dans les chimiothèques réelles certains produits comme les acides carboxyliques ou les bases ne sont pas sous forme moléculaire mais ionisés c'est à dire chargés électriquement. Pour assurer la neutralité électrique globale ils sont associés à un contre ion (un ion de charge électrique opposé). Cet ion, généralement de masse inférieure à la molécule, doit être éliminé lors du traitement chémoinformatique.

4. Le pH du milieu biologique, 7,41, est tel que dans l'organisme certaines molécules ne restent pas sous forme moléculaire mais vont s'ioniser, c'est à dire se charger électriquement en gagnant ou perdant un atome d'hydrogène. Prendre en compte dans un modèle chémoinformatique une telle molécule nécessite de modifier sa structure dans la base de données en fonction de ce phénomène.

Enfin les composés contenant des atomes exotiques n'ont pas été pris en compte dans la sélection final des jeux tests.

2.3 Description des molécules

Les descripteurs déjà présents dans la première version de Screening Assistant (27 au total, dont par exemple des compteurs d'atomes, le score drug-like, le score lead-like, le poids moléculaire, le logP...) ont été calculés pour toutes les molécules ainsi que les descripteurs 2D proposés par MOE. Parmi ces descripteurs, certains sont des binaires (cf. Tableau 2.3), d'autres des variables réelles discrètres (dont 26 compteurs de types d'atomes ou groupes d'atomes, 8 compteurs de types de liaisons, 2 compteurs de cycles et 7 compteurs concernant des règles drug-like) et enfin la majorité sont des variables réelles continues (143 variables au total). Les descripteurs à variance nulle et les variables redondantes de façon certaine (correlation à 1) ont été supprimées. Etant donné que les distances utilisées pour comparer les molécules sont faites soit pour des variables quantitatives soit pour des variables qualitatives, nous avons décidé de supprimer de cet ensemble les descripteurs binaires. Nous possédons donc 186 variables pour décrire les molécules. Leur liste complète est disponible en Annexe A et B.

Nom Descripteurs binaires	Description
lip_druglike	1 si la molécule est drug-like selon les règles de Lipinski, 0 sinon
opr_leadlike	1 si la molécule est lead-like selon les règles d'Oprea, 0 sinon
reactive	1 si la molécule est réactive, 0 sinon
rsynth	
SA_Ispeptide	1 si la molécule est un peptide, 0 sinon
SA_PerfluorinatedChain	1 si la molécule contient une chaîne perfluorée, 0 sinon
SA_Iswarhead	1 si la molécule est un warhead, 0 sinon
SA_SingleChain	1 si la molécule n'a qu'une seule chaîne, 0 sinon

Table 2.3: Liste des descripteurs binaires

Enfin plusieurs types d'informations sont codés par ces descripteurs, nous les classons en trois groupes annotant chaque descripteur dans les tableaux cités plus haut :

- informations sur les propriétés physico-chimiques
- informations topologiques
- informations de constitution de la molécule.

Nous possédons donc 92 descripteurs physico-chimiques (les descripteurs de charge ont été inclus dans cette catégorie), 44 descripteurs constitutifs (majoritairement des compteurs d'atomes et de liaisons) et 50 descripteurs topologiques.

2.4 Jeux tests

Pour effectuer notre analyse des méthodes de sélection par diversité, nous avons construit 6 jeux différents : 3 jeux de 40 000 molécules et 3 jeux issus de ces derniers dont on a extrait les outliers. Pour cela, 3 tirages aléatoires sans remise ont été effectués dans la base totale des 4 millions de composés pour obtenir les jeux distincts J1, J3 et J5. Ensuite les singletons (molécules seules dans leur groupe) détectés par les méthodes de sélection par diversité (cf. Chapitre 4) sont considérés comme outliers et retirés des jeux J1, J3 et J5

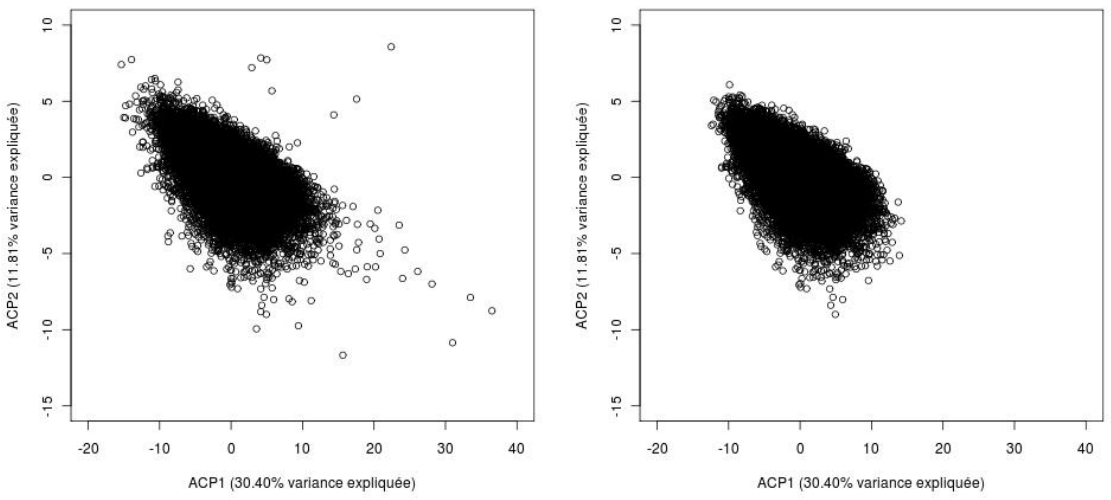
Jeux	Nombre de Composés	% Drug-like
J1	40 000	90.19
J2	39286	90.75
J3	40 000	90. 78
J4	39293	91.35
J5	40 000	90.48
J6	39279	90.92

TABLE 2.4: Description des jeux de test

pour obtenir respectivement les jeux J2, J4 et J6. Nous testerons ainsi la robustesse des méthodes de sélection. En effet nous pourrons vérifier si les résultats des méthodes suivent les mêmes tendances quelque soit le jeu de données et qu'il comporte ou non des outliers.

Le tableau 2.4 donne le nombre de composés par jeu et le pourcentage de structures drug-like pour chacun des jeux. Comme on peut s'y attendre le retrait des outliers donne des jeux (J2, J4 et J6) avec légèrement plus de produits drug-like que les jeux avec outliers. Cependant cette différence est faible et il subsite encore des produits non drug-like dans les jeux sans outliers.

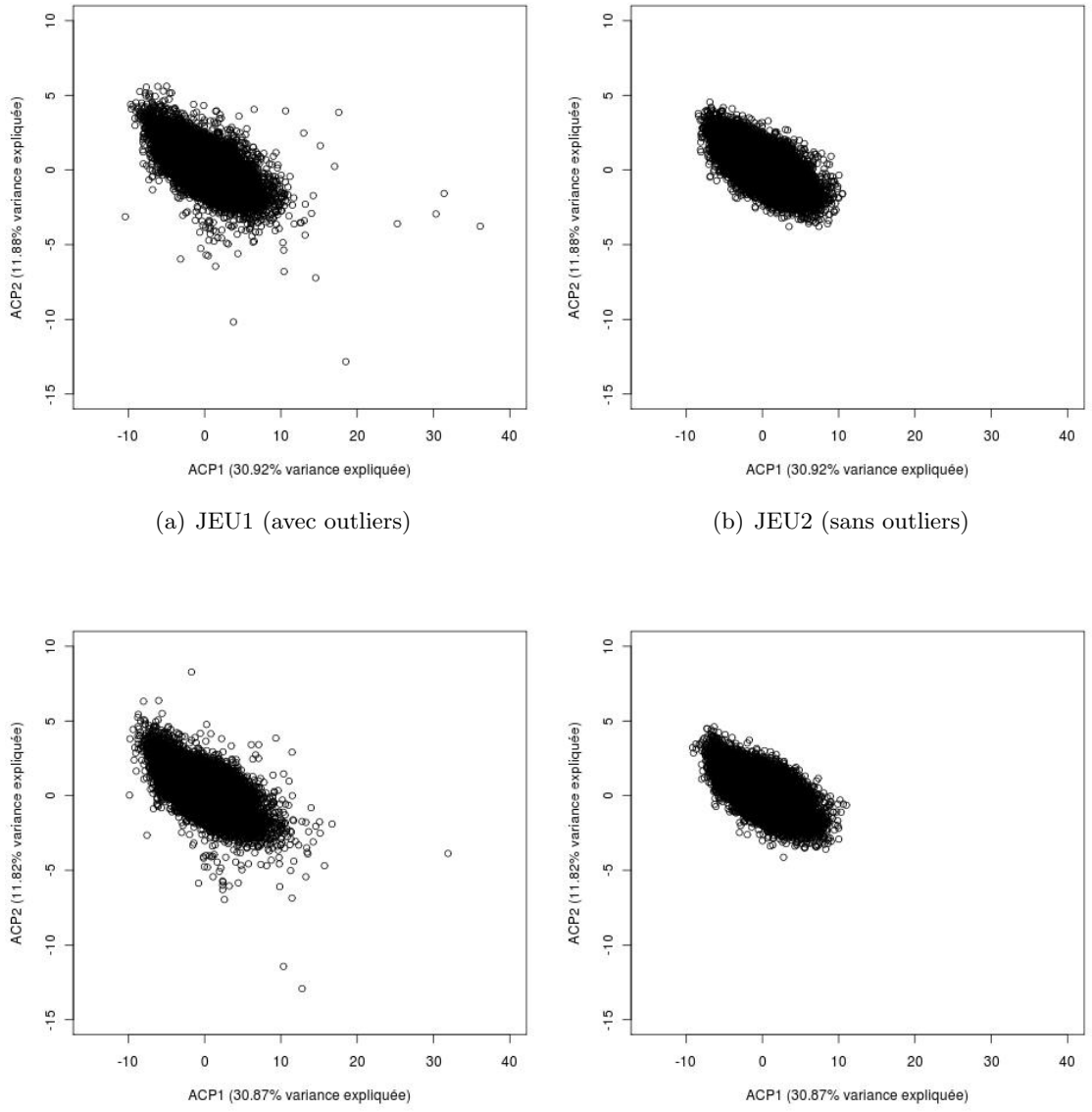
Ensuite, afin de déterminer si nos jeux de données sons homogènement répartis dans l'espace des descriptions, nous pouvons les projeter dans un espace à deux dimensions.Pour cela une ACP a été réalisée sur les jeux J1, J3 et J5. Nous avons ainsi obtenu trois espaces de projection en deux dimensions différents. Les jeux J2, J4 et J6 sont respectivement projetés dans les espaces obtenus pour J1, J3 et J5 (cf. Figure 2.4 et Figure 2.3). Le pourcentage de variabilité des jeux de départs expliqué par les deux premières composantes des 3 ACP équivaut à un peu plus de 40%. Nous pouvons constater pour chaque jeu avec outliers (J1, J3 et J5) qu'il existe une zone dans l'espace très densément peuplée et des zones peu ou pas couvertes par les molécules. Malgré une perte d'informations due à la réduction de dimensions, lorsque deux molécules sont éloignées dans l'espace réduit, elles ont de fortes chances de l'être également dans l'espace complet. Nous pouvons donc dire que les jeux avec outliers sont dispersés de façon non homogène dans l'espace des descriptions. Nous verrons par la suite (cf. Chapitre 4), que cette information est utile pour l'interprétation des résultats de sélection par diversité. Enfin il semble que le retrait d'outliers que nous avons choisi ait l'effet escompté puisque dans cette projection, les jeux sans outliers sont plus concentrés dans l'espace que les jeux avec outliers.



(a) JEU5 (avec outliers)

(b) JEU6 (sans outliers)

FIGURE 2.3: Projection des jeux 5 et 6 dans un espace obtenu par ACP sur J5



(a) JEU1 (avec outliers)

(b) JEU2 (sans outliers)

(c) JEU3 (avec outliers)

(d) JEU4 (sans outliers)

FIGURE 2.4: Projection des jeux 1 et 2 et 3 et 4 dans des espaces obtenus respectivement par ACP sur J1 et sur J3

Chapitre 3

Sélection par diversité : Formalisation

3.1 Notre objectif

On possède un ensemble $\mathcal{M}$ de n molécules (objets, observations, individus, points) : $\mathcal{M}=\{m_i\}_{i=1...n}$. On définit l'espace des descriptions comme l'ensemble des descriptions des molécules que l'on peut faire avec des variables (ou attributs). En chemoinformatique l'espace des descriptions est l'espace multi-dimensionnel engendré par les descripteurs. Celui-ci contient l'ensemble des descripteurs moléculaires (physico-chimiques, topologiques, spaciaux...) qui sont des variables décrivant la constitution, la conformation et les propriétés de chacun de nos objets.

L'espace des descripteurs est donc défini par un ensemble $\mathcal{V}$ de p variables : $\mathcal{V}=\{v_i\}_{i=1...p}$ tel que $v_j(m_i)$ désigne la valeur de la molécule $m_i \in \mathcal{M}$ pour la variable $v_j \in \mathcal{V}$. Chaque objet m_i est donc un vecteur soit de réels et d'entiers soit de binaires à p dimensions (selon le type de descripteurs utilisés, cf. chapitre sur la description des données).

On souhaite trouver une sous-population $C^* \subset \mathcal{M}$ constituée de k objets : $C^* = \{c_l^*\}_{l=1...k}$ (où chaque objet est toujours un vecteur de réels ou binaires à p dimensions), reflétant la diversité de $\mathcal{M}$.

Considérant l'ensemble $\mathcal{M}$ de molécules et l'espace de descriptions qui les caractérise, on peut définir un ensemble de molécules comme divers s'il correspond à un ensemble de représentants de toutes les descriptions possibles dans cet espace. Extraire un sous-ensemble $C^* \subset \mathcal{M}$ de $\mathcal{M}$ divers revient donc à trouver un bon représentant pour chaque molécule de $\mathcal{M}$.

La difficulté réside alors dans le choix de ces représentants. Souhaitant obtenir une sous-population de taille k, sélectionner de bons représentants de l'ensemble $\mathcal{M}$ revient à partitionner l'espace des descriptions de $\mathcal{M}$ de manière homogène en k parties. Puis un représentant est sélectionné dans chacune d'elle. On souhaite que la partition soit un ensemble de parts C_i tels qu'il existe un rayon ϵ tel que l'ensemble $\mathcal{M}$ soit inclus dans l'union des sphères de centre c_i et de rayon ϵ. L'intersection des pshères doit être vide. On minimise alors le rayon de telle sorte que chaque centre soit un bon représentant des molécules incluses dans la sphère concernée. C'est à dire que les molécules contenues dans une sphère doivent toutes être le plus similaire possible à leur centre.

3.1.1 Définition formelle de la sélection par diversité

Rappelons que l'on définit une partition comme un ensemble de parties C_l de $\mathcal{M}$:

Définition 3.1 *Une partition* $\mathcal{C} = \{C_l\}_{l=1...k}$ *de* $\mathcal{M}$ *est telle que :*
$C_l \subset \mathcal{M} \ \forall \ C_l \in \mathcal{C}$
$\bigcup_{l=1..k} C_l = \mathcal{M}$.
et $C_l \neq \emptyset$

C_l est appelé cluster ou sous-groupe. On rappelle (cf. définition 1.6.2.1) qu'à chaque groupe on peut associer un représentant qui peut être un centroïde ou un médoïde. L'ensemble des représentants de la partition $C^* = \{c_l^*\}_{l=1...k}$ (avec $C^* \subset \mathcal{M}$) forme notre sous-ensemble. Considérant la partition précédemment définie, on peut alors dire que le représentant d'un cluster C_l sera l'objet $c_l^* \in C^*$.
Nous notons l'ensemble des objets d'un cluster ainsi : $C_l = \{m_i^l\}_{i=1...|C_l|}$ et la distance calculée entre l'objet m_i^l et l'objet m_j^l : $d(m_i^l, m_j^l)$.

Définition 3.2 *Le* ***rayon relatif*** *d'un cluster noté* $\rho(C_l|m_j^l)$*, est le rayon de la plus petite sphère centrée sur* $m_j^l \in C_l$ *contenant tous les objets du cluster* C_l *:*

$$\rho(C_l|m_j^l) = \underset{i=1...|C_l|}{Max} d(m_i^l, m_j^l)$$

Définition 3.3 *Le* ***rayon absolu d'un cluster*** *noté* $\rho(C_l)$ *est le rayon de la plus petite sphère centrée sur un objet de* C_l *contenant tous les objets du cluster* C_l *:*

$$\rho(C_l) = \underset{j=1...|C_l|}{Min} \rho(C_l|m_j^l)$$

soit

$$\rho(C_l) = \underset{j=1...|C_l|}{Min} \underset{i=1...|C_l|}{Max} d(m_i^l, m_j^l)$$

Le point c_j^l choisi comme représentant du cluster C_l est celui qui minimise $\rho(C_l)$ soit

$$c_j^l = \underset{j=1...|C_l|}{ArgMin} \underset{i=1...|C_l|}{Max} d(m_i^l, m_j^l)$$

Si plusieurs points réalisent ce minimum, on en choisit un au hasard.

Définition 3.4 *Le* ***rayon absolu de la partition*** $\mathcal{C}$ *est alors le rayon de la plus grande sphère de cette partition :*

$$\rho(\mathcal{C}) = \underset{l=1...k}{Max} \rho(C_l)$$

Enfin comme nous l'avons dit précédemment, chaque partition de k groupes peut être associée à un ensemble de k points représentatifs des groupes. On définit alors la partition associée à cet ensemble de représentants noté $C^* = \{c_l^*\}_{l=1...k}$.

Définition 3.5 *La* ***partition associée à un ensemble de k points*** $\mathcal{C} = \{C_l\}_{l=1...k}$ *est définie par :*

$$C_l = \{m_i | d(m_i, c_l^*) \leq d(m_i, c_j^*) \ \forall \ i \neq l \ et \ \forall \ j \neq l\}$$

En cas d'égalité (c'est à dire s'il existe deux affectations possibles), on prend l'un des deux centres au hasard.

Sélection d'un sous-ensemble divers Sachant calculer le rayon absolu d'une partition, sélectionner un sous-ensemble divers de taille k revient donc à trouver l'ensemble $C^* = \{c_l^*\}_{l=1...k}$ satisfaisant l'argument suivant :

$$\underset{\mathcal{C}\ partition\ de\ \mathcal{M}}{ArgMin}\ \rho(\mathcal{C})$$

$$\text{autrement dit :}\ \underset{\mathcal{C}\ partition\ de\ \mathcal{M}}{ArgMin}\ \underset{l=1...k}{Max}\ \underset{j=1...|C_l|}{Min}\ \rho(C_l|m_j^l)$$

Où $\mathcal{C}$ est la partition associée à l'ensemble C^*. Nous rappelons que ArgMin (argument minimum) d'une fonction (ici $\rho(\mathcal{C})$) représente la valeur de la variable pour laquelle la fonction concernée atteint son minimum.
Il s'agit donc de trouver le sous-ensemble de molécules qui minimise le rayon maximum obtenu pour la partition correspondante. Ainsi en minimisant le rayon de la plus grande sphère, on espère tendre vers un ensemble de sphères de taille (rayon) homogène. On espère ainsi que les représentants de ces sphères soient homogènement répartis dans l'espace de descriptions de $\mathcal{M}$ et ainsi que notre sous-ensemble soit représentatif de la diversité de $\mathcal{M}$.

Nous avons trouvé plusieurs problèmes proches de celui qu'on souhaite résoudre :
- les techniques de clustering par partitionnement
- les techniques de location/allocation

Parmi les problèmes de clustering par partitionnement il existe k-means et k-medoïds qui est une variante de k-means. Nous avons également trouvé k-median (également appelé p-median) et k-center (aussi noté p-center) qui sont des problèmes dit location/allocation optimisant deux critères différents (respectivement MinSum et MinMax). Ces deux derniers ne sont pas des problèmes de clustering mais permettent de trouver une partition d'un ensemble d'objets via le jeu de représentants qu'ils fournissent. Enfin le problème de clustering k-means peut être vu comme un critère à optimiser qui permet d'obtenir un ensemble de groupes optimaux. On retrouve ainsi pour les 3 problèmes (k-means, k-median, k-center) des critères différents à optimiser que nous présentons par la suite.

3.2 Les différentes approches liées à notre objectif

Nous définirons rapidement la méthode du k-means qui est la plus connue. Puis nous nous attacherons plus précisément à détailler les méthodes k-median et k-center. Nous verrons ainsi, au fil de nos explications, les méthodes que nous avons retenues pour une comparaison de leurs résultats sur une sélection par diversité.

3.2.1 *k*-means et *k*-medoïds (M3)

Ces deux méthodes sont basées sur le même critère et le même processus, elles ne diffèrent que sur le choix des centres de groupes.
Soit un ensemble de n objets $\mathcal{M}=\{m_i\}_{i=1...n}$. La méthode de clustering k-means [125] consiste à créer une partition $\mathcal{C}$ de $\mathcal{M}$ en k groupes $\mathcal{C}=\{C_j\}_{j=1...k}$ (dans notre cas, k est la taille de l'échantillon souhaitée).
Rappelons que pour chaque cluster C_l un centroïde m^* est calculé ainsi :

$$\forall j = 1...p,\quad v_j(m^*) = \frac{1}{|C_l|}\sum_{m_i \in C_l} v_j(m_i)$$

Il arrive souvent que m^* soit différent de m_i, ce qui signifie que les centroïdes m^* peuvent être virtuels, c'est-à-dire qu'ils appartiennent à l'espace des descriptions mais ne sont pas inclus dans l'ensemble $\mathcal{M}$.

On note $C^* = \{c_l^*\}_{l=1...k}$ l'ensemble des centroïdes des k groupes de la partition (dans notre cas cet ensemble sera l'échantillon divers souhaité à l'arrêt de l'algorithme). Puis chaque individu de $\mathcal{M}$ est affecté à son centroïde le plus proche selon une mesure de similarité choisie. Cet algorithme (cf. algorithme 1) itère ces étapes (calcul des centres/affectation) jusqu'à convergence (les centres sont les mêmes d'une itération à l'autre).
Pour cette méthode le critère optimisé est la variance intra-classes ou critère des moindres carrés que l'on cherche à minimiser comme suit :

$$\underset{C^* \subset \mathcal{M}}{ArgMin} \sum_{l=1...k} \sum_{i=1...|C_l|} d(m_i^l, c_l^*)^2$$

Algorithm 1 La méthode k-means

input k = la taille de la partition
Generate C^{*0} le premier ensemble de centroïdes
$i \leftarrow 0$
while !(*stopping-criterion*) **do**
  produce $\mathcal{C}^{\,i}$ en affectant chaque observation à un centroïde
  produce C^{*i+1} l'ensemble i+1 de centroïdes en recalculant les centroïdes des clusters de la partition $\mathcal{C}^{\,i}$
  $i \leftarrow i + 1$
end while
return C^*(l'ensemble des centroïdes finaux), $\mathcal{C}$ (l'ensemble des clusters)

La méthode k-means est adaptée pour des données sur lesquelles on peut calculer une distance euclidienne. Pour d'autres types de données (binaires ou nominales pour lesquelles la distance euclidienne n'est pas la plus adaptée) ou si l'on souhaite travailler uniquement avec des objets existants, il existe la variante k-medoïds. Avec k-medoïds le centroïde est remplacé par un médoïde, c'est à dire l'objet réel (contenu dans $\mathcal{M}$) le plus proche du centroïde virtuel calculé. Soit mr^* ce médoïde, on le définit comme suit :

$$mr^* = m_i^l \in C_l | underset i = 1..|C_l| Mind(m^*, m_i^l)$$

Dans ce cas, une étape supplémentaire est ajoutée à l'algorithme de k-means après le recalcul des centroïdes.

Complexité : A chaque itération pour affecter un objet à un centroïde, on doit calculer la distance entre chaque objet (n au total) et tous les centroïdes (k). Nous avons donc une complexité en O(kn) pour l'affectation. Puis les centroïdes sont recalculés. Pour cela pour les k clusters, on calcule une moyenne. La complexité passe donc à O(knk) donc O(nk^2) par itération. Lorsque l'on se place dans le cadre de k-medoïds, la complexité est plus élevée car après chaque recalcul de centroïde, il faut trouver l'objet réel le plus proche de celui-ci. Pour cela on calcule la distance entre chaque objet et le centroïde du cluster auquel il appartient, ce qui ajoute n calculs. Nous obtenons donc une complexité en O(k^2n^2).

La méthode k-means et la méthode k-medoïds convergent vers un minimum local mais la convergence vers un minimum global ne peut être garantie.

3.2.2 *k*-median/*k*-center : Formulation

Les problèmes k-center et k-median (aussi appelés p-center et p-median dans la littérature) sont connus comme des problèmes de location/allocation que nous détaillons ensuite. Nous traitons ensemble ces deux problèmes dont l'objectif est similaire malgré un critère à optimiser différent.

Ce problème de location-allocation est formulé dans la section 3.2.2.1 dans le cadre du k-median et dans la section 3.2.2.2 dans le cadre du k-center. Il existe plusieurs façons de formuler ces problèmes notamment en terme de "graphe" ou en terme de "programmation entière". Nous verrons que ces modélisations permettent d'utiliser des algorithmes dédiés à ces approches. Nous verrons donc des solutions qui permettent de résoudre ces approches, et nous verrons qu'elles ne sont pas adaptées aux gros volumes de données que nous devons traiter. Enfin nous détaillerons les heuristiques applicables à k-median et k-center dans la section 3.2.3 et nous verrons que plusieurs méthodes existantes peuvent être comparées. Nous présenterons ensuite la méthode combinant plusieurs heuristiques que nous avons développée pour améliorer les résultats.

3.2.2.1 *k*-median

Formulation du problème Ce problème a été formulé dans le cadre d'allocation de ressources. Soit L un jeu de m emplacements de ressources (en anglais location), U un ensemble de n utilisateurs en demande de ressources (ou clients) et une matrice D de taille n × m contenant les distances entre chaque utilisateur $U_i \in U$ et les emplacements de ressources $L_j \in L$ (ou le coût induit par l'acheminement des ressources à un utilisateur). L'objectif de ce problème pour un fournisseur de ressources (ou produits) est alors de trouver un sous-ensemble de L de taille p qui minimise la somme des distances parcourues pour répondre à la demande des clients :

$$\underset{l \subseteq L \ |l|=p}{ArgMin} \sum_{U_i \in U} \underset{L_j \in l}{Min}\ d(U_i, L_j)$$

Notre problème peut être vu comme un cas particulier du problème k-median dans le cas où L = U. Il s'agirait alors de trouver un sous-ensemble qui vérifie l'argument suivant :

$$\underset{\mathcal{C} \ partition \ de \ \mathcal{M}}{ArgMin} \sum_{i=1...n} \underset{j=1...k}{Min}\ d(m_i, c_j^*)$$

Où $\mathcal{C}$ est la partition associée à l'ensemble $C^* = \{c_l^*\}_{l=1...k}$.
Cela revient à minimiser la somme des distances entre toute molécule et son centre le plus proche. Le problème du k-median est donc différent du nôtre. Cependant un certain nombre d'heuristiques permettant de le résoudre s'appliquant également au problème du k-center, elles seront présentées dans les sections suivantes.

3.2.2.2 *k*-center

Formulation du problème Drezner [142] propose deux formulations pour le problème k-center. Comme pour le problème k-median, nous avons un jeu L de m emplacements et un jeu U de n points de demande.
La première formulation propose de trouver k nouveaux emplacements dans L qui minimisent la distance pondérée maximale entre un point de demande dans U et l'installation

dans L la plus proche.

La deuxième formulation s'effectue en deux phases. Dans un premier temps le jeu U de points de demande est partitionné en k groupes disjoints. Puis pour chacun de ces groupes une installation dans L est choisie pour minimiser la distance maximale entre un point du groupe et l'installation attribuée à celui-ci. Cela permet de sélectionner un sous-ensemble de L. La distance maximale parmi toutes celles obtenues pour la partition de U doit être minimisée par le choix des sous-ensembles de L. Le critère à optimiser s'écrit donc comme suit :

$$\underset{l \subseteq L}{Argmin} \ \underset{U_i \in U}{Max} \ \underset{L_j \in l}{Min} \ d(U_i, L_j)$$

Dans notre cas, comme pour le problème k-median, les points de demande U sont les molécules $\mathcal{M}$ et le jeu d'emplacements L correspond aussi à l'ensemble des molécules. Ces dernières ne sont pas dans un jeu de données séparé mais inclus dans l'ensemble $\mathcal{M}$. Le critère du problème k-center devient donc :

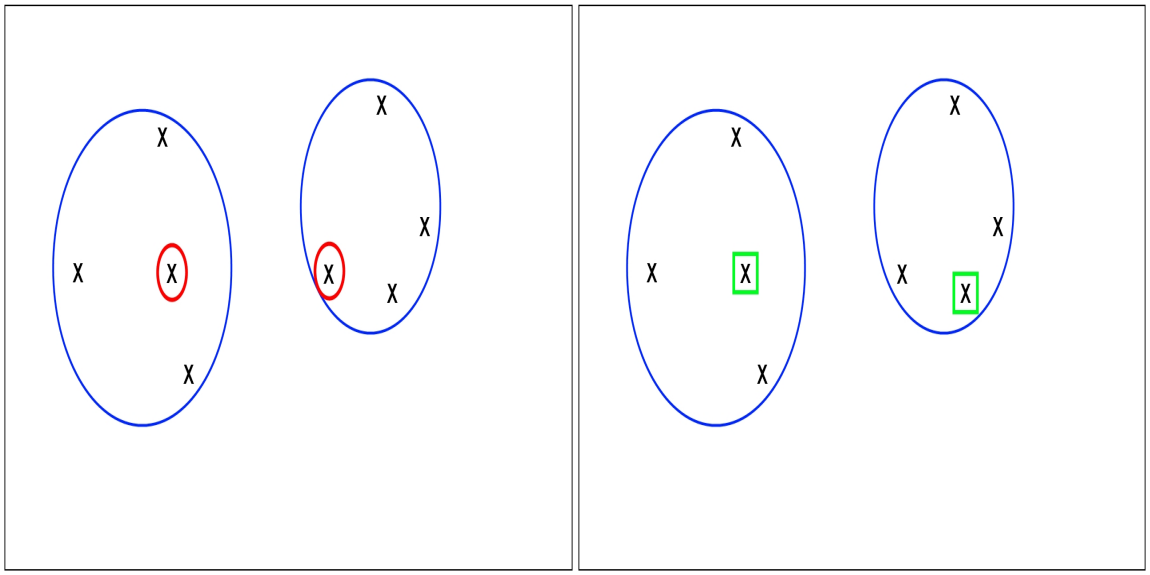
$$\underset{C^* \subset \mathcal{M}}{Argmin} \ \underset{i=1...n}{Max} \ \underset{j=1...k}{Min} \ d(m_i, c_j^*)$$

Ici j qui correspondait à un parcours des centres possibles pour le problème k-center se limitait à un nombre m de centres. Or dans notre cas les centres possibles correspondent à toutes les molécules de l'ensemble de départ soit n molécules (avec n très supérieur à m).

Rappelons notre objectif de départ avec le critère suivant :

$$\underset{\mathcal{C} \ partition \ de \ \mathcal{M}}{ArgMin} \ \underset{l=1...k}{Max} \ \underset{j=1...|C_l|}{Min} \ \rho(C_l | m_j^l)$$

Les deux critères semblent équivalents. Cependant, celui de k-center, en partant de tous les jeux de centres possibles, minimise le rayon maximum d'une partition obtenue à partir d'un ensemble de centres. Or notre critère minimise le rayon maximum d'une partition de $\mathcal{M}$ parmi toutes les partitions possibles, et ceci dans le but d'obtenir l'ensemble de centres correspondant. Nous avons donc deux critères qui auront une valeur identique une fois minimisés mais qui peuvent conduire à des partitions et des ensembles de centres différents. Nous illustrons cette différence par les schémas 3.1(a) et 3.1(b) montrant une partition obtenue après optimisation du critère de k-center et de notre critère respectivement. Le schéma 3.1(a) montre donc qu'à partir d'un jeu de centres donné, une bonne partition peut être trouvée sans pour autant que le centre choisi ne soit le meilleur représentant du groupe (dans le groupe de droite, le centre n'est pas le meilleur représentant du groupe). Le schéma 3.1(b) montre que par notre critère, comme nous cherchons les meilleurs centres à partir d'une partition, une fois la bonne partition trouvée, les centres correspondant seront les meilleurs représentants des groupes. En effet dans le groupe de droite, le centre choisi est plus central dans le groupe que celui choisi par k-center dans le schéma 3.1(a). Or notre objectif premier n'est pas de trouver une bonne partition, mais plutôt un ensemble de centres représentatifs de toute molécule de l'ensemble de départ , et ce afin de traduire la diversité (que l'on peut qualifier de dispersion dans l'espace des descriptions) du jeu de molécules initial. C'est pourquoi malgré des valeurs finales similaires du critère , nous préfeérons adapter les algorithmes conçus pour le problème k-center à notre critère d'optimisation.



(a) Sélection via le critère k-center (b) Sélection via notre critère

FIGURE 3.1: Différence de sélection selon le critère k-center et selon notre critère

Enfin, les schémas précédents illustrent un cas idéal. En effet, dans le critère, c'est le rayon maximum de la partition qui influence le résultat. Donc dans tous les groupes qui ne réalisent pas ce maximum, on peut obtenir des résultats très éloignés de nos attentes. En effet, le schéma 3 montre que dans le cas d'un jeu comportant des outliers, minimiser le rayon maximum pourrait conduire à former des groupes de tailles hétérogènes induisant un mauvais pavage de l'espace.

Il existe donc des heuristiques applicables au problème k-center que nous détaillons ensuite pour aider à pénaliser ou favoriser certains phénomènes au cours de la recherche de centres pour obtenir une bonne solution. Nous nous intéressons donc particulièrement à la résolution du problème k-center.

Cas particulier : le problème 1-center Comme pour le problème k-median, on peut décrire un cas particulier du problème k-center : le problème 1-center étudié en détail par Durier en 1995 [143]. L'objectif de ce cas particulier est de trouver un unique emplacement dans L, qui minimise la plus grande distance entre un point dans un jeu de taille n et cet emplacement.

Autres approches Ce problème peut être résolu soit par des approches graphes, soit par des approches de programmation entière. Cependant, la complexité de ces approches est trop élevée. Pour donner un exemple nous détaillons ici l'approche graphe et une de ses applications. Soit un graphe complet, pondéré et non dirigé G =(V,E) où V est l'ensemble des nœuds (molécules pour nous) et E l'ensemble des arêtes symbolisant la plus petite distance entre deux nœuds. Le poids de chaque arête vérifie l'inégalité triangulaire ; c'est à dire, connaissant 3 nœuds A, B et C, le poids de l'arête connectant le nœud A au noeud B est inférieur ou égal à la somme des poids des arêtes connectant A à C et C à B. Pour tout jeu S inclus dans V et pour tout nœud v appartenant à V, on définit d(v,S) comme étant la longueur de la plus petite arête de v à tout nœud dans S. Soit :

$$\forall S = \{s_j\}_{j=1...|S|} \subseteq V, \quad \forall\ v_i \in V \quad d(v_i, S) = \underset{j=1...|S|}{Min}\ d(v_i, s_j)$$

Il s'agit alors de trouver un jeu $S \subseteq V$ où $|S| \leq k$ qui optimise le critère suivant :

$$\underset{S \subseteq V}{Argmin}\ \underset{v_i \in V}{Max}\ d(v_i, S) \quad soit \quad \underset{S \subseteq V}{Argmin}\ \underset{v_i \in V}{Max}\ \underset{j=1...|S|}{Min}\ d(v_i, s_j)$$

Mihelic et Robic [144] proposent une heuristique, l'algorithme du jeu dominant, qui résout le problème k-center sous sa forme graphe en temps polynomial. Sa complexité pour le problème k-center atteint en effet $O(n^2 log(n))$. Cette complexité semble satisfaisante pour les auteurs qui testent des jeux de 900 nœuds au maximum. Mais elle est trop élevée pour notre problème puisque la taille de notre jeu de données peut atteindre plusieurs millions d'objets.

3.2.3 *k*-median/*k*-center : Stratégies de résolution heuristiques

Pour la suite nous nous concentrons sur notre problème de recherche d'un échantillon de k molécules. Il existe différentes stratégies de recherche comme la stratégie d'ajouts itératifs, de suppression itérative ou encore d'alternance. Nous présentons ces stratégies et quelques heuristiques les utilisant. Notamment nous parlerons des heuristiques classiques citées par Mladenovic et al. [145] applicables aux problèmes k-center et k-median. Nous présenterons également d'autres heuristiques proposées spécifiquement pour le problème du k-center. Enfin nous présenterons l'algorithme que nous avons développé pour approcher une solution du problème k-center appliquée à nos données.

3.2.3.1 Stratégie d'ajout itératif

Cette stratégie consiste à choisir un point comme centre pour initialiser le sous-ensemble divers, puis à ajouter itérativement un centre à ce sous-ensemble. L'algorithme correspondant a donc besoin de deux paramètres en entrée : le choix du point initial et la méthode d'ajout d'un nouveau centre.
Parmi les heuristiques utilisant cette stratégie, il existe l'heuristique gloutonne (ou greedy) dont nous détaillerons 3 algorithmes, et l'heuristique Core-Sets proposée par Badoiu et al. [146].

Heuristiques de type greedy ou gloutonnes On initialise l'ensemble représentatif à un jeu vide de centres. Puis le problème 1-median ou 1-center (cf. section 3.2.2.2) est résolu dans l'ensemble de centres possibles et ajouté à ce jeu. Chaque centre est donc ajouté un par un jusqu'à en obtenir k. A chaque itération le centre qui optimise le critère choisi est sélectionné (nous verrons quels critères sont utilisés en fonction de chaque méthode). L'algorithme 2 présente ces étapes. Pour l'initialisation du premier centre, il existe trois possibilités :

- initialisation par une molécule tirée aléatoirement parmi les molécules de $\mathcal{M}$
- initialisation à la molécule centrale du jeu $\mathcal{M}$
- initialisation à la molécule la plus éloignée des autres molécules de $\mathcal{M}$

Il existe deux exemples de cette heuristique également utilisés en chemoinformatique pour la sélection par diversité : Farthest First Traversal (FFT) et Sphere-exclusion. En 1998, Snarey et al. [129] comparent la méthode Sphere-exclusion à la méthode Maximum

Dissimilarity (dont FFT est une variante) dans le cadre de la chemoinformatique. Nous présentons Maximum Dissimilarity ainsi que sa variante FFT et leur algorithme. Ensuite nous présentons Sphere-Exclusion et enfin l'heurisitique Core-Sets sera évoquée.

Algorithm 2 Algorithme générique de la stratégie Greedy

1: *input k= taille de l'échantillon*
2: **Generate** $C^* = \emptyset$
3: **Etape 1** Initialisation $C^* = C^* + m_i \in \mathcal{M}$, i initialisé selon les méthodes citées plus haut
4: **Etape 2** Répéter k fois
5: Choix de $c^* \in \mathcal{M}$
6: Ajout de c^* dans C^*
7: **return** C^* l'ensemble des individus sélectionnés

Maximum Dissimilarity Pour Maximum Dissimilarity [129] il existe 3 initialisations possibles pour le premier centre (cf. Etape 1 de l'algorithme 2). Une fois le premier individu sélectionné, les suivants peuvent être ajoutés itérativement de deux façons différentes selon le critère que l'on souhaite optimiser (cf. Etape 2 de l'algorithme 2) :

Critère MaxMin (cf. Figure 3.2(a)) : l'objet ajouté est celui qui maximise la distance à son plus proche dans le sous-ensemble déjà sélectionné. Soit $m_i \in \mathcal{M}$ le nouvel élément ajouté dans $C^* = \{c_l^*\}$ l'ensemble des centres sélectionnés :

$$m_i \; est \; tel \; que \underset{l=1...|C^*|}{Min} d(c_l, m_i) = \underset{i=1...n}{ArgMax} \underset{l=1...|C^*|}{Min} d(c_l, m_i)$$

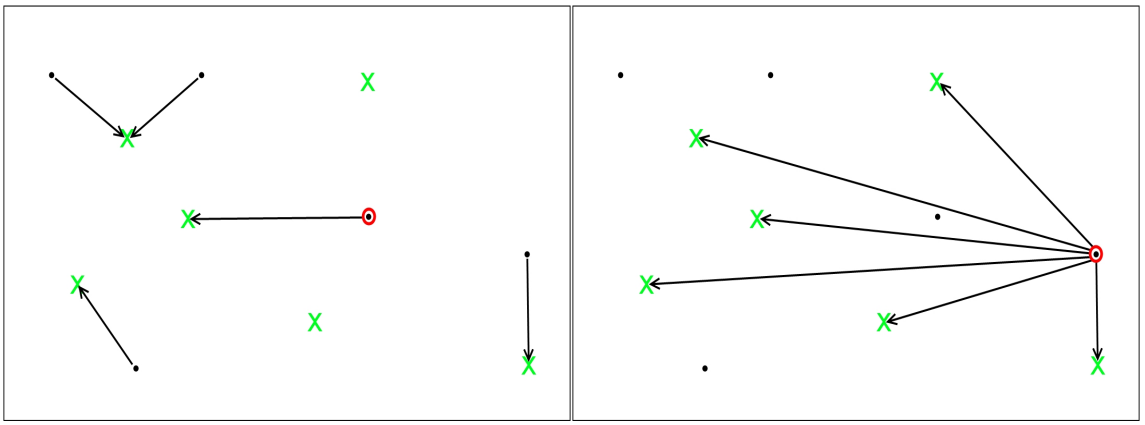
Critère MaxSum (cf. Figure 3.2(b)) : l'objet ajouté est celui qui maximise la somme des distances entre lui-même et chaque centre du sous-ensemble :

$$m_i = \underset{i=1...n}{ArgMax} \sum_{l=1...|C^*|} d(m_i, c_l)$$

Ces critères, associés aux différentes initialisations proposées, donnent les différentes variantes de la méthode Maximum Dissimilarity dont l'algorithme suit. Notons que nous considérons la méthode Farthest First Traversal, présentée par Hochbaum et Shmoys en 1985 [147], comme une variante de Maximum Dissimilarity car elle utilise une initialisation aléatoire du premier centre, puis le critère MaxMin pour choisir les centres suivants.

Complexité : O(k^2*N) avec k=le nombre de molécules à sélectionner et N le nombre total de molécules.

Sphere-Exclusion [148] En entrée de l'algorithme on donne le rayon et l'ensemble de molécules $\mathcal{M}$. Ensuite à l'intialisation, un premier individu est sélectionné selon les trois méthodes d'initialisation vues dans l'algorithme 2. Tous les individus situés dans le rayon, donné en entrée, autour de cet individu sélectionné sont supprimés. Puis le centre suivant



(a) MaxMin (b) MaxSum

FIGURE 3.2: Illustration d'une sélection selon MaxMin et selon MaxSum : En vert, le sous-ensemble déjà sélectionné et entouré en rouge la molécule suivante sélectionnée parmi les molécules noires

Algorithm 3 La méthode Maximum Dissimilarity

input k= taille de l'échantillon
Generate $C^* = \emptyset$
produce $C^* = C^* + m_i \in \mathcal{M}$, i initialisé selon les méthodes citées plus haut
produce $\mathcal{M} = \mathcal{M} - m_i$
$i \leftarrow 0$
while $!(i < k)$ **do**
  $C^* = C^* + (m_i = \underset{i=1...n}{ArgMax} \sum_{l=1...|C^*|} d(m_i, c_l)$ (critère MaxSum)
  OU
  $C^* = C^* + m_i = \underset{i=1...n}{ArgMax} \underset{l=1...|C^*|}{Min} d(c_l^*, m_i)$ (critère MaxMin)
  $i \leftarrow i + 1$
end while
return C^* l'ensemble des individus sélectionnés

peut être sélectionné selon quatre critères différents :

- un individu aléatoire
- l'individu dont la somme des distances au sous-ensemble est la plus petite : critère MinSum (résolution du problème k-median)
- l'individu dont la distance au plus proche du sous-ensemble est la plus petite : critère MinMin
- l'individu dont la distance au plus proche du sous-ensemble est la plus grande : critère MinMax (résolution du problème k-center)

Enfin l'algorithme (cf. algorithme 4) itère jusqu'à obtention de k centres.

On a vu qu'en entrée il faut déterminer un rayon. Or on ne connaît pas le rayon optimal permettant d'obtenir une sélection de k molécules sur le jeu $\mathcal{M}$ de taille n. C'est pourquoi avant le lancement de l'algorithme, il faut déterminer le rayon maximum. On le fixe a priori et on observe combien de molécules sont sélectionnées avec celui-ci. Puis on itère de façon à trouver le rayon optimal par rapport au nombre de molécules à sélectionner.

Algorithm 4 La méthode Sphere-Exclusion

input d_{seuil}
Generate $C^* = \emptyset$
while !($\mathcal{M}$ *non vide (k fois)*) **do**
  $C^* = C^* + m_i \in \mathcal{M}$ i initialisé selon l'une des méthodes citées plus haut
  for $j = 1$ TO n **do**
    $\mathcal{M} = \mathcal{M}$- m_j tel que d(m_i,m_j) > d_{seuil}
  end for
end while
return C^* l'ensemble des individus sélectionnés

Complexité : O(k*N) k nombre d'itérations et N nombre total de molécules.

Pour conclure, les heuristiques gloutonnes que nous venons de présenter permettent de résoudre le problème k-center avec une complexité raisonnable pour nos jeux de données. De plus elles ont été testées avec succès pour la diversité dans le domaine de la chemoinformatique. Etant donné notre objectif très lié au problème k-center, nous avons choisi de comparer Maximum Dissimilarity, FFT et Sphere-Exclusion avec la méthode k-medoïds également très utilisée dans la sélection par diversité en chemoinformatique. Nous ajouterons à cette comparaison une heuristique résolvant le problème k-center que nous avons développée. Celle-ci combinera quelques unes des heuristiques présentées par la suite pour tenter d'obtenir de meilleurs résultats qu'avec les méthodes traditionnellement utilisées en diversité.

Heuristique Core-Sets Enfin Badoiu et al. [146] présentent une heuristique qui permet d'extraire un sous-ensemble approximant le clustering et notamment pour les problèmes k-center et k-median. Leur méthode a une complexité linéaire. Leur but est d'approximer des hyper-sphères entourant les groupes d'individus en faisant l'hypothèse que tous les points ne sont pas utiles pour le calcul. Notamment les points se trouvant à proximité du centre ne définissent pas l'hyper-sphère. Ils utilisent donc moins de points pour le calcul de ces sphères, réduisant ainsi la complexité de l'algorithme.

Gärtner [149] propose lui un algorithme nommé "Mini-Balls" (dont nous nous inspirons

dans notre implémentation) pour résoudre également le calcul de la plus petite hypersphère englobant tous les points d'un groupe en un temps efficace.

3.2.3.2 Stratégies de suppression itérative

Les heuristiques adoptant cette stratégie débutent par la sélection d'un ensemble du jeu de départ. Puis à chaque étape, une molécule est supprimée jusqu'à l'obtention des k molécules souhaitées. Les algorithmes correspondant requièrent donc deux paramètres en entrée : le jeu de départ pour l'initialisation et la méthode de suppression. L'algorithme 5 présente les différentes étapes de cette stratégie. Parmi les heuristiques utilisant cette stratégie nous présenterons l'heuristique avare (stingy).

Algorithm 5 Algorithme générique de la stratégie de suppression itérative

1: *input k= taille de l'échantillon*
2: **Generate** $C^* = \emptyset$
3: **Etape 1** Initialisation $C^* = S$ où $S \subset \mathcal{M}$
4: **Etape 2** Répéter jusqu'à $|C^*| = k$
5: Choix de $m \in C^*$
6: Suppresion de m dans C^*
7: **return** C^* l'ensemble des individus sélectionnés

Heuristique Avare (stingy) [150] Cette heuristique commence avec une initialisation avec m centres ($m < |L|$ où L est l'ensemble des centres potentiels, $L = \mathcal{M}$ dans notre cas), puis ils sont supprimés un par un jusqu'à obtenir les k centres voulus. A chaque itération, le centre qui maximise le coût total est sélectionné et supprimé. Par la molécule qui maximise le coût total, on entend : la molécule qui maximise le critère du problème k-center par exemple. Or comme on souhaite le minimiser, la molécule qui le maximise est supprimée pour ne pas être sélectionnée. Une autre implémentation possible [151] consiste à commencer avec tous les centres potentiels.

3.2.3.3 Stratégie d'alternance

Cette stratégie consiste à choisir k points et à affecter chaque molécule à ces centres. Ensuite certains centres sont modifiés et les molécules sont réaffectées dans le but de trouver une combinaison de centres optimale. Les algorithmes utilisant cette stratégie demandent deux paramètres en entrée : le choix du jeu initial des k points et la méthode de choix des nouveaux centres. L'algorithme 6 résume ces étapes.

Algorithm 6 Algorithme générique de la stratégie d'alternance

1: *input k= taille de l'échantillon*
2: **Generate** $C^* = \emptyset$
3: **Etape 1** Initialisation $C^* = S$ où $S \subset \mathcal{M}\rceil\sqcup|\mathcal{S}| = \|$
4: **Etape 2** Répéter jusqu'à obtention du jeu C^* optimal
5: Affectation de chaque molécule dans $\mathcal{M}$ à son centre le plus proche dans C^*
6: Modification de certains centres dans C^*
7: **return** C^* l'ensemble des individus sélectionnés

Heuristique Alternate Proposée par Maranzana en 1964 [152], l'heuristique Alternate consiste à localiser les centres en k points. Les molécules sont affectées à leur centre le plus proche puis un problème 1-median ou 1-center est résolu pour chaque centre et son jeu de molécules (c'est à dire les molécules dont le centre est le représentant). Cette procédure est itérée avec les nouveaux centres jusqu'à ce qu'il n'apparaisse plus aucun changement d'affectation. Cette alternance de localisation/allocation mène à une méthode exhaustive exacte si tous les jeux de centres proches sont choisis comme solution initiale. Mais ceci n'est pas possible dans la plupart des cas car la complexité de O(m^p) serait trop élevé pour le nombre de données testées.

Heuristique Interchange Teitz et Bart en 1968 [153] proposent la méthode Interchange. Un certain jeu de k centres est donné à l'initialisation. Puis les centres sont changés itérativement par une molécule non-centre dans le but de réduire le coût total. Ce processus de recherche totale est stoppé quand plus aucun mouvement autour d'une installation ne permet de baisser la valeur de l'argument (soit $\underset{l \subseteq L}{Argmin} \underset{U_i \in U}{Max} \underset{L_j \in l}{Min} d(U_i, L_j)$ dans le cadre du problème k-center). Notre grande masse de données ne permet pas d'utiliser ce genre de solution puisque dans notre cas les non-centres potentiels correspondent à toutes les molécules du jeu de départ, la complexité d'une telle heuristique serait donc trop élevée.

D'autres algorithmes comme PAM (Partitioning Around Medoids)et CLARA (Clustering LARge Applications) [154] sont des variantes des k-medoïds. PAM propose de permuter systématiquement un medoïde avec un objet non medoïde de l'ensemble des objets. Cet objet est alors conservé comme nouveau medoïde si la qualité de la classification augmente. C'est à dire si la somme des distances entre chaque objet et son medoïde le plus proche diminue. Lorsque plus aucune permutation n'améliore la classification, l'algorithme s'arrête. La complexité d'un tel algorithme est $O(i \times k(n-k)^2)$ où i est le nombre d'itérations, k le nombre de groupes et n le nombre total d'objets. Pour améliorer cette complexité, l'algorithme CLARA propose de travailler sur plusieurs échantillons de la population en leur appliquant l'algorithme PAM et conserve la meilleur solution. Enfin l'algorithme CLARANS (Clustering Large Applications RANdomized Search) proposé par Ng et Han [155], une variante de CLARA, propose d'étendre la zone de recherche d'objets à permuter. CLARANS donne donc des classes de meilleure qualité que celles produites par PAM et CLARA mais perd sa capacité à traiter de grandes bases de données comme CLARA.

Métaheuristiques Mladenovic et al. [156] présentent également les métaheuristiques "Tabu" et "Neighborhood Search" pour résoudre le problème k-center. Tabu search consiste à construire une liste de substitutions de centres dans le cadre d'une utilisation de l'heuristique interchange. Cette liste correspond à des substitutions à ne plus vérifier à chaque itération car elles ont été calculées comme n'apportant pas d'amélioration. Cette métaheuristique permet ainsi de réduire les temps de calculs des différentes solutions à chaque itération.
L'idée de Variable Neighborhood Search est de procéder systématiquement à un changement de voisin dans un algorithme de recherche locale. Il s'agit de rester dans la même configuration (solution) tant qu'une solution meilleure n'a pas été découverte. Lorsque c'est le cas il passe alors à la solution suivante. Selon les auteurs, cette solution présente deux avantages au regard d'autres méthodes de recherche de minimum local et de trajectoire : en restant dans la même configuration, la recherche continue dans une zone attractive de

l'espace des solutions, et comme les attributs des solutions possèdent déjà leurs valeurs optimales, la recherche locale est moins coûteuse en itérations que si l'initialisation avait été faite aléatoirement.

3.2.3.4 Heuristiques composites

Enfin, Mladonevic et al. [145] présentent plusieurs combinaisons déjà testées de ces heuristiques classiques, comme par exemple une combinaison gloutonne-alternate proposée par Captivo en 1991 ou encore une combinaison alternate-interchange proposée par Pizzolato en 1994.

3.3 Notre implémentation du problème k-center

Nous avons vu que notre objectif correspond à résoudre le problème du k-center dans un cas particulier : celui d'un jeu de centres potentiels confondu au jeu de données de départ. Nous rappelons donc ici notre objectif et les notations le concernant.
Soit $\mathcal{M} = \{m_i\}_{i=1..n}$ l'ensemble de nos objets (molécules) pouvant également être des centres de groupes et $d(m_i, m_j)$ une distance entre deux molécules. Notre objectif est de trouver une sous-population de $\mathcal{M}$ constituée de k molécules : $C^* = \{c_j^*\}_{j=1...k}$ telles que ces k molécules sont considérées comme les représentantes de l'ensemble $\mathcal{M}$, on les appelle alors centres de groupes. Si chaque objet de $\mathcal{M}$ est affecté à son centre le plus proche nous obtenons une partition $\mathcal{C} = \{C_j\}_{j=1...k}$ où chaque groupe C_j est centré sur la molécule c_j^* de la sous-population C^*
Ayant une telle partition on peut alors définir le rayon relatif d'un groupe (cf. définition 3.1.1), son rayon absolu (cf. définition 3.1.1) pour ensuite en déduire le rayon absolu d'une partition que nous rappelons ici :

$$\rho(\mathcal{C}) = \underset{j=1...k}{Max}\ \rho(C_j)$$

Notre objectif est donc de minimiser le rayon absolu de la partition afin que les centres soient de bons représentants de l'ensemble $\mathcal{M}$ selon l'argument suivant :

$$\underset{\mathcal{C}\ partition\ de\ \mathcal{M}}{ArgMin}\ \rho(\mathcal{C})$$

Où C^* est un sous-ensemble de $\mathcal{M}$ et $\mathcal{C}$ la partition correspondante. Notons que le rayon d'un groupe $\rho(C_j)$ est défini par très peu d'objets, en général un seul : celui dont la distance au centre le représentant est la plus grande (égale au rayon). Ces objets seront appelés par la suite des objets critiques.

Pour atteindre notre objectif nous avons choisi un algorithme (cf. algorithme 7) implémentant une combinaison de plusieurs stratégies heuristiques présentées plus haut.

Nous détaillons ici les différentes étapes de cet algorithme, puis nous expliquons les stratégies que nous avons utilisées pour chacune.
1. Initialiser la sous-population C^{*0} aléatoirement ou avec les centres obtenus par FFT (ceci peut permettre de commencer l'algorithme dans un minimum local et ainsi permettre d'améliorer la solution finale et de converger plus vite).
2. Répéter

(a) Affecter chaque objet à un centre pour créer les groupes C_j^0 de la partition $\mathcal{C}^{\ 0}$

Algorithm 7 Notre implémentation du k-center

1: *input* k = *la taille de l'échantillon*
2: **Generate** C^*0 l'ensemble des centres
3: $i \leftarrow 0$
4: **while** !(*critère d'arrêt*) **do**
5: $\mathcal{C}^{\,i}$ la partition obtenue en affectant chaque observation à un centre
6: C^{*i+1} les nouveaux centres de cette partition
7: Conservation en mémoire de la distance entre chaque objet et son centre le plus proche
8: Supprimer p centres dont le coût de suppression est minimum
9: Ajouter p centres correspondant aux p objets critiques
10: Faire varier p
11: Conservation en mémoire de la meilleure solution rencontrée
12: $i \leftarrow i + 1$
13: **end while**
14: **return** C^*, $\mathcal{C}$

(b) Calculer le meilleur centre de chaque groupe (crée C^{*1})
(c) Calculer pour chaque objet le centre le plus proche (crée $\mathcal{C}^{\,i+1}$)
(d) Supprimer p centres (ceux pour lequel le coût de suppression est minimum)
(e) Ajouter p centres correspondant successivement aux objets critiques

Retour au début de l'étape 2.

Les étapes 2 (a) et (b) correspondent à la stratégie alternate puisque nous avons une alternance d'étapes d'affectation et de calcul des centres. Les étapes (d) et (e) correspondent à la stratégie interchange car des centres sont supprimés pour être remplacés par d'autres. Enfin une heuristique gloutonne est utilisée pour l'ajout de nouveaux centres car ce sont les objets critiques qui sont utilisés pour cela comme pour la méthode FFT par exemple.

3.3.1 Etape 1

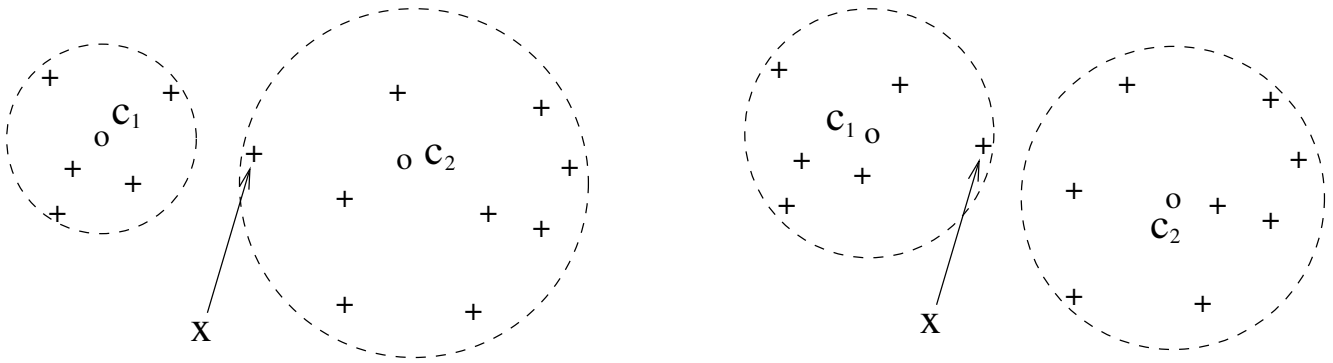
Deux méthodes d'initialisation sont envisagées.
- On initialise les k centres à k objets par un tirage aléatoire suivant une loi uniforme, sans remise dans le jeu de départ
- Ou on les initialise avec l'ensemble de centres obtenus par la méthode FFT

3.3.2 Etape 2 (a) : Affectation des objets aux centres

Pour cette étape, nous avons testé deux techniques d'affectation nommées CLOSE et BEST que nous expliquons ici.

3.3.2.1 Affectation avec CLOSE

CLOSE consiste à affecter une molécule à son centre le plus proche, celui dont la distance à la molécule est la plus petite parmi tous les centres. Cette technique est celle utilisée par k-means. Elle est linéaire avec n le nombre d'objets total, k le nombre d'objets à sélectionner (ou nombre de centres), et D la dimensionnalité ou nombre de variables décrivant les objets.



(a) Avec la règle CLOSE

(b) Avec la règle BEST

FIGURE 3.3: Impact des règles d'affectation CLOSE et BEST sur la qualité des rayons

Si l'on considère la figure 3.3 avec les deux centres c_1 et c_2, l'affectation d'un objet x peut changer le rayon de la partition. En effet, sur la figure 3.3(a), x est affecté à son centre le plus proche c_2 car $d(x, c_2) < d(x, c_1)$. Il en découle les deux sphères englobant les autres objets. Dans le cas de cette figure et/ou si l'on considère que la partition ne change pas, la règle d'affectation CLOSE ne changera pas cette partition au cours des itérations puisque x sera toujours affecté à ce centre et l'algorithme stoppera. Comme nous travaillons avec un algorithme basé sur l'alternance de phases d'affectation et de calcul de centres, il est intéressant de modifier cette règle d'affectation pour forcer les centres à changer même si une partition restait la même.

3.3.2.2 Affectation avec BEST

Nous proposons donc une autre règle d'affectation (BEST) différente afin de réduire l'attractivité des centres de groupes de grand rayon pour obtenir des groupes du type de la figure 3.3(b) . Pour cela rappelons que le rayon relatif d'un groupe C_j^i centré sur c_j^{*i} est noté $\rho(C_j^i|c_j^{*i})$. Le rayon relatif du nouveau groupe C_j^{i+1} est noté ainsi : $\rho(C^{i+1}|c_j^{*i})$ lorsque ce groupe est obtenu à partir du groupe C_j^i en ajoutant une observation x.
Nous définissons le coût d'affectation de x à C_j^i noté $cout(x, C_j^i)$:

- si $d(x, c_j^{*i}) \leq \rho(C_j^i|c_j^{*i})$ et si x est ajouté à C_j^i, alors le rayon relatif ne change pas et le coût est défini ainsi : $cout(x, C_j^i) = d(x, c_j^{*i})$
- si $d(x, c_j^{*i}) > \rho(C_j^i|c_j^{*i})$, le rayon théorique de $C_j^i \bigcup \{x\}$ est $(\rho(C_j^i|c_j^{*i}) + d(x, c_j^{*i}))/2$ comme montré en figure 3.4 (où r est le rayon initial du groupe centré sur c). Le coût devient alors : $cout(x, C_j^i) = (\rho(C_j^i|c_j^{*i}) + d(x, c_j^{*i}))/2$

Ce dernier coût correspond à un rayon théorique, c'est à dire une borne supérieure théorique que peut atteindre le rayon une fois la phase d'affectation terminée. En effet, pour chaque molécule, on calcule le coût d'affectation à chaque groupe et la molécule est affectée au groupe engendrant le coût minimum. On pourrait calculer le rayon réel (c'est à dire rayon en relation au vrai centre du groupe) pour chaque groupe à chaque tentative d'affectation d'un objet. Cependant ce calcul serait trop lourd pour l'ensemble des molécules, car il faudrait pour ce rayon réel, connaître le centre réel. Pour cela, il faudrait qu'à chaque affectation d'un nouvel objet, on recalcule le centre réel du groupe.

C'est pourquoi on calcule ce rayon théorique qui approxime une borne supérieure que pourrait prendre le rayon (ce rayon théorique ou virtuel dépend alors d'un centre virtuel du groupe qui ne change pas tout au long des affectations).

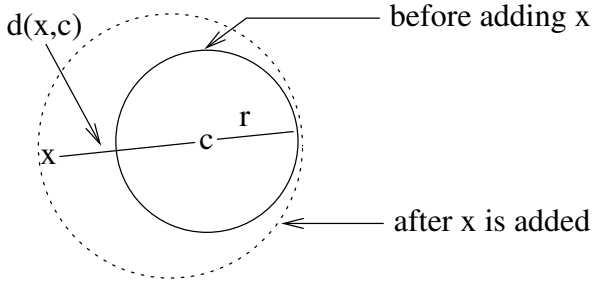


FIGURE 3.4: Pondération du coût d'affectation

La règle d'affectation BEST consiste donc à associer une molécule au centre associé au coût le plus faible. Notons que suite à cette phase d'affectation, le rayon réel de certains groupes pourrait être plus grand que le rayon de la partition et ceci pour deux raisons :

- soit il n'existe pas de molécule proche du nouveau centre virtuel d'un groupe
- soit comme l'affectation est faite pour chaque molécule de façon indépendante des autres, deux affectations successives à un groupe peuvent augmenter le rayon de la partition

Nous pourrions utiliser une méthode de recalcul des centres après chaque affectation d'une molécule pour obtenir un algorithme qui converge à mesure que le rayon de la partition diminue. Cependant nous avons vu que cela demanderait un calcul des centres trop fréquent (autant de fois qu'il y a d'objets) donc un algorithme avec une complexité très élevée. Nous utiliserons donc une phase d'affectation complète sans recalcul des centres durant cette phase. Nous perdons ainsi la garantie d'une convergence mais la complexité de cette phase reste linéaire en n, k et la dimensionnalité.

3.3.2.3 Conclusion

Lors des tests de cet algorithme on a pu constater qu'en utilisant la règle CLOSE, l'algorithme convergeait plus vite mais vers un optimum local moins bon. Si on utilise BEST, le score final est un peu meilleur. Pour notre implémentation finale nous avons choisi d'explorer l'espace avec 15 itérations avec BEST puis pour converger nous faisons les itérations suivantes avec CLOSE.

3.3.3 Etape 2 (b) : Calcul des centres

Calculer le meilleur centre pour chaque groupe revient à résoudre le problème 1-center. C'est à dire que l'on prend chaque objet d'un groupe comme centre possible et on calcule le rayon du groupe pour trouver le centre qui le minimise soit :

$$\underset{l=1...|C_j|}{Argmin}\ \rho(C_j|m_l)$$

On effectue cette opération pour les k groupes. Le problème de cette technique est que pour les grands groupes la complexité explose. Dans le cas de petits groupes, le vrai meilleur centre est important car une approximation pourrait être très mauvaise s'il y a peu de points autour du centre. En revanche pour les grands groupes, une approximation du meilleur centre est acceptable. Pour cela, nous pouvons calculer soit le centre du plus petit hyper-rectangle contenant tous les objets du groupe (H-centers), soit le centre de la plus petite hyper-sphère (S-centers). Ensuite on cherche l'objet réel le plus proche de ce centre.

3.3.3.1 H-centers

Utiliser H-centers revient à choisir le centre du plus petit hyper-rectangle contenant tous les objets du groupe. Pour cela, dans chaque dimension de l'espace, on cherche le minimum et le maximum des valeurs parmi les objets du groupe et on prend la valeur centrale. Ces valeurs dans toutes les dimensions donnent les coordonnées du centre de l'hyper-rectangle. On choisit ensuite le point réel le plus proche de ce centre. Cette méthode est linéaire avec n et la dimensionnalité mais ne donne pas la meilleure approximation du centre du groupe.

3.3.3.2 S-centers

Utiliser S-centers revient à choisir le centre de la plus petite hyper-sphère contenant tous les objets du groupe. Cette opération est plus complexe mais donne une meilleure approximation de la réalité. Il existe des méthodes d'approximation comme les core-sets. Cette méthode considère que parmi tous les points d'un groupe, il existe beaucoup d'objets non déterminants pour le calcul de la sphère (les points autour du centre). Donc cette méthode ne considère que les points du pourtour du groupe. Gartner [149] propose un algorithme avec une complexité de n mais le nombre d'itérations peut être élevé.

3.3.3.3 Conclusion

Pour les groupes avec un effectif inférieur à 500 objets, le vrai centre est calculé. Pour les autres, S-centers est utilisé avec l'algorithme d'approximation de Gärtner.

3.3.4 Etape 2 (c) : Calcul du centre le plus proche

Notons que cette étape ressemble à l'étape (a). Ce n'est pas une vraie phase d'affectation mais on détermine comme pour la phase (a) le centre le plus proche de chaque objet et ceci pour déterminer le rayon de la partition (ou rayon maximum). Ce rayon est utile pour les étapes (d) et (e). Cette phase a donc la même complexité que l'étape (a).

3.3.5 Etape 2 (d) et (e)

Il s'agit de "casser" le jeu de centres pour explorer un peu l'espace des solutions et en trouver une meilleure.

Dans l'étape (d), on supprime un ou plusieurs centres (p que nous définissons ensuite) dont le coût de suppression est le plus faible. Pour calculer ce coût, on supprime un centre, puis on réaffecte les objets du groupe associé à celui-ci, au groupe le plus proche. On recalcule alors le rayon de la partition ainsi obtenue. La différence entre ce rayon et le rayon initial de la partition correspond au coût de suppression du centre qui a engendré cette nouvelle partition. On effectue ce calcul pour la suppression de chaque centre. Puis les p centres dont la suppression engendre le coût minimum sont supprimés.

Pour l'étape (e), on est proche de la stratégie gloutonne utilisée pour la méthode FFT. Quand on a une partition avec un certain rayon maximum, ce rayon est causé par un objet qui est le plus éloigné du centre du plus gros groupe. Donc si on utilise cet objet comme centre, on peut améliorer le rayon maximum. On ajoute donc comme centre les objets critiques pour diminuer la taille des rayons des plus grands groupes. Enfin comme pour l'étape (b), nous n'effectuons pas de réaffectation pendant l'étape (d) et l'étape (e).

Le nombre de centres à supprimer, noté p, est variable. Lorsque l'on supprime un seul centre, cela ne change pas beaucoup la solution initiale, on explore donc mal l'espace des solutions. En revanche si on supprime beaucoup de centres, on explore bien l'espace des

solutions, mais l'algorithme ne converge pas car les solutions sont trop différentes d'une itération à l'autre.
Pour déterminer le p optimal on s'est donc basé sur un critère : le rayon de la partition (ou rayon max). Au départ p a une certaine valeur (10 par exemple). On supprime p groupes (étape (d)) et on en rajoute p (étape (e)). Si ça améliore le rayon de la partition, on augmente p à l'itération suivante ; si le rayon est moins bon, on diminue p ; puis l'algorithme revient à l'étape (a) quelque soit la qualité de la solution. La meilleure solution sera sélectionnée à la fin de l'algorithme car au fur et à mesure des itérations, la meilleure solution rencontrée (c'est à dire la partition et son ensemble de centres associés donnant le rayon de la partition le plus faible) est toujours gardée en mémoire.

3.3.6 Critère d'arrêt

Enfin l'algorithme s'arrête soit au bout de 10 itérations qui n'améliorent pas beaucoup la solution, soit au bout de trois itérations identiques (c'est à dire que l'on obtient trois fois le même rayon de la partition).La meilleure solution trouvée durant les itérations est sélectionnée.
En terme de complexité on a une complexité en O(nkD) pour l'étape (a) d'affectation,multipliée par k pour l'étape (b), l'étape (c) est équivalente à l'étape (a). Enfin l'étape (d) a une complexité plus faible que l'étape d'affectation, et l'étape (e) a une complexité très faible, elle est aussi rapide que p itérations de FFT.

Chapitre 4

Sélection par diversité : Expérimentation

4.1 Plan d'expérimentation

Nous comparons donc 5 méthodes de sélection d'un sous-ensemble divers dont 3 basées sur la méthode k-center :

- Sélection Aléatoire (M1)
- Notre implémentation de k-center (M2) cf. section 3.3 et algorithme 7
- k-means / k-medoïds (M3) cf. section 3.2.1 et algorithme 1
- Maximum Dissimilarity (dont FFT) (M4) cf. section 3.2.3.1, paragraphe "Stratégies Itératives" et algorithme 3
- Sphere-Exclusion (M5) cf. section 3.2.3.1, paragraphe "Stratégies Itératives" et algorithme 4

Plusieurs paramètres sont nécessaires à la mise en place de certaines d'entre elles. Tout d'abord l'initialisation du premier objet ou du premier sous-ensemble de centres peut être effectuée de différentes manières ; nous testerons les suivantes :

- initialisation aléatoire (valable pour un objet ou un ensemble d'objets)
- initialisation à la molécule la plus centrale de l'ensemble de départ notée mc_{init}, elle est définie par la formule suivante :

$$mc_{init} = \underset{i=1...n}{ArgMin} \sum_{j=1...n} d(m_i, m_j)$$

- initialisation à la molécule la plus éloignée des autres molécules de l'ensemble de départ notée ml_{init}, elle est définie avec la formule suivante :

$$ml_{init} = \underset{i=1...n}{ArgMax} \sum_{j=1...n} d(m_i, m_j)$$

- initialisation d'un sous-ensemble avec les résultats de FFT

Ensuite les méthodes Maximum Dissimilarity et Sphere-Exclusion étant des méthodes itératives, elles requièrent un deuxième paramètre concernant le choix de la molécule sélectionnée à chaque itération.

- Maximum Dissimilarity : la molécule suivante est la plus éloignée de l'ensemble déjà sélectionné, ou la plus éloignée de sa plus proche dans l'ensemble sélectionné (cf. paragraphe 3.2.3.1)

– Sphere Exclusion : la molécule suivante est prise aléatoirement, ou celle dont la somme des distances au sous-ensemble est la plus petite, ou la plus proche de sa plus proche dans le sous-ensemble, ou la plus éloignée de sa plus proche dans le sous-ensemble (cf. paragraphe 3.2.3.1).

Et enfin nous avons choisi de tester deux paramètres différents pour k-medoïds. A chaque itération, lorsqu'un centre est calculé il peut être virtuel. D'un côté nous travaillons avec ces centres virtuels jusqu'à convergence de la méthode. Puis pour obtenir le sous-ensemble final, on prendra une molécule réelle la plus proche de chaque centre virtuel. Dans un second temps nous testerons la vraie méthode k-medoïds où un centre réel est substitué à un centre virtuel à chaque itération.

Le tableau 4.1 résume les paramétrages testés pour chaque méthode.

Méthodes	Aléatoire M1	k-center M2	k-medoïds M3	Maximum Dissimilarity M4	Sphere-Exclusion M5
Init Aléatoire	X	X	X	X	X
Init Centre			X	X	
Init le plus éloigné				X	X
Init FFT		X	X		
MaxSum				X	
MaxMin				X	
Aléatoire					X
MinSum					X
MinMin					X
MinMax					X
Centres réels		X	X		
Centres virtuels		X	X		

TABLE 4.1: Récapitulatif des paramètres utilisés pour chaque méthode comparée

Enfin rappelons que pour cette étude, nous utilisons 3 jeux différents composés chacun de 40 000 molécules (Cf. chapitre présentation des données) : jeu 1, jeu 3 et jeu 5. Ensuite les outliers de chacun de ces jeux ont été supprimés pour obtenir les jeu 2, jeu 4 et jeu 6 (respectivement issus des jeu 1, jeu 3 et jeu 5). Nous avons choisi de prendre tous les singletons identifiés par les 5 méthodes comme outliers pour commencer. L'échantillonnage par diversité a été réalisé pour 100, 500 et 1000 molécules (respectivement 0,25%, 1.25% et 2.5% des jeux initiaux).
Enfin pour toutes les méthodes utilisant des tirages aléatoires, 10 runs ont été effectués pour éprouver la stabilité des résultats.

4.2 Critères d'évaluation des méthodes comparées

Les méthodes évaluées ne sont pas toutes basées sur la formation de groupes de molécules avant sélection. Néanmoins il est possible de toutes les comparer avec les mêmes critères. En effet pour les méthodes basées sur les distances, nous avons considéré que les molécules sélectionnées étaient des centres de groupes. Ainsi une fois la sélection effectuée, chaque molécule non sélectionnée est affectée à sa molécule la plus proche parmi les sélectionnées. Nous obtenons alors des groupes comme pour les méthodes de clustering. Nos critères peuvent donc être appliqués à toutes les méthodes de sélection. Rappelons que

nous formons autant de groupes que de molécules à sélectionner.
Enfin pour une bonne compréhension des critères nous redonnons les notations :

- L'ensemble de n molécules de départ $\mathcal{M} = \{m_i\}_{i=1...n}$ avec n = 40000
- L'ensemble de k molécules sélectionnées par diversité (également appelé l'ensemble des centres de groupes) $C^* = \{c_j^*\}_{j=1...k}$ avec k = 1000, 500 ou 100
- L'ensemble des k groupes formant la partition des n molécules $\mathcal{C} = \{C_l\}_{l=1...k}$
- L'ensemble des p variables $\mathcal{V}=\{v_i\}_{i=1...p}$ telles que $v_j(m_i)$ désigne la valeur de la molécule $m_i \in \mathcal{M}$ pour la variable $v_j \in \mathcal{V}$ avec p = 200
- Le rayon d'un cluster C_l noté $\rho(C_l)$ (= le rayon de la plus petite sphère centrée sur un objet de C_l contenant tous les objets du cluster C_l)
- Le rayon absolu d'une partition $\mathcal{C}$ noté $\rho(\mathcal{C})$ (= le rayon de la plus grande sphère de cette partition)
- L'argument que nous souhaitons optimiser : $\underset{C^* \subset \mathcal{M}}{ArgMin}\ \rho(\mathcal{C})$

Nous présentons ici les critères numériques sur lesquels se base notre analyse des méthodes de sélection par diversité. Ces critères seront illustrés par des schémas (pour une meilleure compréhension de ceux-ci, une unité de mesure appelée Unité U pourra être visualisée sur certains d'entre eux). Beaucoup des critères numériques sont basés sur des sommes de distances et ne rendent pas bien compte de la réalité. Ceux-ci sont donc complétés par des critères graphiques concernant les histogrammes de distribution des distances. Les histogrammes dit "théoriques" que nous présentons par la suite ne sont pas issus d'un calcul théorique mais sont des exemples de l'idéal que l'on souhaite approcher avec les distributions issues des différentes méthodes comparées.

4.2.1 Critères en rapport aux groupes formés

4.2.1.1 Le rayon

- Rayon maximum noté rayon max
- Rayon moyen
- Ecart type des rayons noté EC Rayon

Le Rayon maximum est le plus grand rayon parmi tous les rayons des groupes, il s'agit du rayon absolu de la partition $\mathcal{C}$ de l'ensemble $\mathcal{M}$ dont nous redonnons la formule :

Définition 4.1 *Le **rayon maximum** se définit ainsi :*

$$\rho(\mathcal{C}) = \underset{l=1...k}{Max}\ \rho(C_l)$$

La rayon moyen est la moyenne des rayons absolus de chaque groupe de la partition $\mathcal{C}$ de $\mathcal{M}$, k étant le nombre de groupes de la partition. Cette moyenne s'effectue uniquement sur les groupes contenant plus d'un objet.

Définition 4.2 *Le **Rayon Moyen** ($\bar{\rho}$) se définit comme suit :*

$$\bar{\rho} = (\mathcal{C})\ \tfrac{1}{Nb} \sum_{l=1...k} \rho(C_l)$$

Où Nb est le nombre de groupes de cardinal supérieur à 1, soit : $Nb = |\{C_l|\ |C_l| > 1\}|$

Définition 4.3 *L'**écart type des rayons** se calcule ainsi :*

$$ECRayon(\mathcal{C}) = \sqrt{\tfrac{1}{Nb} \sum_{i=1,\ |C_i|>1}^{k} (r(\rho_i) - \bar{\rho})^2}\ \textit{avec}\ Nb = |\{C_l||C_l| > 1\}|$$

L'écart type des rayons et le rayon moyen se calculent sur tous les rayons des groupes ayant un effectif supérieur à 1. Ce qui signifie que les groupes singletons ne sont pas pris en compte pour ne pas fausser les résultats.

Rappelons que notre objectif est de minimiser le rayon absolu de la partition (critère rayon max, cf. section 3.1.1), et ce dans le but d'obtenir un ensemble de rayons de tailles homogènes pour que les groupes ainsi formés couvrent l'espace des descriptions de façon la plus homogène possible. Il s'agit donc de minimiser le critère "rayon max" mais également d'obtenir un "rayon moyen" proche du rayon max ainsi qu'un "EC Rayon" faible. Donc entre deux méthodes, celle qui a le rayon moyen le plus petit avec un écart type faible sera la méthode répondant le mieux à nos attentes.

On souhaite obtenir le Rayon moyen le plus petit possible et le plus proche possible du Rayon max, un EC Rayon le plus petit possible et un Rayon Max le plus petit possible.

4.2.1.2 EC Ind

Définition 4.4 *L'* ***écart type des Individus par groupe*** *est défini par la formule suivante :*

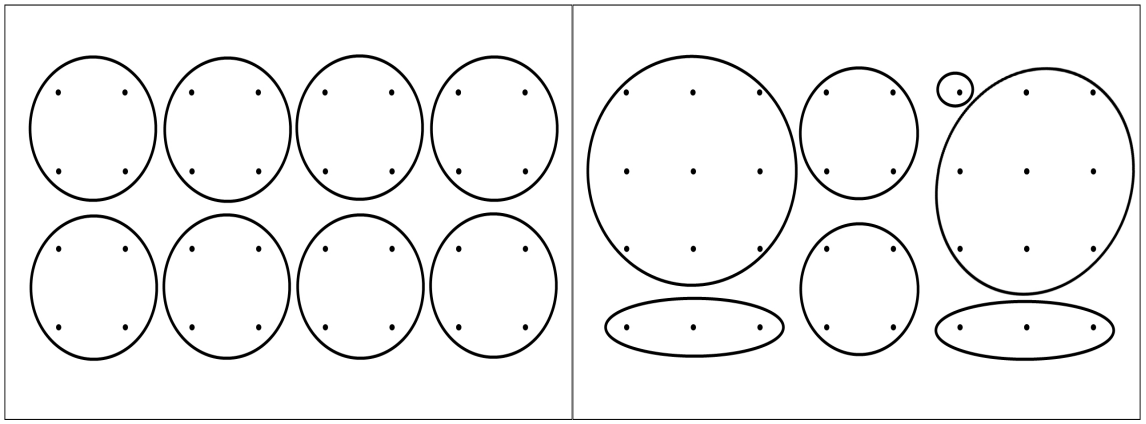
$$ECInd(\mathcal{C}) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(|C_i| - \overline{|C|})^2}$$

Il se calcule sur tous les groupes et renseigne sur l'écart entre les groupes sur le nombre de molécules contenues par ceux-ci.

Les schémas 4.1(a) et 4.1(b) montrent deux partitions différentes d'un jeu de données réparti uniformément dans un espace des descriptions en 2 dimensions.

Dans ce cas, un quadrillage homogène (c'est à dire une taille équivalente de tous les rayons) de l'espace (cf. Figure 4.1(a)) conduit à des groupes dont le nombre d'individus est le même ou quasiment le même pour chaque groupe.

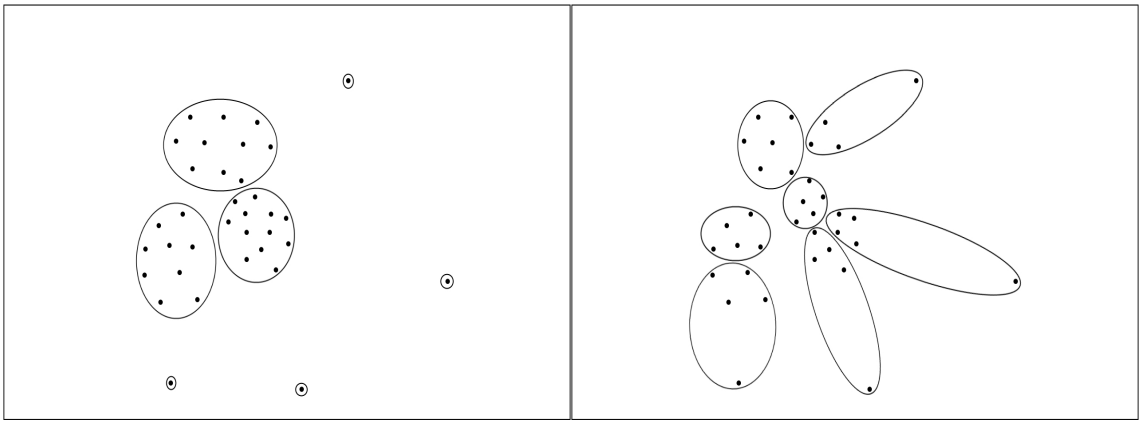
En revanche si le quadrillage n'est pas homogène (c'est à dire que la taille des rayons des groupes est non homogène), on remarque que le nombre d'individus par groupe (cf. Figure 4.1(b)) peut être très variable.



(a) Quadrillage homogène (rayons uniformes) (b) Quadrillage non homogène (rayons non uniformes)

FIGURE 4.1: Visualisation de l'écart du nombre d'individus par groupe pour un jeu de données uniformément réparti dans l'espace

Donc dans le cas d'un jeu initial homogènement réparti dans l'espace des descriptions, un partitionnement conduisant à un critère EC Ind de valeur élevée pourra être considéré comme non homogène et donc mauvais pour la diversité de l'échantillon.
D'un autre côté dans le cas d'un jeu initial non homogènement réparti dans l'espace (c'est



(a) Quadrillage homogène (rayons uniformes) (b) Quadrillage non homogène (rayons non uniformes)

FIGURE 4.2: Visualisation de l'écart du nombre d'individus par groupe pour un jeu de données non uniformément réparti dans l'espace

à dire avec des zones denses et des zones peu denses, cf. Figures 4.2(a) et 4.2(b)), on peut observer la tendance inverse.

En effet la figure 4.2(a) montre un partitionnement homogène de l'espace (c'est à dire : taille des rayons équivalente pour tous les groupes sauf les singletons). Les zones denses sont couvertes par quasiment autant de groupes que les zones peu denses. De ce fait le nombre d'individus dans les groupes est plus élevé dans les groupes couvrant les zones denses que dans les groupes couvrant les zones peu denses.

En revanche lorsque le partitionnement ne conduit pas à un quadrillage homogène de l'espace (cf. Figure 4.2(b)), les zones denses peuvent être couvertes par moins de groupes que les zones peu denses (3 groupes couvrent uniquement la zone dense en plus de 4 groupes couvrant les zones peu denses et la zone dense à la fois). Le nombre d'individus par groupe s'équilibre donc quelque soit la zone où se trouve le groupe.

Intuitivement, on pourra donc penser qu'un partitionnement qui aura tendance à donner un critère EC Ind de valeur faible sera moins bon qu'un partitionnement donnant un EC Ind de valeur plus élevée (dans le cas d'un jeu initial non homogènement réparti dans l'espace).

Conclusion
L'EC Ind doit être faible pour un jeu homogènement distribué dans l'espace et élevé pour un jeu non homogènement distribué.

4.2.1.3 Singletons

Les singletons sont des groupes composés d'une seule molécule.

L'histogramme 4.3 correspond à la distribution des distances entre chaque molécule et sa molécule la plus proche. En bleu sont projetées ces mêmes distances pour les singletons identifiés par les méthodes de sélection.

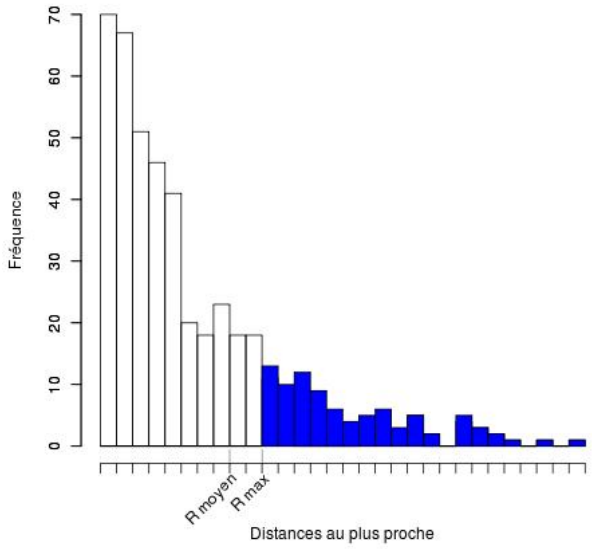


FIGURE 4.3: Histogramme théorique pour le critère Singletons : distribution des distances entre chaque molécule $m_i \in \mathcal{M}$ et sa molécule la plus proche.

Toute molécule appartenant à un groupe doit être à une distance de sa molécule la plus proche inférieure ou égale au rayon maximum de la partition. Ces molécules possèdent alors un représentant différent dans le jeu initial de molécules. Lorsqu'une molécule est à une distance de sa plus proche voisine, supérieure au rayon maximum de la partition, elle devrait alors être un singleton et nous la considérons alors comme un outlier puisqu'elle n'a pas de meilleur représentant qu'elle même et qu'elle est très éloignée de toute autre molécule. Dans ce cadre, toutes les molécules dont la distance à leur plus proche est supérieure au rayon maximum doivent être des singletons, et ceux-ci à cette condition seront considérés outliers. Si certains singletons ont une distance à leur molécule plus proche inférieure au rayon maximum, alors cela signifie qu'ils auraient pu former un groupe avec cette molécule. Ce ne sont donc pas de bons indicateurs d'outliers. Enfin si certaines molécules non singletons ont une distance à leur plus proche supérieure à celle que certains singletons ont eux-même avec leur voisine, alors ces derniers ne sont pas de bons indicateurs d'outliers. En effet cela signifie que des molécules appartenant à un groupe sont plus éloignées entre elles que certains singletons avec leur plus proche voisine. Dans ce cas, le singleton aurait pu former un groupe avec sa voisine et n'est donc pas un bon outlier.

4.2.2 Critères en rapport aux distances

4.2.2.1 Diversité-Recouvrement de l'espace

Pour évaluer la diversité et le recouvrement de l'espace nous avons 4 indicateurs qui sont calculés parmi les molécules de l'échantillon sélectionné :

- Somme des dissimilarités au plus proche voisin (ou Sum of Dissimilarity of Nearest Neighbor) notée SomDissimNN ou SDNN
- Dissimilarité minimum entre une molécule et son plus proche voisin notée SDNN Min
- Dissimilarité maximum entre une molécule et son plus proche voisin notée SDNN

Max
– Ecart type des dissimilarités au plus proche voisin noté EC SDNN
Où on définit NN par :

$$NN = \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1...k}{ArgMin}\ d(c_i^*, c_j^*)\}$$

Ce qui signifie qu'on considère deux individus c_i^* et c_j^* sachant qu'ils appartiennent à l'ensemble des représentants qui est aussi notre échantillon C^*, qu'ils sont deux individus différents et que c_j^* est le plus proche voisin du centre c_i^*. La somme des dissimilarités au plus proche voisin concerne donc l'échantillon divers sélectionné. Les distances entre chaque molécule de l'échantillon et sa molécule la plus proche dans l'échantillon C^* sont additionnées.

Définition 4.5 *La **somme des dissimilarités au plus proche voisin SDNN** est définie comme suit :*

$$SDNN(C^*) = \sum_{l=1...k} \underset{j=1...k, j\neq l}{Min} d(c_l^*, c_j^*)$$

SDNN Min est la plus petite distance entre une molécule de l'échantillon et sa plus proche dans l'échantillon divers C^*.

Définition 4.6 *La **dissimilarité minimum au plus proche voisin SDNN Min** se calcule ainsi :*

$$SDNNMin(C^*) = \underset{c_j^* \in C^*}{Min}\ \underset{l=1...k, j\neq l}{Min} d(c_l^*, c_j^*)$$

SDNN Max est la plus grande distance entre une molécule de l'échantillon et sa plus proche dans l'échantillon divers C^*.

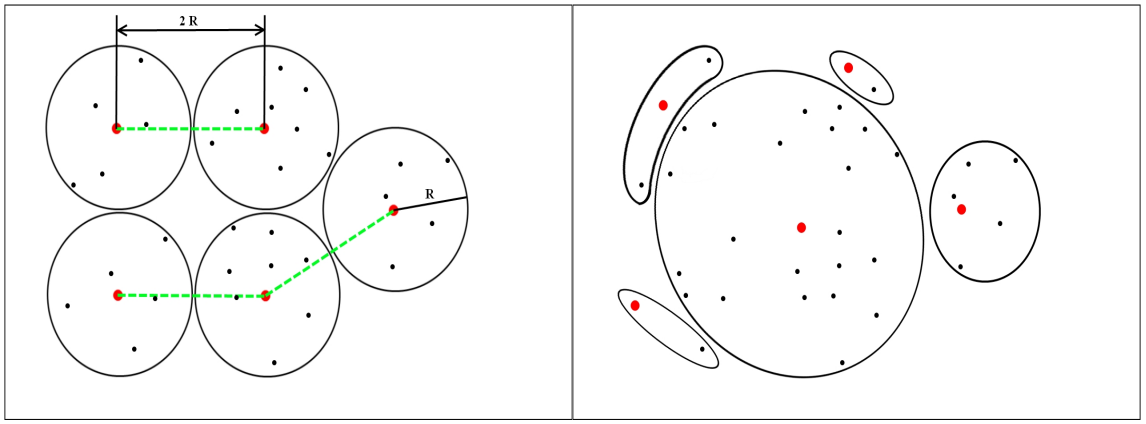
Définition 4.7 *La **dissimilarité maximum au plus proche voisin SDNN Max** est défini par la formule suivante :*

$$SDNNMax(C^*) = \underset{c_j^* \in C^*}{Max}\ \underset{l=1...k, j\neq l}{Min} d(c_l^*, c_j^*)$$

Définition 4.8 *L' **écart type** des dissimilarités (au nombre de k) au plus proche voisin EC SDNN se calcule ainsi :*

$$ECSDNN(C^*) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(\underset{j=1...k}{Min}\ d(c_i^*, c_j^*) - \overline{\underset{j=1...k}{Min}\ d(c_i^*, c_j^*)})^2}$$
$$où\ \overline{\underset{j=1...k}{Min}\ d(c_i^*, c_j^*)} = \frac{SDNN(C^*)}{k}$$

Interprétation numérique Nous calculons ici les distances entre chaque centre (ou objet de l'échantillon) et son centre le plus proche. Le schéma 4.4(a) montre un partitionnement idéal d'un ensemble de points répartis dans un espace à deux dimensions. En effet, cet espace est quadrillé de façon homogène par les groupes, les rayons de ceux-ci étant équivalents entre eux. De plus les groupes se juxtaposant les uns aux autres, on peut dire que l'ensemble des points est couvert de façon homogène par la partition. Dans ce cas les centres peuvent être considérés comme des bons représentants non seulement des groupes mais également de l'ensemble initial de points. Or dans ce cas précis, on observe que la distance séparant chaque centre et son plus proche vaut environ deux fois le rayon moyen. On peut donc dire que la configuration de la figure 4.4(a) est celle de l'échantillon



(a) Partition idéale (b) Partition non homogène

FIGURE 4.4: Visualisation de la différence entre deux partitionnements de l'espace conduisant au même rayon moyen et au même résultat de SDNN

qui produit le critère SDNN = $k \times 2 \times Rmoyen$. Cependant ce critère ne suffit pas pour assurer un échantillon divers.

En effet, il convient de nuancer le critère SDNN en fonction de la distance minimum et de la distance maximum entre deux centres proches. En effet, le schéma 4.4(b) montre un exemple où une partition différente, du même espace et du même jeu de points, conduirait au même rayon moyen et au même résultat du critère SDNN [1]. Or ce partitionnement donne des centres qui sont de moins bons représentants de la dispersion du jeu dans l'espace. Si on ne regarde que la valeur du critère SDNN, on pourrait conclure à un échantillon divers comme pour le schéma 4.4(a). Mais si on prend en compte la distance minimum (SDNN Min : très faible, inférieure à 1 fois le rayon moyen) et la distance maximum (SDNN Max : égale à 4 fois le rayon moyen), alors on peut dire que malgré une valeur correcte du critère SDNN, le quadrillage dans l'espace n'est pas idéal, l'échantillon n'est pas aussi divers que l'idéal attendu (c'est à dire le schéma 4.4(a)) et surtout les individus sélectionnés couvrent plus les extrémités de l'ensemble de points que la partie centrale. Or c'est un écueil que nous avons déjà évoqué pour bon nombre de méthodes de sélection 1 et que nous souhaitons éviter.

Ensuite, le rayon d'un groupe est régi par le point central de ce groupe et un seul autre point : celui qui dicte le rayon du groupe (le plus éloigné du centre). Il se peut donc, comme le montre le schéma 4.5, qu'en traçant les sphères englobant les points des groupes, celles-ci se chevauchent (précisons que malgré cet effet "graphique", les points ne sont affectés qu'à un seul groupe). Dans ce cas, on remarque alors que malgré une bonne partition de l'espace, le critère SDNN peut être inférieur à $k \times 2 \times Rmoyen$. On cherchera donc à maximiser le critère SDNN tout en maximisant également le critère SDNN Min.

Enfin dans le cas de jeux comportant des outliers, certains centres sont très éloignés de leur centre le plus proche. La valeur de SDNN Max pourra donc être supérieure à $2 \times Rmax$. On cherchera alors à maximiser le critère SDNN Max. En effet, si ce critère est élevé, cela signifiera que les outliers sont des objets sélectionnés dans l'échantillon. En

1. Le rayon moyen et le critère SDNN sont calculés sur la base d'unités arbitraires communes aux deux schémas où 1R = 2Unités dans le schéma 4.4(a). Pour les deux figures le rayon moyen vaut 2Unités et SDNN vaut 20 unités

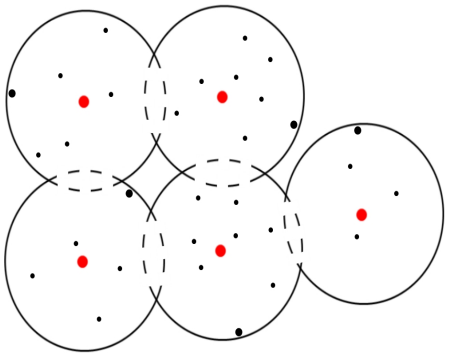


FIGURE 4.5: Visualisation d'une partition chevauchante

effet nous souhaitons que les outliers soient représentés même s'ils ne doivent pas être les seuls individus constituant l'échantillon, puisqu'ils font partie de l'espace chimique.

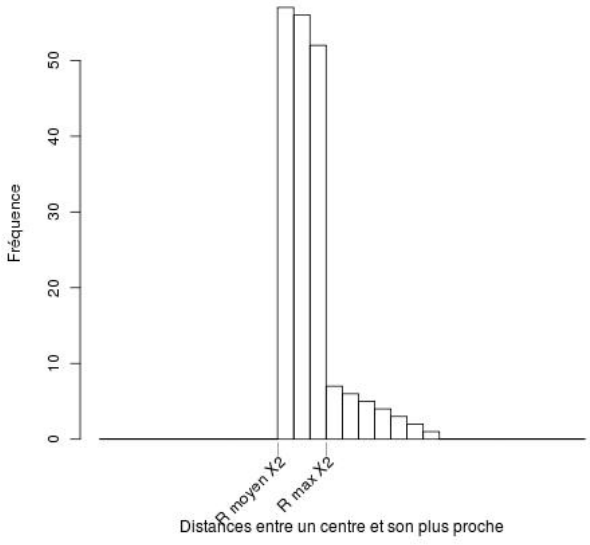


FIGURE 4.6: Histogramme théorique pour le critère SDNN : distribution des distances entre chaque centre et son centre le plus proche

Interprétation graphique L'histogramme de la figure 4.6 reflète la distribution des distances entre chaque centre et son centre le plus proche que l'on souhaiterait approcher dans l'idéal. Comme nous l'avons vu précédemment avec la figure 4.4(a), dans l'idéal, ces distances sont toutes supérieures à $2 \times Rmoyen$. De plus pour un quadrillage homogène de l'espace, nous avons vu avec les figures 4.4(a) et 4.4(b) montrant respectivement une bonne et une mauvaise partition de l'espace, que la majorité des distances entre un centre et son plus proche doivent être équivalentes à $2 \times Rmoyen$ voire $2 \times Rmax$. Ensuite nous avons vu qu'il est possible que certaines valeurs soient inférieures à $2 \times Rmoyen$, dans ce cas, nous minimiserons le nombre de ces distances. Enfin lorsque le jeu comporte des outliers, nous

avons vu que certaines distances peuvent être très supérieures à $2 \times Rmax$. Cependant le nombre de ces distances doit être faible et en rapport avec le nombre d'outliers.

Conclusion On cherche à maximiser le critère SDNN (Somme des dissimilarités entre chaque centre et son plus proche) et le critère SDNN Min. Le critère SDNN Max pourra être élevé et supérieur à $2 \times Rmax$ si le jeu initial comporte des outliers.

Enfin pour l'histogramme de la distribution de ces distances, on cherche à minimiser la partie se situant avant $2 \times Rmoyen$ et à maximiser la partie comprise entre $2 \times Rmoyen$ et $2 \times Rmax$. La partie se trouvant après $2 \times Rmax$ doit être présente dans une moindre mesure dans le cas d'un jeu constitué d'outliers.

4.2.2.2 Représentativité

Pour évaluer la représentativité, nous calculons les distances entre chaque molécule de l'ensemble de molécules de départ et leur molécule la plus proche dans l'échantillon sélectionné, nous obtenons les quatre indices suivant :

- Somme des dissimilarités entre chaque molécule du jeu de départ et son plus proche voisin dans l'échantillon (respectivement son centre le plus proche) notée SDDep
- Dissimilarité minimum entre une molécule du jeu de départ et sa plus proche dans l'échantillon (son centre le plus proche) notée SDDep Min
- Dissimilarité maximum entre une molécule du jeu de départ et sa plus proche dans l'échantillon (son centre le plus proche) notée SDDep Max
- Ecart type des dissimilarités entre chaque molécule du jeu de départ et son centre le plus proche noté EC SDDep

Où on définit Dep par :

$$Dep = \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1...k}{ArgMin}\ d(m_i, c_j^*),\ m_i \neq c_j^*\}$$

Ceci signifie que l'on considère les paires d'individus m_i et c_j^*, le premier appartenant à l'ensemble des molécules $\mathcal{M}$ et le second à l'échantillon ou ensemble des centres C^* et c_j^* étant le centre le plus proche de m_i. Enfin la molécule m_i ne doit jamais être un centre pour ne pas calculer leur distance à eux-même. La somme SDDep permet d'exprimer la représentativité de l'échantillon divers C^*. Les distances entre chaque molécule et sa molécule du jeu de départ $\mathcal{M}$ la plus proche dans l'échantillon C^* sont additionnées.

Définition 4.9 *La **somme des dissimilarités entre une molécule et son centre le plus proche SDDep** s'exprime ainsi :*

$$SDDep(\mathcal{M}, \mathcal{C}^*) = \sum_{i=1...n} \underset{j=1...k, j\neq i}{Min} d(m_i, c_j^*)$$

SDDepMin est la plus petite distance entre une molécule et son représentant dans l'échantillon divers C^*.

Définition 4.10 *La **dissimilarité minimum entre une molécule et son centre le plus proche SDDep Min** est défini ainsi :*

$$SDDepMin(\mathcal{M}, \mathcal{C}^*) = \underset{i=1...n}{Min} \underset{j=1...k, j\neq i}{Min} d(m_i, c_j^*)$$

SDDepMax est la plus grande distance entre une molécule et son représentant dans l'échantillon divers C^*.

Définition 4.11 *La **dissimilarité maximum entre une molécule et son centre le plus proche SDDep Max** est donné par :*

$$SDDepMax(\mathcal{M}, \mathcal{C}^*) = \underset{i=1...n}{Max} \underset{j=1...k, j\neq i}{Min} d(m_i, c_j^*)$$

Cette formule est la même que celle du rayon maximum, ceci s'explique par le fait que ces deux critères possèdent la même définition. Nous avons donc la propriété suivante :

$$SDDepMax(\mathcal{M}, \mathcal{C}^*) = RayonMax(\mathcal{C}) = \rho(\mathcal{C})$$

Définition 4.12 *L'**écart type** des dissimilarités au centre le plus proche EC SDDep se calcule ainsi :*

$$ECSDDep(\mathcal{M}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (d(m_i^l, c_l^*) - \overline{d(m, c^*)})^2}$$

où $\overline{d(m, c^*)}$ est la distance moyenne entre toutes les molécules et leur centre le plus proche.

Interprétation numérique Nous calculons ici les distances entre chaque molécule du jeu initial et son centre le plus proche (c'est à dire son représentant dans l'échantillon). La somme de ces distances nous permet d'évaluer la représentativité de l'échantillon. En effet, le schéma 4.4(a) montre une partition menant à un échantillonnage divers de l'espace. Sur ce schéma, on observe que toute molécule est à une distance maximum de son centre d'environ le rayon moyen. La molécule la plus éloignée est à une distance de un rayon max.

En revanche, le schéma 4.4(b) montre une partition avec un rayon moyen équivalent mais les molécules se trouvant dans le plus gros groupe sont pour la plupart plus éloignées de leur représentant (leur centre) que la valeur du rayon moyen. La distance maximum entre une molécule et son centre devient alors plus élevée puisque le rayon maximum est plus grand que pour la partition de la figure 4.4(a).

On souhaite donc minimiser les critères SDDep et SDDep Min. Mais on souhaite également minimiser SDDep Max (= au Rayon Max.).

Interprétation graphique L'histogramme 4.7 reflète la distribution des distances entre chaque molécule du jeu initial et son centre le plus proche que l'on souhaiterait approcher dans l'idéal. On a vu que la majorité des distances doivent se trouver entre 0 et la valeur du rayon moyen. Quelques unes peuvent se trouver entre le rayon moyen et le rayon maximum. Aucune distance ne doit se trouver après ce rayon maximum.

Conclusion Pour obtenir un échantillon représentatif, on souhaite minimiser les critères SDDep (Somme des distances entre chaque molécule du jeu initial et son centre le plus proche), SDDep Min et SDDep Max.

Dans le même temps, on souhaite maximiser le nombre de distances comprises entre 0 et le rayon moyen. A valeur du critère SDDep équivalent, un échantillon sera considéré plus représentatif si les distances entre chaque molécule du jeu initial et leur représentant sont plus proches de 0 que pour un autre échantillon.

4.2.2.3 Dissimilarité totale de l'échantillon

Nous évaluons la dissimilarité de l'échantillon avec les quatre indices qui suivent :

– Somme des dissimilarités entre toutes les molécules de l'échantillon deux à deux notée SomDissimTot et SDTot

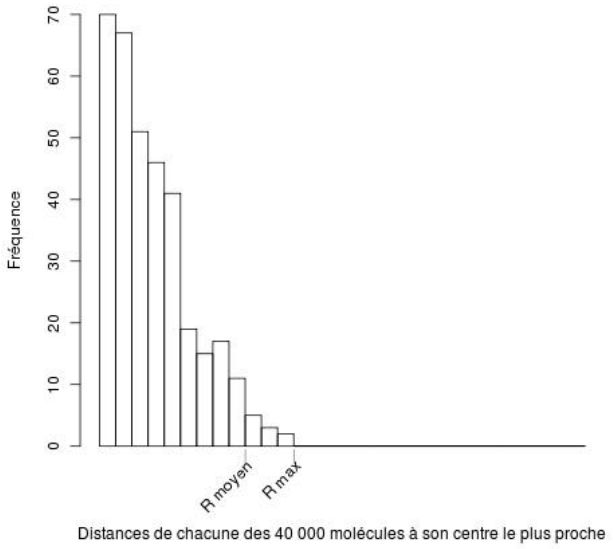


FIGURE 4.7: Histogramme théorique pour le critère SDDep : distribution des distances entre chaque molécule $m_i \in \mathcal{M}$ et son centre le plus proche $c_j^* \in C^*$

- Dissimilarité minimum entre une molécule de l'échantillon et une autre notée SDTot Min
- Dissimilarité maximum entre une molécule de l'échantillon et une autre notée SDTot Max
- Ecart type des dissimilarités : EC SDTot

Où Tot est défini par :

$$Tot = \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$$

Ce qui signifie que l'on considère les paires de molécules c_i^* et c_j^* sachant qu'elles appartiennent à l'échantillon ou ensemble des centres C^* et qu'aucun d'entre eux n'est comparé avec lui-même. SDTot est le critère le plus couramment utilisé dans la littérature pour évaluer la diversité d'un échantillon.

Définition 4.13 *La **Somme des dissimilarités entre toutes les molécules de l'échantillon** est défini par la formule suivante :*

$$SDTot(C^*) = \sum_{l=1...k} \sum_{j=1...k} d(c_l^*, c_j^*)$$

SDTot Min, parmi toutes les distances entre molécules de l'échantillon, est la plus petite distance entre deux centres.

Définition 4.14 *La **dissimilarité minimum** est défini comme suit :*

$$SDTotMin(C^*) = \underset{l\ et\ j=1...k\ et\ l\neq j}{Min} d(c_l^*, c_j^*)$$

Ce critère est donc égal à SDNN Min.

SDTot Max, parmi toutes les distances entre molécules de l'échantillon, est la plus grande distance entre deux centres.

Définition 4.15 *La **dissimilarité maximum** se calcule ainsi :*

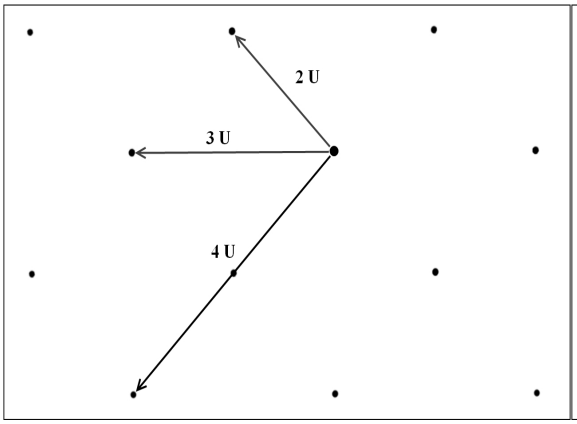
$$SDTotMax(C^*) = \underset{l \ et \ j=1...k}{Max} d(c_l^*, c_j^*)$$

Ce critère donne une indication sur la distance entre les deux extrêmes de l'échantillon. Il nous renseigne donc sur l'amplitude de l'échantillon dans l'espace chimique et donc également sur le fait que des outliers soient présents ou non dans cet échantillon.

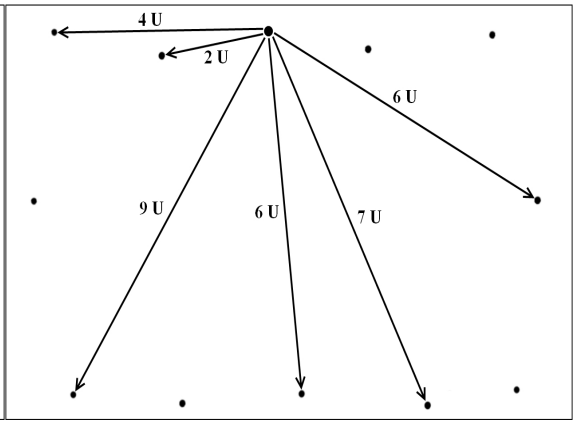
Définition 4.16 *L' **écart type** des dissimilarités dans l'échantillon EC SDTot se calcule ainsi :*

$$ECSDTot(\mathcal{M}) = \sqrt{\frac{1}{ntot} \sum_{i \ et \ j=1}^{n} tot(d(c_i^*, c_j^*) - \overline{d(c_i^*, c_j^*)})^2}$$

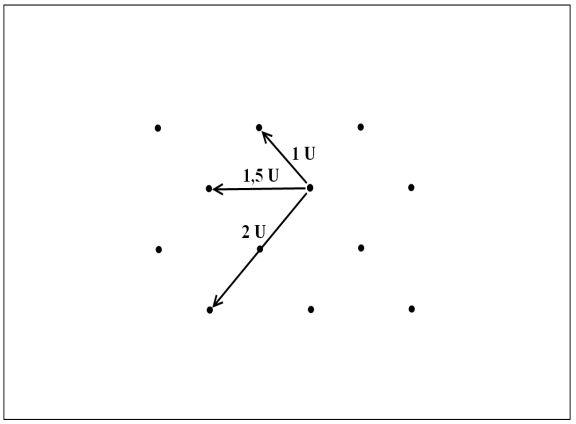
Où ntot est le nombre de dissimilarités calculées pour la somme : $ntot = k!$



(a) Echantillon divers SDTot=33 Unités $\times k$



(b) Echantillon d'outliers SDTot =62 Unités $\times k$



(c) Echantillon trop concentré SDTot=16.5 Unités $\times k$

FIGURE 4.8: Visualisation de la dissimilarité totale de différents échantillons (les chiffres sont expliqués par la suite)

Interprétation numérique On calcule les distances entre chaque centre (objet de l'échantillon) deux à deux. La somme de ces distances est le critère généralement maximisé pour sélectionner un échantillon divers dans le domaine de la chemo-informatique. Les schémas 4.8(a), 4.8(b) et 4.8(c) montrent trois échantillons différents que l'on pourrait

obtenir d'un même jeu de données. Pour chaque figure, les distances en Unité arbitraires sont indiquées et permmetent le calcul de la somme des dissimilarités totale. La figure 4.8(a) montre un échantillon divers, la figure 4.8(b) montrant un échantillon ne couvrant que les extrêmes de l'espace et la figure 4.8(c) montre un échantillon très concentré au centre de l'espace. Rappelons que dans l'idéal on souhaite obtenir un échantillon dont les propriétés s'approchent de celles de la figure 4.8(a).

On peut remarquer que l'échantillon le moins divers (cf. Figure 4.8(c)) donne une faible valeur du critère SDTot [2] ($16.5U \times k$).

En revanche, l'échantillon que l'on considère comme étant le plus divers (cf. Figure 4.8(a)) n'est pas celui qui donne la valeur la plus élevée du critère SDTot [3] ($33U \times k$). En effet, ce critère comporte un biais. Le maximiser peut revenir à ne favoriser que les extrêmes (cf. Figure 4.8(b), SDTot [4] $\simeq 62U \times k$). On souhaite certes que les outliers soient représentés dans l'échantillon, mais on souhaite également que le centre de l'espace soit représenté dans l'échantillon. Une somme des dissimilarités totale dans l'échantillon maximale ne sera donc pas forcément un bon indicateur de la diversité de celui-ci. Cependant on souhaite que SDTot Max soit assez élevé pour indiquer que des extrêmes sont couverts.

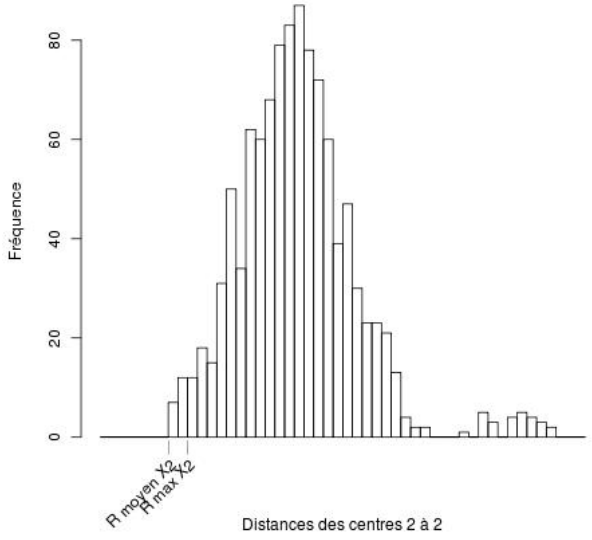


FIGURE 4.9: Histogrammes théorique pour le critère SDTot : distribution des distances entre tous les centres deux à deux

Interprétation graphique L'histogramme 4.9 reflète la distribution des distances entre toutes les molécules de l'échantillon que l'on souhaiterait approcher dans l'idéal. On s'attend à un histogramme en cloche homogène indiquant que les distances entre centres vont en augmentant de manière régulière. Ceci traduira un quadrillage homogène de l'espace. Dans l'idéal, les distances les plus faibles ne seront pas inférieures à $2 \times Rmoyen$. On autorisera une queue à droite de l'histogramme dans le cas d'un jeu avec outliers.

2. $SDTot \simeq (4 \times 1U + 3 \times 1.5U + 4 \times 2U) \times k$
3. $SDTot \simeq (4 \times 2U + 3 \times 3U + 4 \times 5U) \times k$
4. $SDTot \simeq (2 \times 2U + 2 \times 4U + 3 \times 6U + 2 \times 7U + 2 \times 9U) \times k$

Conclusion On cherche à maximiser SDTot Max. Dans le même temps une valeur élevée du critère SDTot ne sera pas un indicateur d'une bonne diversité de l'échantillon. Pour cela, on cherchera à approcher la forme de l'histogramme vu précédemment, tout en tentant d'obtenir un centre de la cloche la plus homogène possible.

4.3 Distances inter-molécules

Pour évaluer la similarité/dissimilarité entre molécules, il faut utiliser une mesure de distance. Nous avons vu dans le premier chapitre qu'il existe plusieurs possibilités pour cela. Etant donné que nous travaillons avec les descripteurs physico-chimiques qui sont des valeurs réelles, nous avons choisi d'utiliser la distance euclidienne pour comparer les molécules entre elles.

Pour l'étude des résultats, nous présentons dans un premier temps une étude des différentes méthodes pour l'échantillonnage de 1000 molécules d'un jeu avec outliers. Puis nous étudierons l'impact du jeu initial, ainsi que celui de la taille de l'échantillon sur les résultats des méthodes. Ensuite nous verrons si les jeux sans outliers ont une influence sur la qualité de l'échantillonnage pour chaque méthode. Et enfin nous verrons quel impact peut avoir le jeu initial sans outlier sur la qualité des échantillons.

4.4 Etude des résultats pour l'échantillonnage d'un jeu avec outliers (jeu : J1 et k = 1000)

Pour chacune des méthodes nous étudions les tendances qui ressortent pour chaque critère en comparant les résultats de différents paramétrages pour un même échantillonnage (1000 molécules) et un même jeu (J1 contenant 40 000 molécules). Puis nous comparons les performances des méthodes les unes par rapport aux autres.

4.4.1 Sélection Aléatoire (M1)

La première méthode de sélection à être testée est la sélection aléatoire uniforme sans remise.

4.4.1.1 Stabilité sur 10 runs

Rappelons que pour cette méthode 10 runs différents ont été effectués pour chaque échantillonnage afin de s'assurer que les résultats sont reproductibles d'un tirage à l'autre. Sur ces 10 runs les échantillons se révèlent toujours différents, c'est à dire qu'aucune molécule n'est sélectionnée dans plus d'un jeu. Certaines valeurs comme SDTot et SDTot Max (cf. Tableau 4.2) varient beaucoup d'un run à l'autre et cela peut s'expliquer par la sélection d'un outlier différent à chaque run. Cependant l'ordre de grandeur reste le même. Les autres critères restent stables, nous comparons donc le résultat de la moyenne des runs pour chaque jeu et pour chaque taille d'échantillon dans la suite de l'étude.

	Critère	Résultats pour chacun des 10 runs pour M1
	Rayon max.	127.69 158.18 164.93 172.01 165.62 149.33 165.90 164.99 159.23 164.66
	Rayon moyen	14.03 15.25 14.12 14.50 14.87 14.24 14.06 13.87 14.64 14.52
	EC Rayon	8.03 10.96 9.95 10.47 10.10 8.22 9.36 8.48 9.50 10.72
	EC Ind	32.04 31.72 30.49 30.34 31.76 30.50 32.68 32.74 29.53 31.25
NN [a]	Somme	8.81×10^3 8.51×10^3 8.50×10^3 8.59×10^3 8.60×10^3 8.57×10^3 8.74×10^3 8.63×10^3 8.58×10^3 8.61×10^3
	Min	5.01 1.66 3.59 4.02 2.27 4.11 4.30 4.18 4.35 3.64
	Max	57.85 27.56 25.90 44.45 82.14 59.02 59.58 43.29 48.13 52.12
	EC	3.37 2.29 2.23 2.78 3.17 2.90 2.99 2.82 3.04 2.98
Dep [b]	Somme	334.77×10^3 336.03×10^3 334.91×10^3 335.36×10^3 335.73×10^3 335.21×10^3 335.77×10^3 334.76×10^3 335.66×10^3 335.24×10^3
	Min	0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01
	Max	127.69 158.18 164.93 172.01 165.62 149.33 165.90 164.99 159.23 164.66
	EC	3.03 3.39 3.33 3.29 3.21 3.19 3.23 3.14 3.29 3.23
Tot [c]	Somme	9.39×10^6 9.027×10^6 9.022×10^6 9.018×10^6 9.046×10^6 9.091×10^6 9.26×10^6 9.088×10^6 9.005×10^6 9.13×10^6
	Min	5.01 1.66 3.59 4.02 2.27 4.11 4.30 4.18 4.35 3.64
	Max	104.40 62.52 53.77 83.72 89.79 80.68 73.23 76.12 80.64 74.67
	EC	7.65 5.61 5.58 6.17 6.12 6.22 6.13 6.16 6.34 6.14
	Nb singletons	2.00 4.00 5.00 2.00 7.00 2.00 2.00 4.00 1.00 5.00

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.2: Tableau des résultats pour les 10 runs d'un tirage aléatoire

4.4.1.2 Résultats sur la moyenne des 10 runs (jeu 1 et k = 1000)

Rayon et densité des groupes Le rayon maximum est très élevé (159.25, cf. Tableau 4.3) par rapport au rayon moyen (14.41). Ceci signifie qu'il existe de grandes disparités dans les rayons des groupes. On le confirme d'ailleurs avec la valeur de l'écart-type du nombre d'individus par groupe qui est faible (31.31). En effet, cette valeur signifie qu'en moyenne il y a un écart de 31 individus contenus dans les groupes. Or nous savons que notre jeu de données est réparti de façon non homogène dans l'espace des descriptions notamment avec des valeurs extrêmes (cf. chapitre description des données). Certaines parties de l'espace étant peu représentées, comme nous l'avons vu dans la section 4.2.1.2 on peut dire que les groupes doivent avoir des rayons de tailles très disparates pour contenir un nombre homogène d'individus.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind
M1	159.25	14.41	9.58	31.31

TABLE 4.3: Résultats du critère Rayon pour un tirage aléatoire (M1) - Jeu 1, k=1000

Diversité/Recouvrement de l'espace Nous étudions ici les distances entre chaque centre de groupe (aussi appelé représentant) et son centre le plus proche.On remarque que la distance minimum entre deux centres est très faible : de 3.7 (cf. Tableau 4.4), soit le quart du rayon moyen. De plus la distance maximum entre deux centres , 50.0 n'est pas très élevée (seulement 3.5 fois le rayon moyen et très inférieure au rayon maximum). Nous notons que le rayon du plus grand groupe est trois fois plus élevé que la distance entre les deux centres les plus éloignés. Ceci peut s'expliquer par le fait que les représentants choisis aléatoirement ne sont pas de bons centres de groupes. Certains individus comme vu plus haut (cf. paragraphe 4.4.1.2) sont donc très éloignés de leur représentant alors que ceux-ci peuvent être proches entre eux. Nous le confirmerons dans le paragraphe suivant en étudiant les critères de représentativité de l'échantillon.
Enfin lorsqu'on observe la distribution des distances entre chaque représentant et son plus proche (cf. Figure 4.10), on remarque d'une part qu'il existe beaucoup de centres très proches (distance comprise entre 0 et 5) et d'autre part la majorité des centres ont un voisin proche à une distance plus faible que le rayon moyen des groupes. Ceci confirme donc nos observations précédentes qui indiquait une mauvaise répartition des représentants les uns par rapport aux autres.

NN [a]	Somme	Min	Max	EC
M1	8.61×10^3	3.71	50.00	2.86

a. $\text{NN} = \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

TABLE 4.4: Résultats du critère SDNN pour un tirage aléatoire (M1) - Jeu 1, k=1000

Représentativité Comme nous l'avons vu précédemment et notamment avec le rayon maximum, la molécule la plus loin de son représentant en est très éloignée (distance de 159.25 pour SDDep Max = Rayon Max, cf. Tableau 4.5). Lorsqu'on étudie la distribution des distances entre chaque molécule et son représentant le plus proche (cf. Figure 4.11), on remarque que la majorité des molécules sont très proches de leur représentant

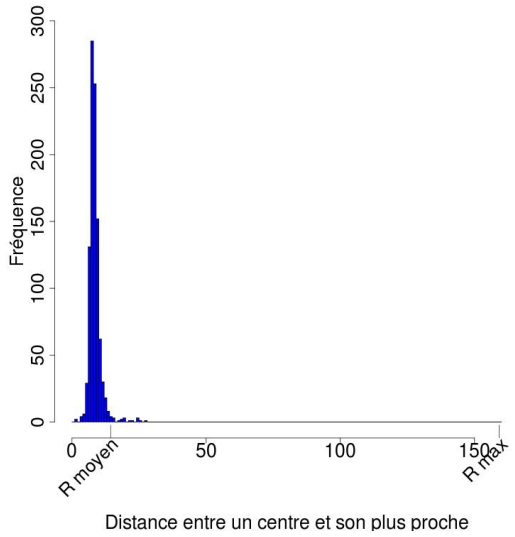


FIGURE 4.10: Pour M1 tirage aléatoire, Distribution des dissimilarités entre chaque centre et son centre le plus proche (SDNN)

(distance inférieure au rayon moyen). Une bonne partie des molécules est à deux fois le rayon moyen de leur représentant. Le tirage aléatoire permet donc d'obtenir un échantillon globalement représentatif du jeu de départ. Nous observons cependant des distances très élevées (supérieures à 50) (cf. Figure 4.11(b)) marquant quelques individus mal représentés par l'échantillon.

Dep [a]	Somme	Min	Max	EC
M1	335.34×10^3	0.01	159.25	3.23

a. $\text{Dep} = \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

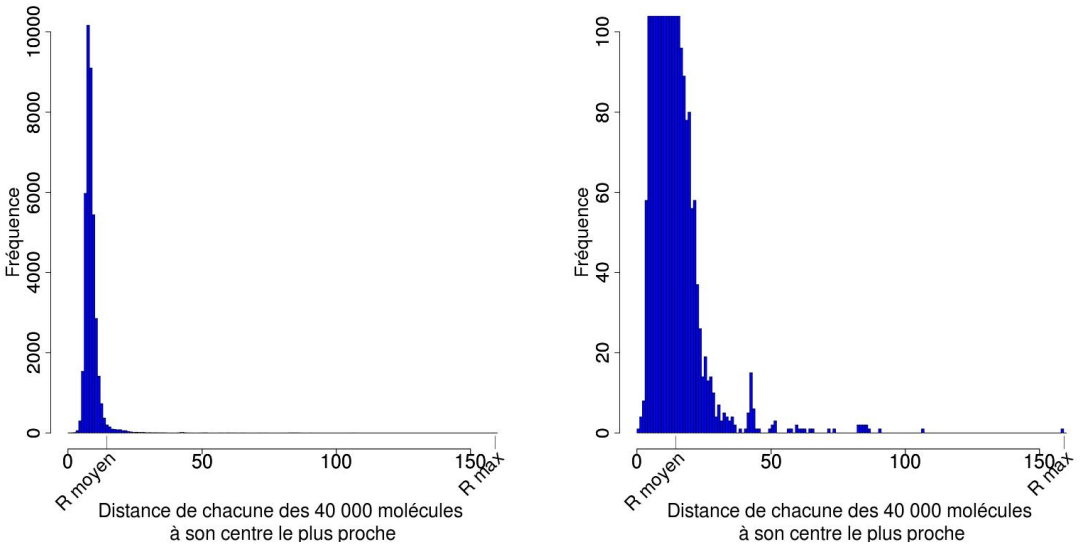
TABLE 4.5: Résultats du critère SDDep pour un tirage aléatoire (M1) - Jeu 1, k=1000

Dissimilarité totale dans l'échantillon Encore une fois on observe que la distance maximum entre deux centres - SDTot Max- est très faible (77.95, cf. Tableau 4.6) au regard du rayon maximum et de la dispersion connue du jeu de données dans l'espace des descriptions. En effet, si les extrêmes de cet espace sont couverts par l'échantillon, on s'attend à ce que la distance maximum entre deux centres soit très élevée. On observe d'ailleurs que toutes les distances entre les centres deux à deux (cf. Figure 4.12) sont très concentrées autour du rayon moyen. Ceci confirme donc une faible dispersion des représentants dans l'espace.

Tot [a]	Somme	Min	Max	EC
M1	9.10×10^6	3.71	77.95	6.21

a. $\text{Tot} = \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.6: Résultats du critère SDTot pour un tirage aléatoire (M1) - Jeu 1, k=1000



(a) M1 SDDep histogramme complet

(b) M1 SDDep échelle réduite

FIGURE 4.11: Pour M1 tirage aléatoire, Distribution des dissimilarités entre chaque molécule et son centre le plus proche (SDDep)

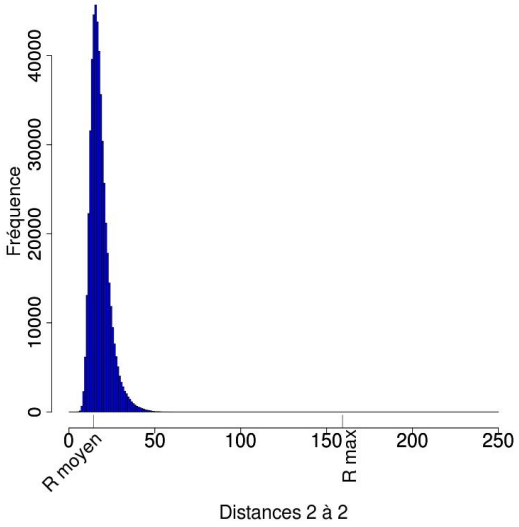
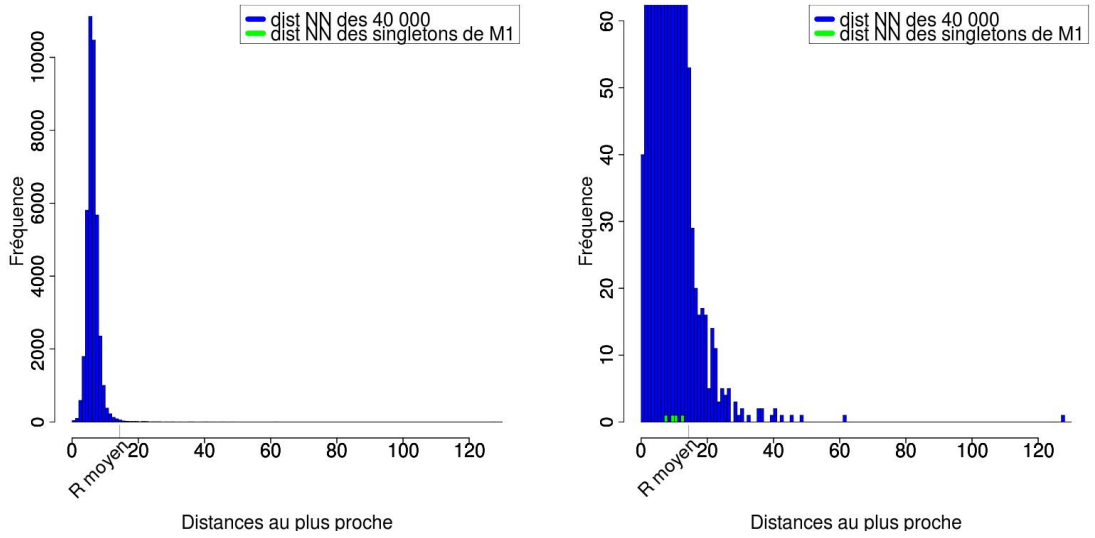


FIGURE 4.12: Pour M1 tirage aléatoire, Distribution des dissimilarités entre tous les centres (SDTot)

Singletons Enfin le tirage aléatoire ne donne que très peu de singletons (groupes contenant un seul individu), cf. Tableau 4.7. En observant la distance entre chaque singleton et sa molécule la plus proche (cf. Figure 4.13 en vert), on remarque que ces singletons sont très proches de leur molécule voisine (aux alentours du rayon moyen). Et de plus ces distances ne sont pas parmi les plus grandes dans cette distribution de distances au plus proche voisin. Ces singletons ne sont donc pas de bons indicateurs d'outliers.

	Singletons
M1	3.40

TABLE 4.7: Résultats du critère Singletons pour un tirage aléatoire (M1) - Jeu 1, k=1000



(a) M1 Singletons histogramme complet

(b) M1 Singletons échelle réduite

FIGURE 4.13: Pour M1 tirage aléatoire, Distribution des dissimilarités entre chaque molécule et sa molécule la plus proche et projections de ces distances pour les singletons

Conclusion Le tirage aléatoire donne donc des représentants peu dispersés dans l'espace des descriptions, contrairement au jeu de départ, et très représentatifs des zones les plus denses couvertes par ce jeu. Ceci est cohérent avec la théorie du tirage aléatoire uniforme. Rappelons qu'une propriété de ce tirage est qu'il donne un échantillon représentatif (au sens statistique du terme) de la distribution initiale au sens que les zones de l'espace les plus denses sont mieux représentées que les zones peu denses.

De plus ces représentants également centres de groupes, induisent des groupes de taille très variable contenant un nombre homogène d'individus. Nous sommes loin de notre objectif voulant quadriller l'espace de façon homogène avec des groupes de rayons similaires. Enfin les outliers ne sont pas ou peu présents dans l'échantillon.

Le tirage aléatoire ne donne donc pas un échantillon très divers.

4.4.2 Sélection avec *k*-center (M2)

Rappelons que pour cette méthode nous testons deux initialisations différentes : M2A est initialisée avec un tirage aléatoire des centres et M2B est initialisée avec les centres résultant de la sélection par la méthode FFT.

4.4.2.1 Stabilité sur 10 runs

Pour les initialisations aléatoires, 10 runs sont effectués afin de s'assurer que les résultats obtenus après cette initialisation soient comparables quel que soit le tirage effectué au départ. Tous les runs engendrent des valeurs similaires pour les différents critères (cf.

Tableau 4.8). Nous nous fixons donc sur l'étude de la moyenne des runs.

	Critère	Résultats des 10 runs pour M2A
	Rayon max.	12.66 12.66 12.74 12.70 12.65 12.69 12.68 12.69 12.66 12.73
	Rayon moyen	11.86 11.77 11.84 11.79 11.80 11.80 11.79 11.80 11.82 11.84
	EC Rayon	0.85 1.03 1.04 1.10 0.98 1.05 1.04 1.02 1.06 1.00
	EC Ind	194.73 162.31 171.06 181.68 178.22 176.77 180.88 171.32 187.77 172.87
NN [a]	Somme	14.04×10^3 13.99×10^3 14.05×10^3 14.09×10^3 14.05×10^3 14.06×10^3 14.06×10^3 14.05×10^3 13.99×10^3 14.10×10^3
	Min	4.99 7.18 7.89 7.20 7.88 7.39 7.70 5.06 6.95 7.48
	Max	127.69 127.69 127.69 127.69 127.69 127.69 127.69 127.69 127.69 127.69
	EC	5.93 5.98 5.97 5.95 5.95 5.97 5.95 5.97 6.00 5.95
Dep [b]	Somme	364.10×10^3 360.97×10^3 361.89×10^3 363.07×10^3 359.51×10^3 361.85×10^3 359.21×10^3 366.21×10^3 359.70×10^3 364.91×10^3
	Min	0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10
	Max	12.66 12.66 12.74 12.70 12.65 12.69 12.68 12.69 12.66 12.73
	EC	1.89 1.89 1.89 1.89 1.88 1.89 1.89 1.89 1.91 1.88
Tot [c]	Somme	18.59×10^6 18.51×10^6 18.43×10^6 18.52×10^6 18.51×10^6 18.51×10^6 18.60×10^6 18.61×10^6 18.46×10^6 18.52×10^6
	Min	4.99 7.18 7.89 7.20 7.88 7.39 7.70 5.06 6.95 7.48
	Max	201.77 201.77 201.77 201.77 201.77 201.77 201.77 201.77 201.77 201.77
	EC	16.78 16.81 16.60 16.76 16.77 16.60 16.82 16.80 16.62 16.61
	Nb singletons	439.00 414.00 418.00 423.00 428.00 431.00 435.00 425.00 434.00 430.00

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.8: Tableau des résultats sur 10 runs pour la méthode k-center (M2A)

4.4.2.2 Résultats sur la moyenne des 10 runs (jeu 1 - k = 1000)

Rayon et densité des groupes On observe un rayon maximum faible (12.68, cf. Tableau 4.9) et très proche du rayon moyen (11.81) tant pour l'initialisation aléatoire (M2A) que pour l'initialisation avec FFT (M2B). Ceci indique un ensemble de rayons homogènes quelque soit le type d'initialisation utilisée. L'écart moyen du nombre d'individus par groupe (EC Ind) approche les 200 individus. Ceci confirme des groupes de rayons homogènes. Comme vu en section 4.2.1.2, certains groupes couvrant des zones plus denses que d'autres ; avec un écart des rayons homogènes on peut observer un écart important du nombre d'individus par groupe.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	
M2A	12.68	11.81	1.02	177.76	initialisation aléatoire
M2B	12.80	11.90	1.02	175.89	initialisation avec FFT

TABLE 4.9: Résultats du critère Rayon pour notre méthode k-center (M2) - Jeu 1, k=1000

Diversité/Recouvrement de l'espace On observe que les deux centres les plus proches sont à une distance inférieure de moitié au rayon moyen et maximum (6.97 et 6.09, cf. Tableau 4.10). Il existe donc des groupes plus petits que d'autres et donc des centres plus proches l'un de l'autre que peut le laisser croire le rayon moyen. Cependant les deux groupes les plus éloignés sont à une grande distance l'un de l'autre (127.69) indiquant qu'au moins un outlier doit être contenu dans l'échantillon. De plus nous voyons dans la distribution de ces distances (cf. Figure 4.14) que la plupart des centres ont un voisin au delà du rayon moyen voire du rayon maximum. Et par ailleurs, ces distances restent concentrées entre le rayon moyen et deux fois ce rayon. Ainsi on confirme des groupes de taille homogène par la méthode k-center, quadrillant l'espace de façon régulière comme nous l'attendons pour une sélection par diversité. Ces histogrammes nous permettent d'observer des distances plus élevées notamment supérieures à 40. Ces distances nous indiquent la sélection d'outliers. Enfin nous voyons que quelque soit l'initialisation du jeu de centres, les résultats sont similaires.

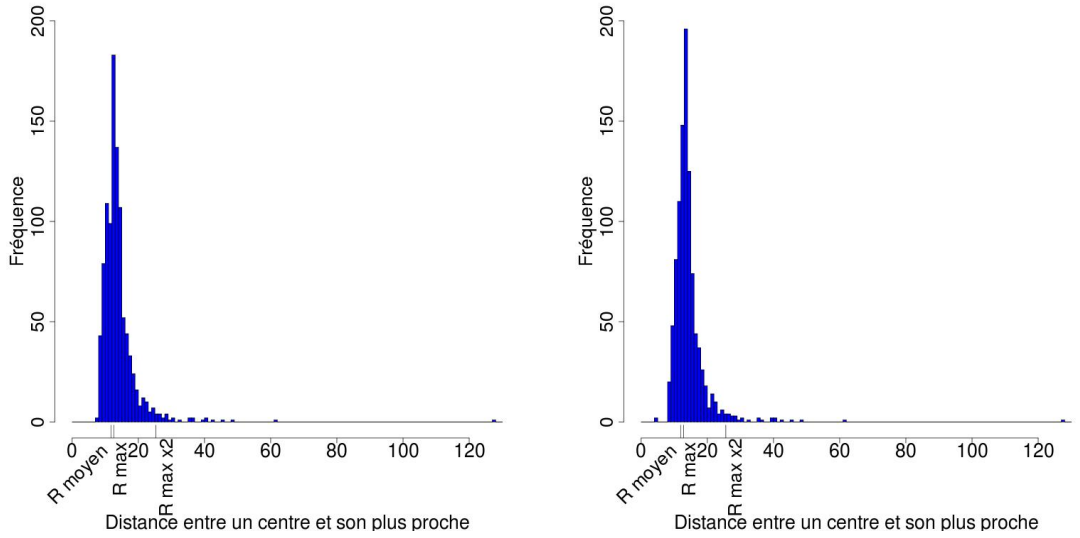
NN [a]	Somme	Min	Max	EC	
M2A	14.05×10^3	6.97	127.69	5.96	initialisation aléatoire
M2B	14.34×10^3	6.09	127.69	5.88	initialisation avec FFT

a. $\text{NN} = \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

TABLE 4.10: Résultats du critère SDNN pour notre méthode k-center (M2) - Jeu 1, k=1000

Représentativité Comme nous pouvons le voir dans le tableau 4.11, les molécules sont très proches de leur représentant (de 0.1 à 12.68-12.80 égal au rayon maximum). Cette tendance se confirme en observant la distribution des distances entre toute molécule et son représentant (cf. Figure 4.15). En effet, la majorité des distances se situe avant le rayon maximum voire est inférieure à 10. Cette méthode permet donc de sélectionner un échantillon représentatif des molécules du jeu de départ.

Dissimilarité totale dans l'échantillon Nous remarquons ici que la distance maximum entre deux centres (SDT Max= 201.77, cf. Tableau 4.12) est élevée au regard de la



(a) M2A SDNN (initialisation aléatoire)

(b) M2B SDNN (initialisation avec FFT)

FIGURE 4.14: Pour M2 k-center, Distribution des dissimilarités entre chaque centre et son centre le plus proche (SDNN)

Dep [a]	Somme	Min	Max	EC	
M2A	362.14×10^3	0.10	12.68	1.89	initialisation aléatoire
M2B	368.25×10^3	0.10	12.80	1.89	intialisation avec FFT

a. Dep $= \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

TABLE 4.11: Résultats du critère SDDep pour notre méthode k-center (M2) - Jeu 1, k=1000

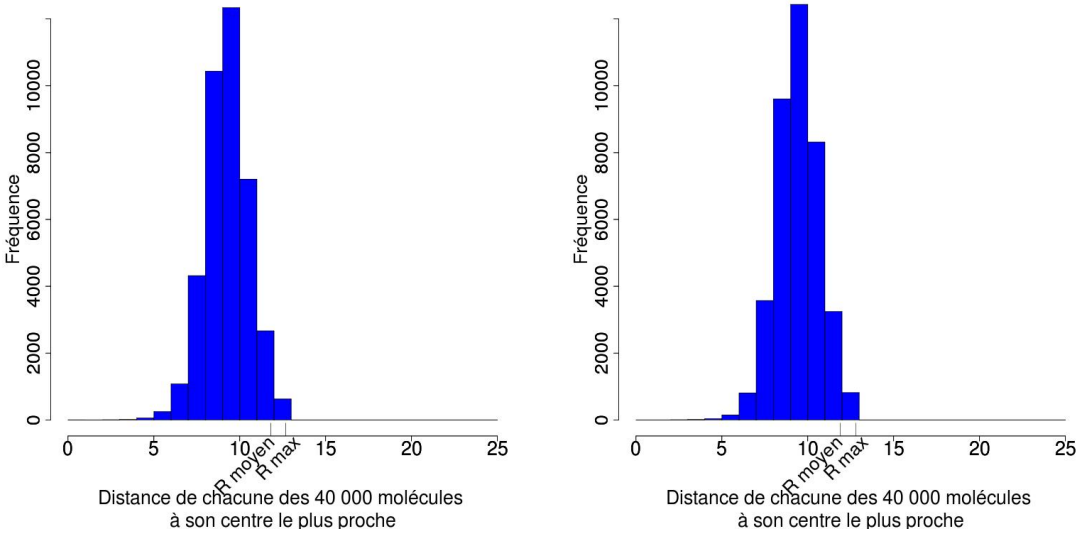
distance minimum. Cette distance maximale montre donc que les représentants couvrent les extrêmes de l'espace comme ils couvrent le centre de celui-ci. Cette observation se confirme par les histogrammes de la figure 4.16 où la plupart des distances entre centres deux à deux, se concentre entre 0 et 50 montrant ainsi la couverture de l'espace des descriptions de façon homogène. Mais également, les extrêmes semblent couverts car on observe des "îlots" de distances aux alentours de 100 et après 150 dénotant que certains centres sont très éloignés les uns des autres.

Tot [a]	Somme	Min	Max	EC	
M2A	18.52×10^6	6.97	201.77	16.72	initialisation aléatoire
M2B	18.51×10^6	6.09	201.77	16.55	initialisation avec FFT

a. Tot $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

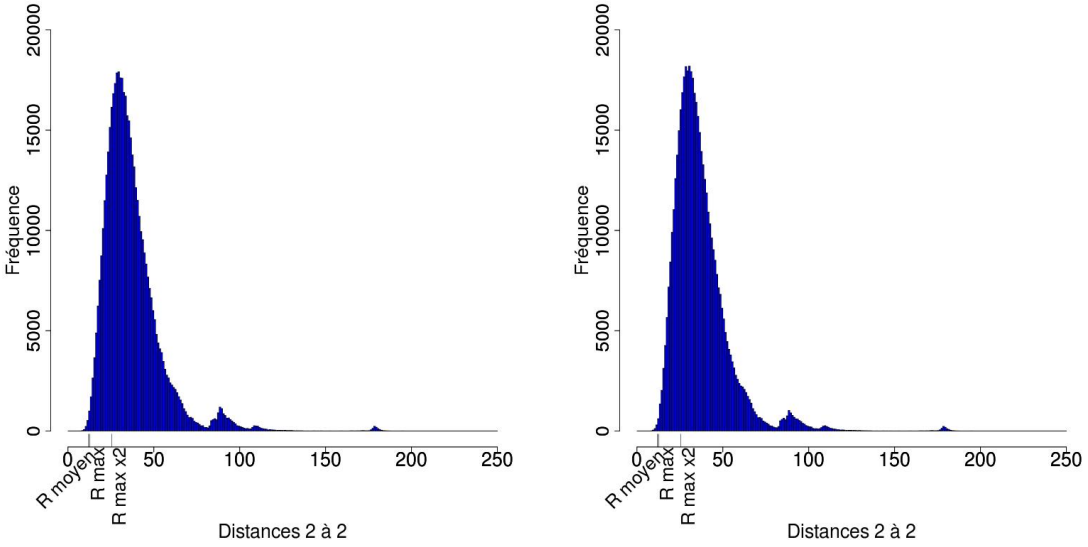
TABLE 4.12: Résultats du critère SDTot pour notre méthode k-center (M2) - Jeu 1, k=1000

Singletons Enfin cette méthode produit beaucoup de groupes ne contenant qu'un individu (environ 420, cf. Tableau 4.13). En observant les histogrammes de la figure 4.17, on remarque que la plupart des singletons ont des distances élevées avec leur molécule plus proche. De plus, toutes les distances entre une molécule et son plus proche qui sont



(a) M2A SDDep (b) M2B SDDep

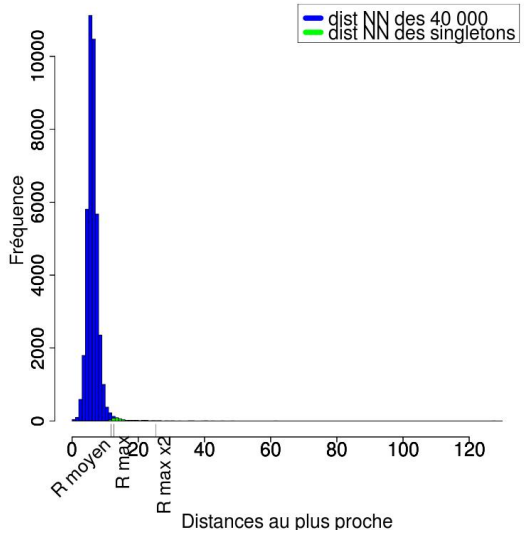
FIGURE 4.15: Pour M2 k-center, Distribution des dissimilarités entre chaque molécule et son centre le plus proche (SDDep)



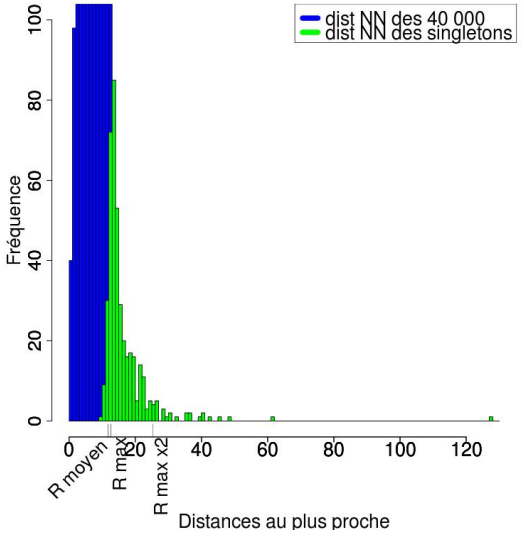
(a) M2A SDTot (initialisation aléatoire) (b) M2B SDTot (initialisation avec FFT)

FIGURE 4.16: Pour M2 k-center, Distribution des dissimilarités entre tous les centres (SDTot)

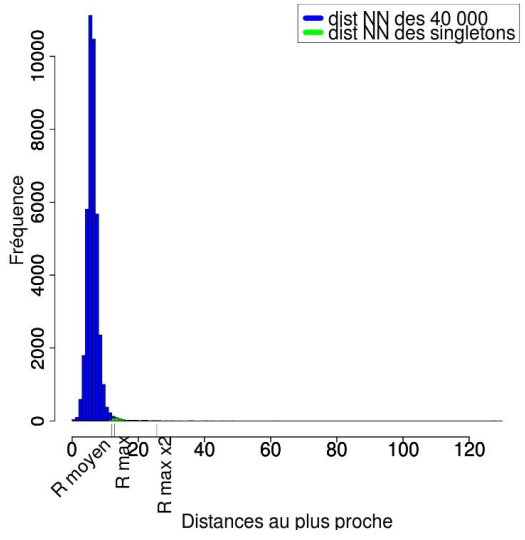
extrêmes sont couvertes par les distances entre les singletons et leur plus proche. La majorité d'entre eux sont donc de bons indicateurs d'outliers. Notons toutefois que la méthode initialisée avec FFT (M2B, cf. Figure 4.17(d)) produit quelques singletons dont la molécule la plus proche n'est qu'à 5 de distance (très inférieure au rayon maximum). De tels singletons ne sont pas de bons indicateurs d'outliers.



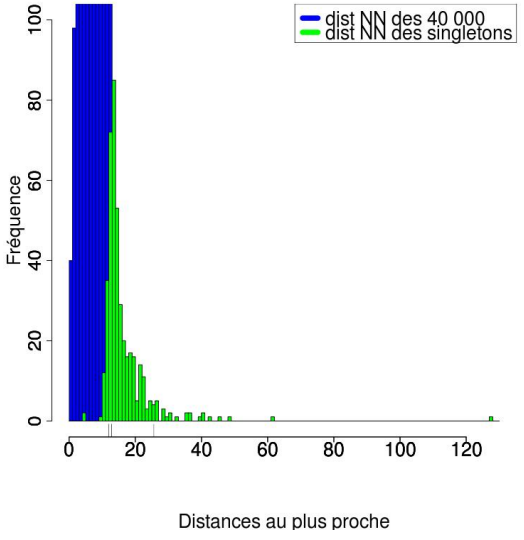
(a) M2A Singletons histogramme complet



(b) M2A Singletons échelle réduite



(c) M2B Singletons histogramme complet



(d) M2B Singletons échelle réduite

FIGURE 4.17: Pour M2 k-center, Distribution des dissimilarités entre chaque molécule et sa molécule la plus proche et projection de ces distances pour les singletons

	Singletons	
M2A	427.70	intialisation aléatoire
M2B	418.90	initialisation avec FFT

TABLE 4.13: Résultats du critère Singletons pour k-center (M2) - Jeu 1, k=1000

Conclusion Cette méthode k-center permet d'obtenir un échantillon représentatif de la diversité du jeu de départ. L'espace des descriptions semble être couvert homogènement. Les résultats sont en adéquation avec notre objectif : un rayon maximum faible, une faible dispersion de la taille des rayons et une bonne représentativité. Enfin les singletons peuvent être de bons indicateurs d'outliers.

Concernant le temps de calcul, pour une sélection de 1000 molécules, la méthode k-center met autant de temps quelque soit l'initialisation utilisée, soit environ 53 minutes en temps utilisateur.

4.4.3 Sélection avec k-medoïds (M3)

Nous rappelons que plusieurs initialisations ont été testées avec cette méthode : aléatoire ou avec les résultats de FFT.
Ensuite pour chacune d'elle, l'algorithme a utilisé soit des centres réels à chaque itération, soit des centres virtuels jusqu'à l'arrêt des itérations suivi d'un calcul des centres réels en fin d'algorithme. Nous obtenons donc 4 paramétrages différents de la méthode k-medoïds auxquels nous attribuons par la suite le code couleur indiqué dans le tableau 4.14 pour une meilleure compréhension :

M3A	initialisation
M3B	aléatoire
M3C	initialisation
M3D	avec FFT

TABLE 4.14: Code couleur des différents paramètres utilisés pour la méthode k-medoïds (M3) ; centres réels, centres virtuels

4.4.3.1 Stabilité sur 10 runs

Seules les variantes M3A et M3B sont concernées par l'initialisation aléatoire. Nous effectuons 10 initialisations différentes dans ce cas pour s'assurer de la reproductibilité des résultats quelque soit le tirage. Or les 10 runs de M3A engendrent des valeurs similaires pour les différents critères (cf. tableau 4.15). Nous obtenons le même genre de résultats pour M3B. Certes le rayon maximum a une certaine variance mais reste dans le même ordre de grandeur (entre 143 et 170). Nous nous fixons donc sur l'étude de la moyenne des runs pour la suite de l'analyse.

Critère		Résultats pour chacun des 10 runs pour M3A
Rayon max.		143.78 143.78 148.63 143.78 143.78 170.35 165.34 147.31 165.90 165.13
Rayon moyen		13.63 13.94 14.44 14.10 13.52 14.79 13.79 14.84 14.27 13.72
EC Rayon		7.98 9.05 10.52 8.87 7.75 11.56 9.89 11.01 9.65 9.59
EC Ind		30.12 29.22 30.10 30.53 30.92 30.29 30.21 29.10 30.11 30.76
NN [a]	Somme	8.43×10^3 8.59×10^3 8.50×10^3 8.58×10^3 8.60×10^3 8.51×10^3 8.38×10^3 8.41×10^3 8.40×10^3 8.57×10^3
	Min	4.48 5.00 4.26 2.42 3.58 4.32 3.02 4.32 3.23 3.27
	Max	84.88 84.10 45.47 84.98 65.27 40.49 42.28 40.10 41.94 39.13
	EC	3.62 3.48 2.52 3.87 3.35 2.44 2.45 2.53 2.34 2.56
Dep [b]	Somme	329.62×10^3 329.08×10^3 329.41×10^3 329.55×10^3 330.49×10^3 331.06×10^3 329.56×10^3 330.48×10^3 329.25×10^3 329.65×10^3
	Min	0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01
	Max	143.78 143.78 148.63 143.78 143.78 170.35 165.34 147.31 165.90 165.13
	EC	2.99 3.02 3.17 3.01 2.93 3.29 3.14 3.30 3.15 3.08
Tot [c]	Somme	8.96×10^6 9.11×10^6 9.06×10^6 9.09×10^6 9.16×10^6 9.04×10^6 8.93×10^6 8.90×10^6 9.01×10^6 9.04×10^6
	Min	4.48 5.00 4.26 2.42 3.58 4.32 3.02 4.32 3.23 3.27
	Max	105.29 100.78 95.42 129.91 99.79 75.19 67.40 70.12 61.04 61.35
	EC	6.65 6.85 6.25 7.42 6.83 6.17 5.75 5.94 5.81 5.85
	Nb singletons	5.00 4.00 5.00 1.00 4.00 5.00 5.00 1.00 5.00 7.00

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.15: Tableau des résultats sur 10 runs pour la méthode k-medoïds (M3A)

4.4.3.2 Résultats sur la moyenne des 10 runs (jeu 1 - k = 1000)

Rayon et densité des groupes Nous remarquons dans un premier temps que le rayon maximum (cf. Tableau 4.16) est très différent selon le type d'initialisation. En effet pour une initialisation aléatoire des centres (M3A-M3B), les rayons maximum sont très élevés (153.78 et 109.18 respectivement). Alors que pour une initialisation avec FFT (M3C-M3D), les rayons maximums sont faibles (16.67). En revanche les rayons moyens sont similaires quelque soit l'initialisation. Ils sont donc très éloignés du rayon maximum pour une initialisation aléatoire contrairement à l'initialisation avec FFT. L'initialisation aléatoire produit donc des groupes de rayons peu homogènes. Ceci est confirmé par l'écart moyen du nombre d'individus par groupe (cf. section 4.2.1.2). Comme noté pour la méthode M1, cet écart est faible. Etant donné notre jeu non homogènement réparti dans l'espace et ce faible écart du nombre d'individus par groupe, on peut en déduire que les groupes sont de rayons très variés, certains couvrant de trop grandes parties de l'espace. Cela induit le risque que les centres ne soient pas de bons représentants.

En revanche pour l'initialisation avec FFT, le rayon maximum et le rayon moyen sont proches (16.67 pour 12.19 avec M3C) et l'écart du nombre d'individus par groupe est plus important. Cette initialisation permet vraisemblablement à la méthode k-medoïds de produire des groupes plus homogènement répartis sur l'espace des descriptions.

Enfin pour l'initialisation aléatoire, l'utilisation de k-medoïds avec des centres virtuels (M3B) donne de meilleurs résultats qu'avec l'utilisation de centres réels pour le rayon maximum. Cette tendance ne se confirme pas pour l'initialisation avec FFT.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	
M3A	153.78	14.10	9.59	30.14	initialisation
M3B	109.18	12.91	6.32	22.99	aléatoire
M3C	16.67	12.19	1.18	162.93	initialisation
M3D	16.67	12.29	1.23	75.34	avec FFT

TABLE 4.16: Résultats du critère Rayon pour la méthode k-medoïds (M3) - Jeu 1, k=1000 ; centres réels, centres virtuels

Diversité/Recouvrement de l'espace Lorsque l'on étudie les distances entre chaque centre et son plus proche, et connaissant la qualité des rayons maximums de chaque paramétrage, il semble encore une fois que l'initialisation avec FFT (M3C-M3D) donne de meilleurs résultats que l'initalisation aléatoire (M3A-M3B). En effet la première a une somme des distances plus élevée (environ 13 000, cf. Tableau 4.17) que la deuxième (environ 8000). De plus la distance maximum entre deux centres voisins est très élevée également pour l'intialisation avec FFT (127.69) montrant une couverture des extrêmes. Nous observons cependant pour la distance minimum que l'initialisation n'est pas le seul paramètre à prendre en compte. En effet pour l'initialisation avec FFT et le travail avec les centres virtuels (M3D), au moins un centre est très proche de son voisin (2.31). Or pour l'initialisation alétoire (M3A-M3B), ce critère est moins faible (3.79-4.09). Il en ressort que l'initialisation avec FFT couplée avec un travail en centres réels donne des résultats plus proches de nos attentes. M3B (travail en centres virtuels) reste meilleur que M3A(travail en centres réels).

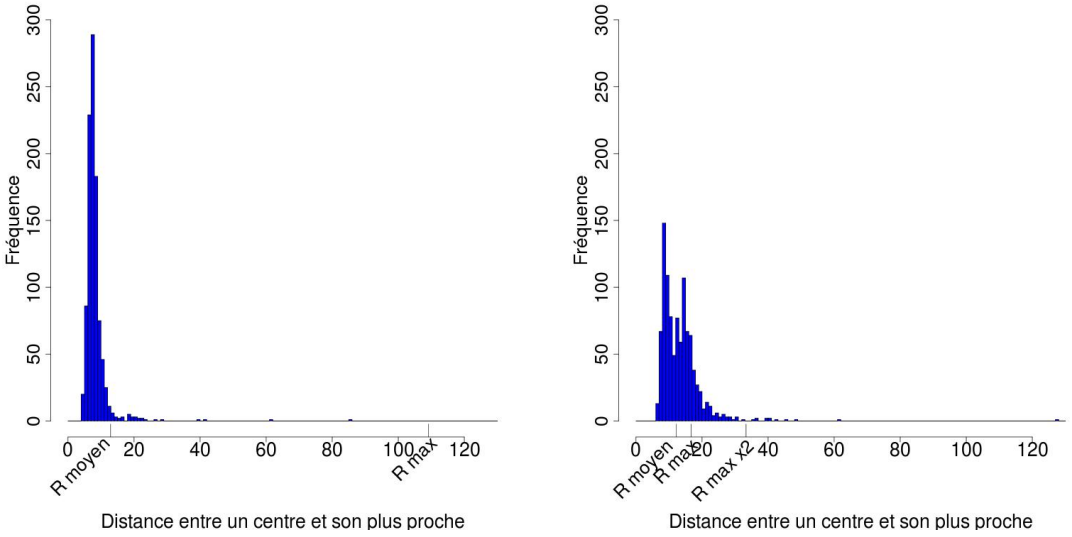
Ces observations se confirment par l'étude de la distribution de ces distances (cf. Figure 4.18). La majorité des distances se trouve proches de 0 et inférieure au rayon moyen pour l'initialisation aléatoire (Figure 4.18(a)). Alors que l'initialisation avec FFT (cf. Figure

4.18(b)) donne une distribution des distances plus proches du rayon moyen et du rayon maximum mais également plus dispersées. De plus un certain nombre de centres sont plus proches de leur voisin que le rayon moyen comme pour l'initialisation aléatoire. Cela indique donc un espace couvert de façon peu homogène quelque soit le type d'initialisation. L'échantillon couvre un espace plus divers avec M3C qu'avec M3B.

NN [a]	Somme	Min	Max	EC	
M3A	8.50×10^3	3.79	56.86	2.92	initialisation
M3B	8.27×10^3	4.09	106.17	4.70	aléatoire
M3C	13.35×10^3	6.15	127.69	6.52	initialisation
M3D	12.73×10^3	2.31	127.69	6.87	avec FFT

a. $\text{NN} = \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

TABLE 4.17: Résultats du critère SDNN pour la méthode k-medoïds (M3) - Jeu 1, k=1000 ; centres réels, centres virtuels



(a) M3B SDNN (initialisation aléatoire)

(b) M3C SDNN (initialisation avec FFT)

FIGURE 4.18: Pour M3 k-medoïds, Distribution des dissimilarités entre chaque centre et son centre le plus proche (SDNN)

Représentativité La somme des distances entre chaque molécule et son représentant (SDDep, cf. Tableau 4.18) est légèrement plus faible pour une initialisation aléatoire (M3A : environ 320 000 et M3B : environ 315 000) que pour une initialisation avec FFT (M3C-M3D). En étudiant ce seul critère, l'initialisation aléatoire semble donner un échantillon plus représentatif. Mais la molécule la plus éloignée de son représentant est à une grande distance pour M3A et M3B. Enfin la somme des distances SDDep est plus faible et donc meilleure pour le travail en centres virtuels (M3B est inférieur à M3A donc plus représentatif).

Par ailleurs, en observant les histogrammes de la figure 4.19 nous confirmons cette tendance. L'intialisation aléatoire (Figure 4.19(b)) montre que la plupart des distances

se situent avant le rayon moyen mais que quelques unes se trouvent supérieures à ce rayon. Ceci montre que l'échantillon ne comporte pas de bons représentants pour toutes les molécules du jeu de départ. En revanche pour l'initialisation avec FFT (4.19(c)), peu de distances s'éloignent du rayon moyen.

Si on considère donc tous les critères d'évaluation de la représentativité, il s'avère que l'initialisation avec FFT donne un échantillon globalement plus représentatif. Certes avec l'intialisation aléatoire, une partie du jeu est très proche de ses représantants ; mais une autre partie en est en revanche très éloignée. L'initialisation aléatoire, malgré une meilleure valeur pour le critère SDDep, ne donne pas l'échantillon le plus représentatif.

Dep [a]	Somme	Min	Max	EC	
M3A	329.82×10^3	0.01	153.78	3.11	initialisation
M3B	315.5×10^3	0.01	109.18	2.70	aléatoire
M3C	341.61×10^3	0.01	16.67	1.93	initialisation
M3D	324.12×10^3	0.01	16.67	2.00	avec FFT

a. Dep $= \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

TABLE 4.18: Résultats du critère SDDep pour la méthode k-medoïds (M3) - Jeu 1, k=1000 ; centres réels, centres virtuels

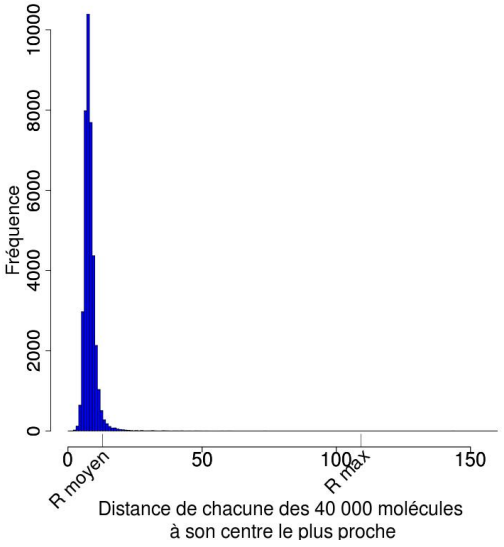
Dissimilarité totale dans l'échantillon Encore une fois l'initialisation aléatoire (M3A-M3B) donne des distances maximales entre deux centres plus faibles (86.63-148.57, cf. Tableau 4.20) que pour l'initialisation avec FFT (201.77). Ceci montre que cette dernière initialisation permet à la méthode k-medoïds de produire des échantillons couvrant mieux les extrêmes. De plus les distances entre deux centres sont très concentrées autour du rayon moyen avec quelques distances extrêmes pour l'initialisation aléatoire (Figure 4.20(a)). Alors que pour l'initilisation avec FFT (Figure 4.20(b)) la distribution des distances est plus large et couvre plus de valeurs extrêmes. On a donc un échantillon plus éclaté et moins redondant comme présenti précédemment avec l'initialisation avec FFT.

Tot [a]	Somme	Min	Max	EC	
M3A	9.03×10^6	3.79	86.63	6.35	initialisation
M3B	9.23×10^6	4.09	148.57	9.11	aléatoire
M3C	17.73×10^6	6.15	201.77	16.89	initialisation
M3D	17.25×10^6	2.31	201.77	17.30	avec FFT

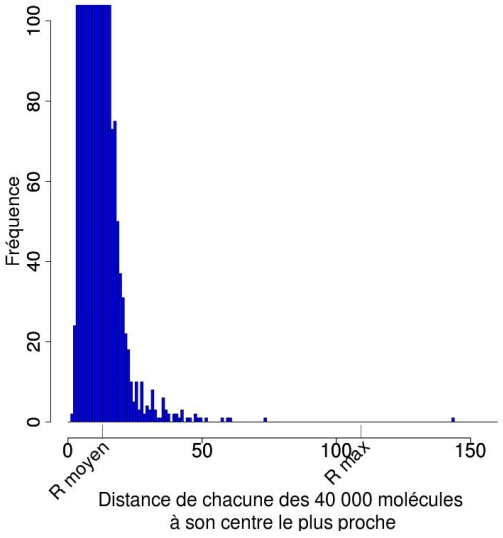
a. Tot $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.19: Résultats du critère SDTot pour la méthode k-medoïds (M3) - Jeu 1, k=1000 ; centres réels, centres virtuels

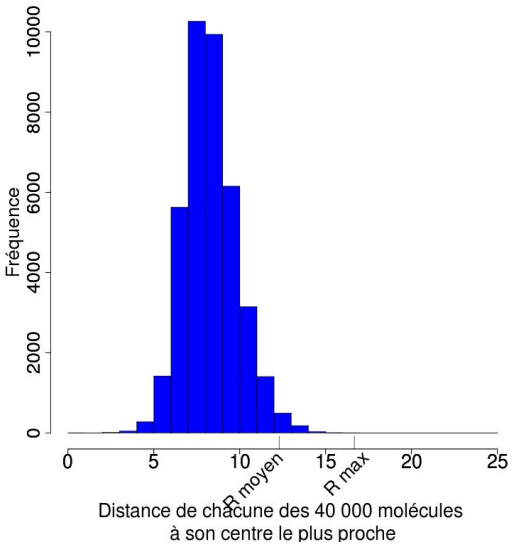
Singletons k-medoïds avec une initialisation aléatoire (M3A-M3B) produit très peu de singletons (cf. Tableau 4.20) et ceux-ci ont des distances à leur molécule la plus proche (cf. Figure 4.21(b)) très faible. En revanche k-medoïds (M3C-M3D) avec une initialisation avec FFT donne beaucoup de singletons dont la majorité des distances à leur molécule voisine (cf. Figure 4.21(d)) se situe dans les extrêmes. Ils sont donc de meilleurs indicateurs d'outliers.



(a) M3B SDDep histogramme complet

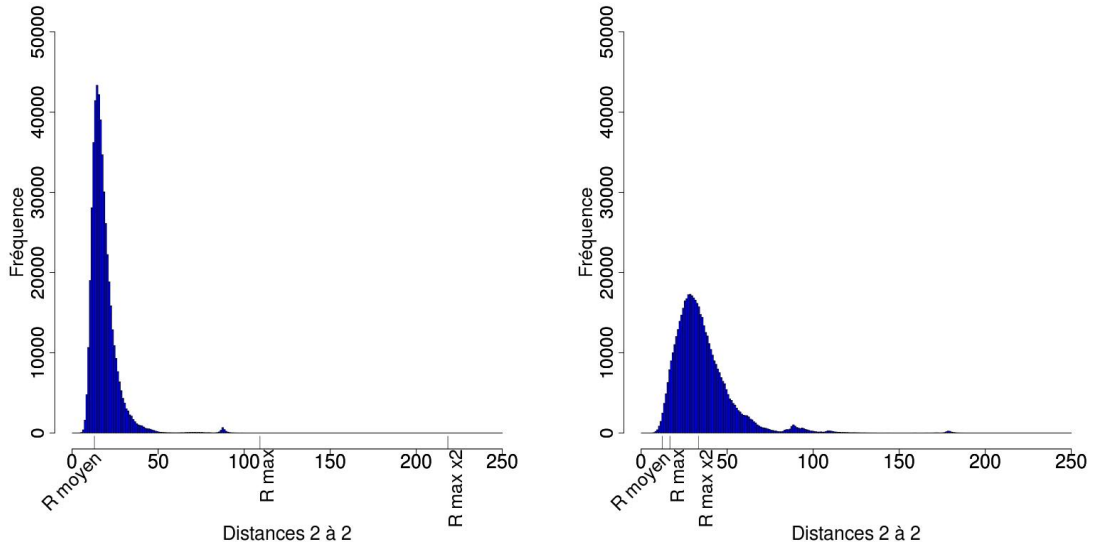


(b) M3B SDDep échelle réduite



(c) M3D SDDep

FIGURE 4.19: Pour M3 k-medoïds, Distribution des dissimilarités entre chaque molécule et son centre le plus proche (SDDep)



(a) M3B SDTot (initialisation aléatoire)

(b) M3C SDTot (initialisation avec FFT)

FIGURE 4.20: Pour M3 k-medoïds, Distribution des dissimilarités entre tous les centres (SDTot)

	Singletons	
M3A	4.20	initialisation
M3B	3.60	aléatoire
M3C	326.00	initialisation
M3D	309.00	avec FFT

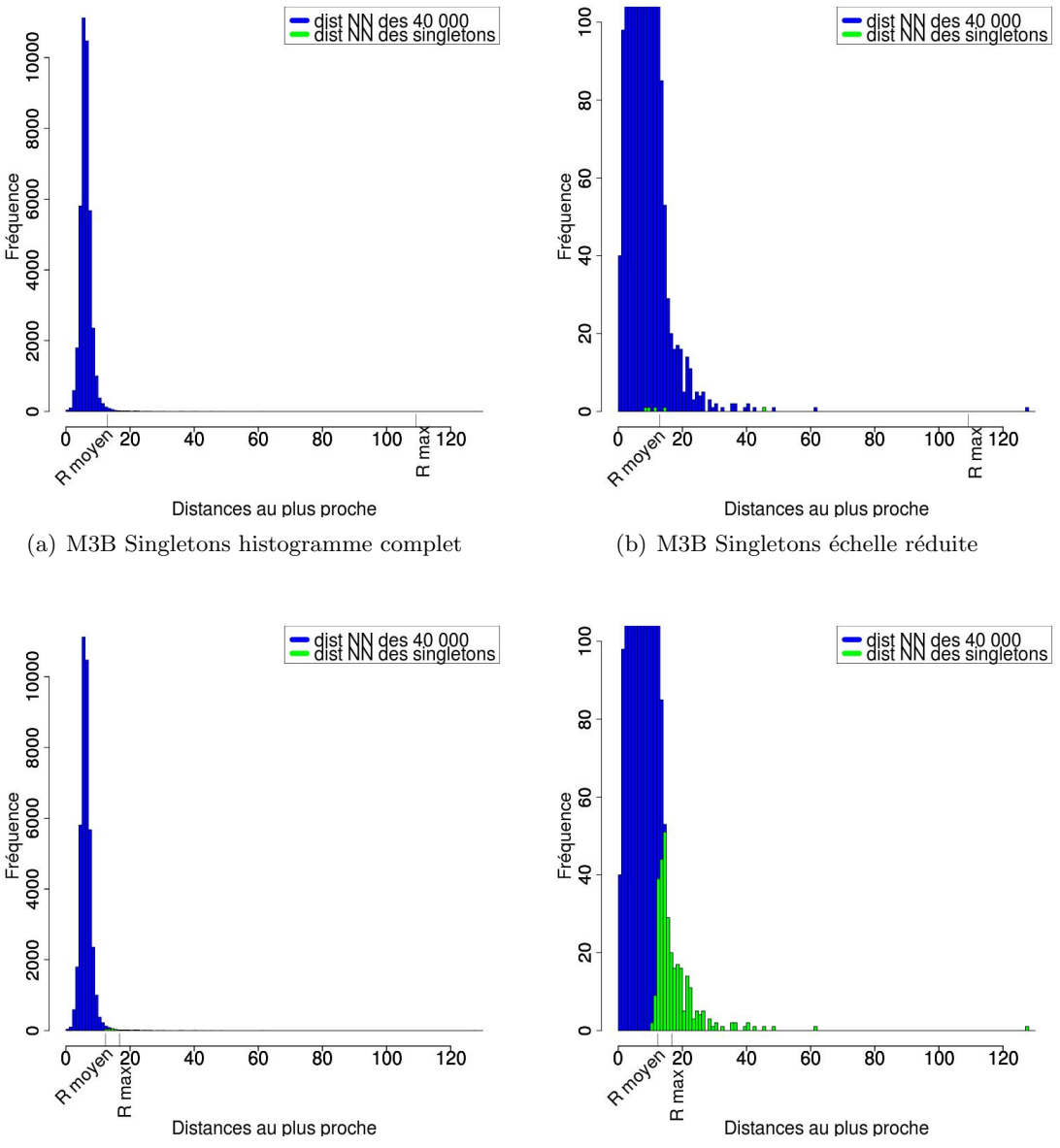
TABLE 4.20: Résultats du critère Singletons pour la méthode textitk-medoïds (M3) - Jeu 1, k=1000 ; centres réels, centres virtuels

Conclusion La méthode k-medoïds est très sensible au type d'initialisation. Il semble que l'initialisation avec FFT donne des résultats proches de notre objectif. Quant au choix de travailler en centres virtuels ou réels, il semble que le travail avec des centres virtuels donne de meilleurs résultats surtout avec une initialisation aléatoire.

Pour cette méthode, avec une initialisation aléatoire, le temps de calcul pour le travail en centres réels est d'environ 3 minutes, contre 17 minutes pour le travail en centres virtuels. L'algorithme étant plus complexe en travail avec centres réels, on en déduit donc une convergence nettement plus rapide pour le travail en centres réels que pour celui en centres virtuels. Enfin pour l'initialisation avec FFT, le temps de calcul est d'environ 3 minutes 30 pour le travail en centres réels contre 7 minutes 30 pour le travail en centres réels. La convergence plus rapide grâce à l'utilisation des centres réels est donc confirmée. En revanche, l'intialisation avec FFT n'a pas permis de converger plus vite qu'avec l'initialisation aléatoire.

4.4.4 Sélection avec Maximum-Dissimilarity (M4)

Nous rappelons que plusieurs initialisations ont été testées avec cette méthode : aléatoire, à la molécule centrale du jeu de départ, à la molécule la plus éloignée de toutes. Ensuite pour chacune d'elle, deux critères différents de choix de la molécule suivante à



(a) M3B Singletons histogramme complet

(b) M3B Singletons échelle réduite

(c) M3D Singletons histogramme complet

(d) M3D Singletons échelle réduite

FIGURE 4.21: Pour M3 k-medoïds, Distribution des dissimilarités entre chaque molécule et sa molécule la plus proche et projection de ces distances pour les singletons

sélectionner ont été appliqués : le critère MaxSum (la molécule la plus éloignée du sous-ensemble) et le critère MaxMin (la molécule la plus éloignée de sa plus proche dans le sous-ensemble). Nous obtenons donc 6 paramétrages différents de la méthode Maximum-Dissimilarity, auxquels nous attribuons par la suite le code couleur indiqué dans le tableau 4.21 pour une meilleure compréhension :

M4A	initialisation
M4B	aléatoire
M4C	initialisation
M4D	au centre
M4E	initialisation
M4F	au plus loin

TABLE 4.21: Code couleur des différents paramètres utilisés pour la méthode Maximum-Dissimilarity (M4) ; MaxSum, MaxMin

4.4.4.1 Stabilité sur 10 runs

Seules les variantes M4A et M4B sont concernées par l'initialisation aléatoire. Encore une fois ces initialisations ont été réalisées 10 fois de manière à vérifier que l'algorithme donnait des résultats similaires quelque soit le tirage effectué. Or les 10 runs de M4A engendrent des valeurs similaires pour les différents critères (cf. tableau 4.22) sauf pour SSDep. Nous obtenons le même genre de résultats pour M4B, et le critère SDDep ne varie pas autant que pour M4A. La moyenne de la somme SDDep rendra quand même compte correctement de la tendance pour M4A. Nous nous fixons donc sur l'étude de la moyenne des runs pour la suite de l'analyse.

	Critère	Résultats pour chacun des 10 runs pour M4A
	Rayon max.	23.27 23.27 23.27 23.27 23.27 24.02 23.27 22.86 23.27 23.27
	Rayon moyen	16.87 15.99 16.60 16.17 17.00 18.22 17.73 16.43 16.81 16.83
	EC Rayon	3.35 2.51 2.96 2.63 3.11 3.83 3.71 2.79 3.14 2.95
	EC Ind	1098.47 1383.84 1202.47 1350.50 1001.34 248.92 496.38 1277.88 1050.16 1087.15
NN [a]	Somme	12.44×10^3 12.45×10^3 12.45×10^3 12.45×10^3 12.44×10^3 12.43×10^3 12.45×10^3 12.44×10^3 12.45×10^3 12.45×10^3
	Min	2.38 2.38 2.38 2.38 2.38 2.38 2.38 2.38 2.38 2.38
	Max	127.69 127.69 127.69 127.69 127.69 127.69 127.69 127.69 127.69 127.69
	EC	6.65 6.65 6.65 6.65 6.65 6.65 6.65 6.65 6.65 6.65
Dep [b]	Somme	599.96×10^3 519.55×10^3 571.92×10^3 531.39×10^3 617.84×10^3 690.65×10^3 671.88×10^3 556.80×10^3 591.55×10^3 590.58×10^3
	Min	0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01
	Max	23.27 23.27 23.27 23.27 23.27 24.02 23.27 22.86 23.27 23.27
	EC	3.54 2.96 3.21 2.97 3.29 3.91 3.83 3.04 3.49 3.22
Tot [c]	Somme	21.61×10^6 21.61×10^6 21.61×10^6 21.61×10^6 21.61×10^6 21.62×10^6 21.62×10^6 21.61×10^6 21.61×10^6 21.61×10^6
	Min	2.38 2.38 2.38 2.38 2.38 2.38 2.38 2.38 2.38 2.38
	Max	201.77 201.77 201.77 201.77 201.77 201.77 201.77 201.77 201.77 201.77
	EC	16.28 16.28 16.28 16.28 16.27 16.27 16.27 16.27 16.28 16.27 16.27
	Nb singletons	543.00 547.00 541.00 551.00 543.00 532.00 536.00 549.00 543.00 543.00

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.22: Tableau des résultats sur 10 runs pour la méthode k-medoïds (M4A)

4.4.4.2 Résultats sur la moyenne des 10 runs (jeu 1 - k = 1000)

Rayon et densité des groupes Dans un premier temps nous étudions les conséquences d'une initialisation différente. Les variantes M4A, M4C et M4E (cf. Tableau 4.23, couleur :) ont le même paramètre de recherche de la molécule suivante à sélectionner et sont basées respectivement sur une initialisation aléatoire, à l'individu central du jeu et à l'individu le plus éloigné. Globalement le rayon maximum est sensiblement le même (24.90, 24.34 et 25.10) quelle que soit l'initialisation utilisée. En revanche pour le rayon moyen on observe des différences. En effet, l'initialisation à l'individu central (M4C) produit un rayon moyen de 16.23 meilleur que les autres intialisations. L'initialisation à l'individu le plus éloigné de tous (M4E) donne le rayon moyen le moins bon (18.83). En ce qui concerne l'écart du nombre d'individus par groupe, il est plus élevé aussi pour l'initialisation au centre avec 1614.58 contre 1218.95 et 275.46 pour une initialisation aléatoire et une initialisation au plus éloigné respectivement. En conclusion pour le paramètre de choix de la molécule suivante à sélectionner : "MaxSum", le type d'initialisation utilisée a des effets sur les résultats. Il en ressort que l'initialisation au centre permet à la méthode Maximum-Dissimilarity de produire de meilleurs résultats pour le critère du rayon. L'initialisation à la molécule la plus éloignée donne les moins bons résultats.

Pour le paramètre de choix de la molécule suivante avec le critère "MaxMin", les variantes M4B, M4D et M4F sont basées respectivement sur une initialisation aléatoire, à l'individu central du jeu et à l'individu le plus éloigné de tous (cf. Tableau 4.23, couleur :). Dans ce cas, les différences sont moins flagrantes. Le rayon maximum (14.00, 13.99 et 13.94) est sensiblement le même, ainsi que le rayon moyen (12.96, 12.97, 12.92). Nous notons cependant des différences pour l'écart du nombre d'individus par groupe (190.29, 585.84, 275.46). La tendance observée pour M4A, M4C et M4E se confirme alors avec un meilleur résultat pour l'initialisation au centre.

Enfin nous étudions l'impact du critère de choix de la molécule suivante à sélectionner ; soit pour une même initialisation nous comparons M4A à M4B, M4C à M4D et M4E à M4F. Nous notons alors une différence significative sur les effets de ces deux critères de choix quelle que soit l'initialisation. Pour illustrer notre propos nous comparons M4A à M4B (critères MaxSum et MaxMin respectivement). Le rayon maximum (M4A=24.90,M4B=14.00) et le rayon moyen (M4A=17.20 et M4B=12.96) montrent que le critère de choix MaxMin donne de biens meilleurs résultats. Cependant l'écart du nombre d'individus dans les groupes est beaucoup plus faible pour M4B que pour M4A. Il faudra donc étudier si le quadrillage est effectivement homogène pour M4B comme le laisse à penser son rayon maximum.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	
M4A	24.90	17.20	3.24	1218.95	initialisation
M4B	14.00	12.96	1.07	190.29	aléatoire
M4C	24.34	16.23	2.57	1614.58	initialisation
M4D	13.99	12.97	1.03	585.84	au centre
M4E	25.10	18.83	4.17	275.46	initialisation
M4F	13.94	12.92	1.08	150.49	au plus loin

TABLE 4.23: Résultats du critère Rayon pour la méthode Maximum-Dissimilarity (M4) - Jeu 1, k=1000 ; MaxSum, MaxMin

Diversité/Recouvrement de l'espace Pour le critère de choix de la molécule suivante : MaxSum (M4A, M4C et M4E, cf. Tableau 4.24, couleur :), les résultats sont ici sensiblement les mêmes (par exemple : SDNN = 8551.73, 8579.25, 8550.32) pour tous les critères d'évaluation quelque soit l'initialisation. On observe une distance minimum entre deux centres voisins quasi nulle (0.01) et une distance maximale entre deux centres voisins faible (26.84). Le critère de choix MaxSum ne donne donc pas de bons résultats pour un quadrillage homogène de l'espace. Les centres ne sont pas dispersés de façon régulière dans l'espace. En revanche pour le critère MaxMin (M4B, M4D, et M4F, cf. Tableau 4.24, couleur :), la plus petite distance entre deux centres voisins correspond au rayon maximum. Ceci signifie qu'aucun centre n'est plus proche de son voisin que la taille du plus grand groupe. On peut donc penser que le quadrillage de l'espace est homogène. De plus, la plus grande distance entre deux centres voisins est élevée (127.69) indiquant que les extrêmes de l'espace sont couverts.

Enfin si l'on compare les deux critères de choix MaxSum et MaxMin entre eux (M4A versus M4B par exemple), on remarque la même tendance que précédemment pour le rayon. Le critère MaxMin donne de meilleurs résultats que MaxSum. En effet la somme des distances au plus proche est deux fois plus élevée pour M4B (16177.79) que pour M4A (8551.73).

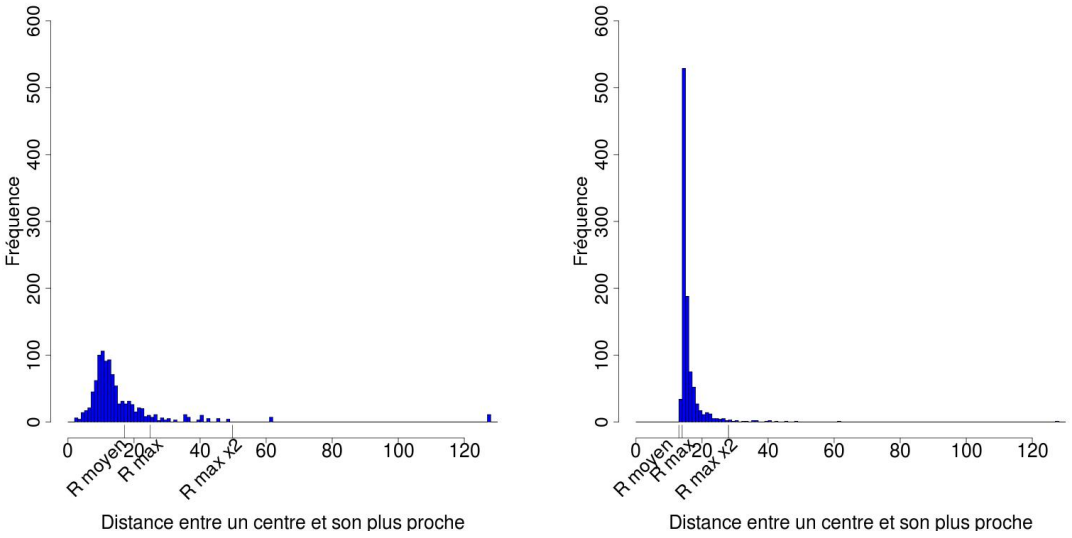
Ceci se confirme avec l'étude de la distribution des distances (cf. Figure 4.22). En effet pour le critère "MaxSum" (Figure 4.22(a)), la majorité des distances se trouve avant le rayon moyen et proche de 0. De plus, elles sont très dispersées, indiquant une couverture non homogène de l'espace, voire une certaine redondance pour les centres proches dont la distance est inférieure au rayon moyen. Alors que pour le critère "MaxMin" (cf. Figure 4.22(b)) toutes les distances sont après le rayon moyen et sont extrêmement concentrées autour de celui-ci et du rayon maximum. L'espace est donc couvert par des représentants espacés homogènement. La diversité est bien représentée puisque les extrêmes sont égalements couverts.

NN [a]	Somme	Min	Max	EC	
M4A	8.55×10^3	0.01	26.84	6.18	initialisation
M4B	16.17×10^3	14.00	127.69	5.21	aléatoire
M4C	8.57×10^3	0.01	26.84	6.18	initialisation
M4D	16.17×10^3	13.99	127.69	5.22	au centre
M4E	8.55×10^3	0.01	26.84	6.17	initialisation
M4F	16.16×10^3	13.95	127.69	5.22	au plus loin

a. $\text{NN} = \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

TABLE 4.24: Résultats du critère SDNN pour la méthode Maximum-Dissimilarity (M4) - Jeu 1, k=1000 ; MaxSum, MaxMin

Représentativité Pour le critère MaxSum (M4A, M4C, M4E, cf. Tableau 4.25, couleur :), on observe de nettes différences dans la somme des distances entre une molécule et son représentant. Encore une fois l'initialisation au centre (M4C 482695.28) donne un échantillon plus représentatif que les autres initialisations car la somme des distances est plus faible. Et l'initialisation au plus éloigné (M4E 730887.69) donne un échantillon moins représentatif. Pour le critère MaxMin (M4B, M4D et M4E, cf. Tableau 4.25, couleur :) on observe la même tendance de façon moins significative avec une somme des distances plus



(a) M4A SDNN (b) M4B SDNN

FIGURE 4.22: Pour M4 Maximum-Dissimilarity, Distribution des dissimilarités entre chaque centre et son centre le plus proche (SDNN)

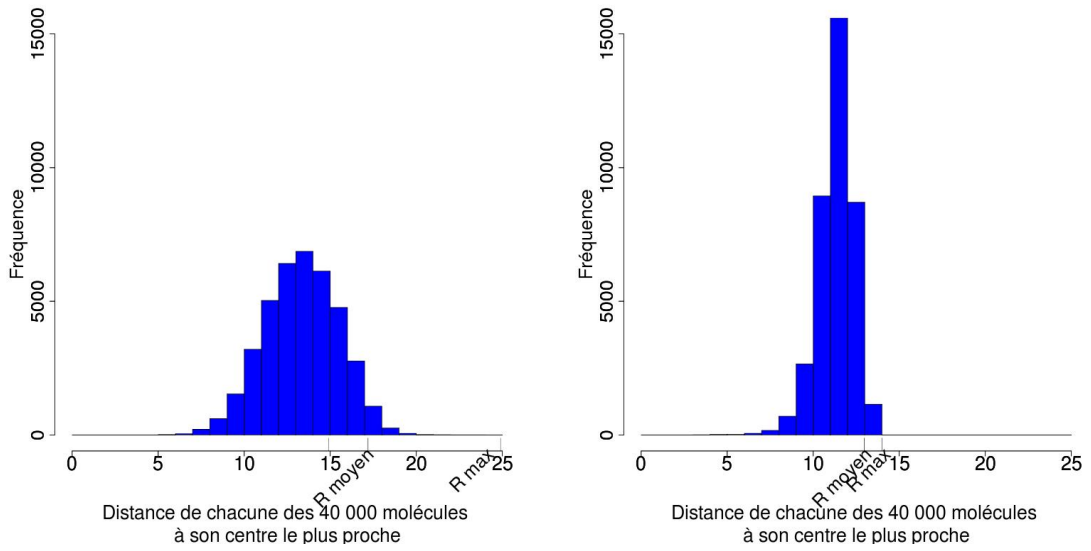
faible pour l'initialisation au centre (M4D 419579.38) que pour les autres initialisations (M4B 441 648.44 et M4F 440 230.53). Cependant pour ce critère l'initialisation aléatoire ne se détache pas de l'initialisation au plus éloigné. Enfin en comparant les deux critères, on observe toujours un net avantage à choisir le critère MaxMin (M4B 441 648.44) plutôt que le critère MaxSum (M4A 598 66.94). En effet le premier donne un échantillon plus représentatif. Cette observation se confirme avec l'étude de la distribution des distances entre chaque molécule et son représentant (cf. Figure 4.23). Pour le critère MaxSum (M4A, Figure 4.23(a)), ces distances s'étendent jusqu'à 20 alors que pour le critère MaxMin elles se concentrent avant 15.

Dep [a]	Somme	Min	Max	EC	
M4A	598.66×10^3	0.01	24.90	3.32	initialisation
M4B	441.64×10^3	0.01	14.00	2.05	aléatoire
M4C	482.69×10^3	0.01	24.34	3.08	initialisation
M4D	419.57×10^3	0.01	13.99	2.20	au centre
M4E	730.88×10^3	0.01	25.10	3.93	initialisation
M4F	440.23×10^3	0.01	13.94	2.04	au plus loin

a. $\text{Dep} = \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

TABLE 4.25: Résultats du critère SDDep pour la méthode Maximum-Dissimilarité (M4) - Jeu 1, k=1000 ; MaxSum, MaxMin

Dissimilarité totale dans l'échantillon Pour le critère MaxSum (M4A, M4C et M4E, cf. Tableau 4.26, couleur :) les résultats sont sensiblement les mêmes quelle que soit l'initialisation ainsi que pour le critère MaxMin (M4B, M4D et M4F, cf.Tableau 4.26, couleur :). On observe notamment que la plus grande distance entre deux centres est



(a) M4A SDDep (b) M4B SDDep

FIGURE 4.23: Pour M4 Maximum-Dissimilarity, Distribution des dissimilarités entre chaque molécule et son centre le plus proche (SDDep)

élevée, ce qui signifie que les extrêmes de l'espace sont inclus dans l'échantillon. Comme précédemment, il y a une nette différence entre les deux critères. Le critère MaxSum (M4A 28 million environ) donne une somme des distances totales entre centres, plus élevée que pour le critère MaxMin (M4B 14 millions environ). Dans une étude classique, seul ce critère aurait été étudié et le critère MaxSum aurait donc été considéré comme meilleur que le critère MaxMin. Or étant donné les résultats obtenus pour le rayon des groupes, le recouvrement de l'espace et la représentativité, il s'avère que maximiser la somme des dissimilarités totale n'est pas un bon critère d'évaluation de la diversité d'un échantillon. Du moins, il ne peut être utilisé comme seul critère d'optimisation.

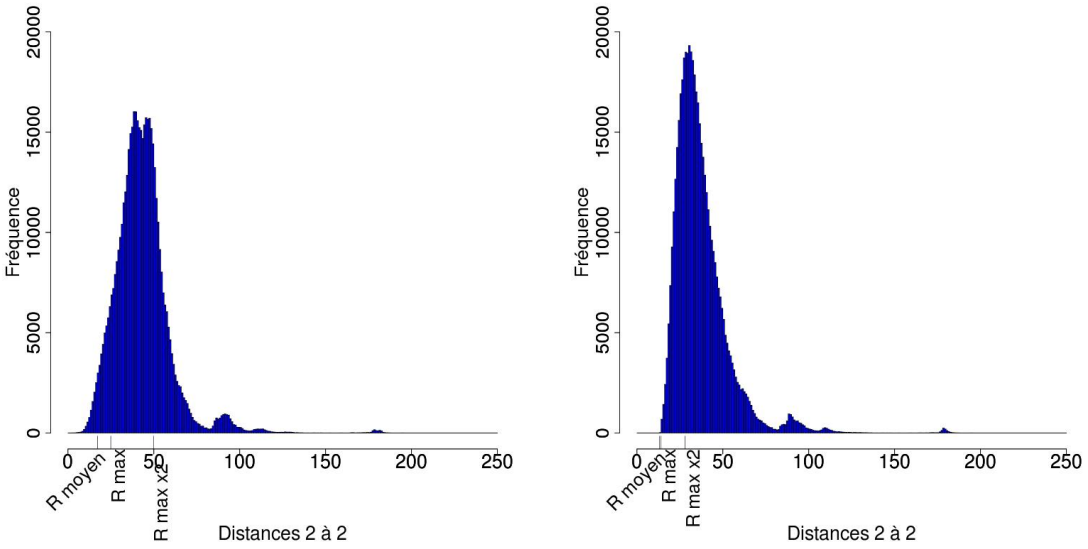
Enfin nous confirmons ces observations par l'étude de la distribution des distances (cf. Figure 4.24). Pour le critère MaxSum (Figure 4.24(a)) la répartition des distances est plus étendue que pour le critère MaxMin (Figure 4.24(b)) et notamment avant le rayon maximum. De plus, il y a des grandes distances couvertes par M4A (notamment autour de 100 et entre 150 et 200). La méthode Maximum Dissimilarity couvre donc certainement plus d'extrêmes que FFT (M4B) tout en donnant un échantillon plus redondant dans les zones les plus denses de l'espace (car beaucoup de petites distances). Ceci confirme que la maximisation du critère SDTot ne donne pas forcément un échantillon divers. Il a tendance à favoriser les extrêmes au détriment de la couverture homogène de l'espace central.

Singletons Comme nous le présentions précédemment, le critère MaxSum (M4A, M4C, M4E, cf. Tableau 4.27, couleur :) produit plus de singletons que le critère MaxMin (M4B, M4D et M4F, cf. Tableau 4.27, couleur :). Cependant lorsqu'on observe la distribution des distances entre chaque singleton et sa molécule la plus proche (cf. Figure 4.25(b)), on s'aperçoit que les singletons produits pour M4A ne couvrent pas toutes les distances les plus élevées et couvrent même des distances faibles à leur plus proche molécule. Ceux-ci ne sont donc pas de bons indicateurs d'outliers. En revanche pour le critère MaxMin (M4B, cf. Figure 4.25(d)), les singletons couvrent toutes les distances les plus grandes et aucun

Tot [a]	Somme	Min	Max	EC	
M4A	27.72×10^6	0.01	201.77	28.69	initialisation
M4B	18.72×10^6	14.00	201.77	16.03	aléatoire
M4C	27.71×10^6	0.01	201.77	28.69	initialisation
M4D	18.73×10^6	13.99	201.77	16.04	au centre
M4E	27.74×10^6	0.01	201.77	28.68	initialisation
M4F	18.66×10^6	13.95	201.77	16.04	au plus loin

a. Tot $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.26: Résultats du critère SDTot pour la méthode Maximum-Dissimilarity (M4) - Jeu 1, k=1000 ; MaxSum , MaxMin



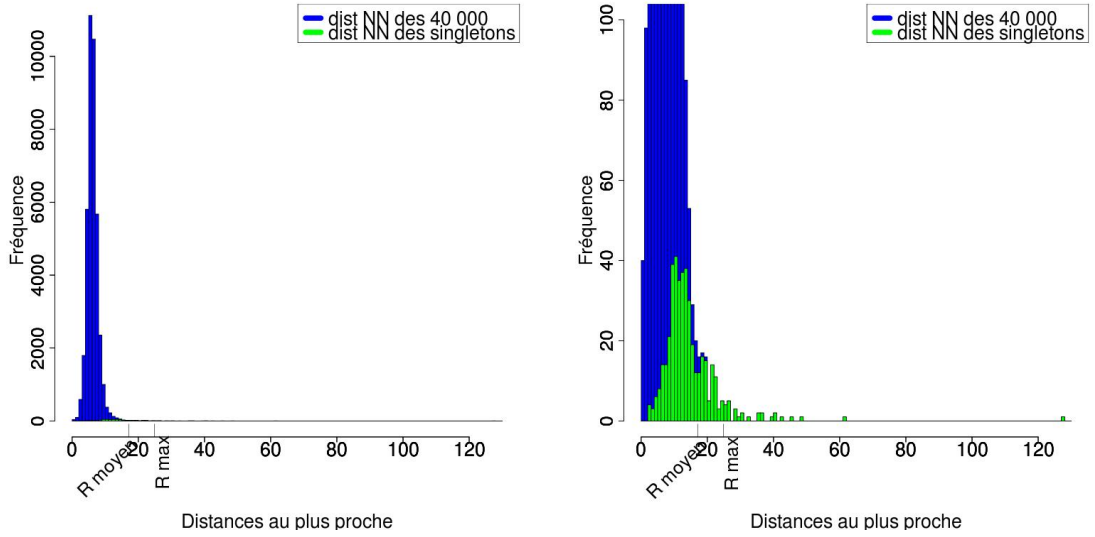
(a) M4A SDTot (b) M4B SDTot

FIGURE 4.24: Pour M4 Maximum-Dissimilarity, Distribution des dissimilarités entre tous les centres (SDTot)

ne couvre de faibles distances (inférieur à 10). Ils peuvent donc être , pour la plupart, de bons indicateurs d'outliers.

	Singletons	
M4A	428.10	initialisation
M4B	307.80	aléatoire
M4C	436.00	initialisation
M4D	309.00	au centre
M4E	422.00	intialisation
M4F	314.00	au plus loin

TABLE 4.27: Résultats du critère Singletons pour Maximum-Dissimilarity (M4) - Jeu 1, k=1000 ; MaxSum , MaxMin



(a) M4A Singletons histogramme complet

(b) M4A Singletons échelle réduite

(c) M4B Singletons histogramme complet

(d) M4B Singletons échelle réduite

FIGURE 4.25: Pour M4 Maximum-Dissimilarity, Distribution des dissimilarités entre chaque molécule et sa molécule la plus proche et projection de ces distances pour les singletons

Conclusion En conclusion, il semble que l'initialisation ait un impact sur certains critères d'évaluation notamment pour le rayon et la représentativité où une initialisation à l'individu central du jeu donne de meilleurs résultats. Ensuite il est clair que le critère de choix de la molécule suivante à sélectionner a des conséquences sur l'échantillon. En effet lorsque le critère MaxMin est utilisé, l'échantillon semble plus divers et plus proche de notre objectif qu'avec le critère MaxSum. La méthode FFT utilisant le premier critère est donc une bonne méthode de sélection par diversité, d'autant plus avec une initialisation à l'individu central du jeu de départ.

Quel que soit le paramétrage de la méthode (type d'intialisation et choix de la molécule suivante), le temps de calcul est très faible (la solution est trouvée presque instantanément), soit environ 30 à 40 secondes.

4.4.5 Sélection avec Sphere-Exclusion (M5)

Nous rappelons que plusieurs initialisations ont été testées avec cette méthode : aléatoire, à la molécule centrale du jeu de départ, à la molécule la plus éloignée de toutes. Ensuite pour chacune d'elle, quatre critères différents de choix de la molécule suivante à sélectionner ont été appliqués : choix aléatoire, le critère MinSum (la molécule la plus proche du sous-ensemble), le critère MinMin (la molécule la plus proche de sa plus proche dans le sous-ensemble) et le critère MaxMin (la molécule la plus éloignée de sa plus proche dans le sous-ensemble). Nous obtenons donc 12 paramétrages différents de la méthode Sphere-Exclusion, auxquels nous attribuons par la suite le code couleur indiqué dans le tableau 4.28 pour une meilleure compréhension.

M5A	initialisation aléatoire
M5B	
M5C	
M5D	
M5E	initialisation au centre
M5F	
M5G	
M5H	
M5I	initialisation au plus loin
M5J	
M5K	
M5L	

TABLE 4.28: Code couleur des différents paraméètres utilisés pour la méthode Sphere-Exclusion (M5) ; aléatoire, MinSum, MinMin, MaxMin

4.4.5.1 Stabilité sur 10 runs

Seules les variantes M5A, M5B, M5C, M5D et M5I sont concernées par l'initialisation aléatoire. Nous effectuons donc 10 initialisations différentes dans ces cas précis. Or les 10 runs de M5A engendrent des valeurs stables pour les différents critères (cf. Tableau 4.29). Ces observations sont également valables pour M5B, M5C, M5D et M5I. Nous nous fixons donc sur l'étude de la moyenne des runs pour la suite de l'analyse.

	Critère	Résultats pour chacun des 10 runs pour M5A
	Rayon max.	14.15 14.15 14.15 14.15 14.15 14.15 14.15 14.15 14.15 14.15
	Rayon moyen.	12.97 12.95 12.94 12.94 12.95 12.96 12.98 13.02 12.91 12.93
	EC Rayon	1.06 1.09 1.06 1.09 1.08 1.07 1.07 1.08 1.08 1.07
	EC Ind	181.47 274.32 446.69 151.08 183.84 354.62 176.14 198.36 373.63 205.91
NN [a]	Somme	16.02×10^3 16.19×10^3 16.11×10^3 15.98×10^3 16.07×10^3 15.92×10^3 16.07×10^3 16.05×10^3 16.01×10^3 16.18×10^3
	Min	14.16 14.16 14.16 14.16 14.16 14.16 14.16 14.16 14.16 14.16
	Max	127.69 127.69 127.69 127.69 127.69 127.69 127.69 127.69 127.69 127.69
	EC	5.22 5.19 5.21 5.22 5.21 5.23 5.21 5.21 5.22 5.19
Dep [b]	Somme	442.52×10^3 438.90×10^3 429.73×10^3 442.35×10^3 439.17×10^3 433.29×10^3 445.43×10^3 444.10×10^3 432.80×10^3 437.66×10^3
	Min	0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01
	Max	14.15 14.15 14.15 14.15 14.15 14.15 14.15 14.15 14.15 14.15
	EC	2.07 2.07 2.10 2.05 2.06 2.10 2.08 2.10 2.08 2.09
Tot [c]	Somme	18.91×10^6 19.29×10^6 19.10×10^6 18.80 18.99×10^6 18.69×10^6 18.99×10^6 18.95×10^6 18.86×10^6 19.28×10^6
	Min	14.16 14.16 14.16 14.16 14.16 14.16 14.16 14.16 14.16 14.16
	Max	201.77 201.77 201.77 201.77 201.77 201.77 201.77 201.77 201.77 201.77
	EC	16.04 15.97 16.01 16.06 16.04 16.07 16.04 16.01 16.03 15.99
	Nb singletons	298.00 306.00 304.00 311.00 309.00 304.00 301.00 306.00 307.00 309.00

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.29: Tableau des résultats sur 10 runs pour la méthode k-medoïds (M5A)

4.4.5.2 Résultats sur la moyenne des 10 runs (jeu 1 - k = 1000)

Rayon et densité des groupes Nous remarquons (cf. Tableau 4.30) que pour n'importe quel paramétrage, les résultats sont quasiment identiques pour le rayon maximum et le rayon moyen. En revanche pour l'écart moyen du nombre d'individus par groupe, il existe des différences. Pour une même initialisation (respectivement aléatoire, au centre, au plus éloigné), l'EC Ind garde le même ordre de grandeur (environ 200 pour l'initialisation aléatoire, environ 500 pour une initialisation au centre, et entre 140 et 230 pour l'initialisation au plus éloigné). Le choix de la molécule suivante ne modifie donc pas les résultats. En revanche pour un même critère de choix (par exemple : MinSum), l'initialisation au centre donne un écart d'individus dans les groupes plus élevé qu'avec les autres initialisations. Comme expliqué en section 4.2.1.2, l'initialisation aléatoire semble donc donné, par cet écart plus élevé un meilleur quadrillage de l'espace.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	
M5A	14.15	12.95	1.07	254.61	initialisation aléatoire
M5B	14.15	12.96	1.06	197.52	
M5C	14.15	12.96	1.06	226.98	
M5D	14.15	12.96	1.07	186.31	
M5E	14.15	12.94	1.08	575.35	initialisation au centre
M5F	14.15	12.91	1.08	559.71	
M5G	14.15	12.91	1.08	559.71	
M5H	14.15	12.91	1.08	559.71	
M5I	14.16	12.96	1.07	232.65	initialisation au plus loin
M5J	14.17	13.00	1.09	141.17	
M5K	14.17	13.00	1.09	141.17	
M5L	14.17	13.00	1.09	141.17	

TABLE 4.30: Résultats du critère Rayon pour la méthode Sphere-Exclusion (M5) - Jeu 1, k=1000 ; aléatoire, MinSum, MinMin, MaxMin

Diversité/Recouvrement de l'espace Nous remarquons ici (cf. Tableau 4.31) que pour n'importe quel paramétrage (initialisation et choix de la molécule suivante) les résultats sont quasiment identiques : SDNN = 16100 environ, SDNN Min = 14, SDNN Max = 127.69 et SDNN EC = 5.20 environ. Ni l'initialisation, ni le critère de choix de la molécule suivante à sélectionner n'ont donc d'impact sur le critère d'évaluation SDNN. De plus, la distance minimum entre deux centres voisins (SDNN Min = 14.16) est supérieure ou égale au rayon maximum et supérieur au rayon moyen. Ceci signifie qu'aucun centre n'est plus proche de son voisin que la taille du plus grand groupe. Ceci peut donc nous indiquer qu'il n'y a pas de redondance dans les représantants et que le quadrillage de l'espace est homogène. Enfin la distance maximum entre deux centres voisins (SDNN Max = 127.69) est élevée, indiquant que les extrêmes de l'espace sont également couverts par l'échantillon.

Lorsque l'on étudie la distribution de ces distances entre centres voisins (cf. Figure 4.26), on confirme les observations précédentes. En effet pour avoir un quadrillage homogène de l'espace il faut que la majorité des centres soit à une distance homogène de leur plus proche. Or sur cette figure nous observons que la majorité des distances se situe autour du rayon maximum et du rayon moyen et de plus aucune ne se situe avant ce rayon moyen.

Enfin on remarque que M5F, M5G et M5H ont exactement les mêmes résultats comme M5J, M5K et M5L.

NN [a]	Somme	Min	Max	EC	
M5A	16.06×10^3	14.16	127.69	5.21	intialisation aléatoire
M5B	16.10×10^3	14.16	127.69	5.21	
M5C	16.13×10^3	14.16	127.69	5.20	
M5D	16.14×10^3	14.16	127.69	5.20	
M5E	16.06×10^3	14.15	127.69	5.21	initialisation au centre
M5F	16.10×10^3	14.15	127.69	5.20	
M5G	16.10×10^3	14.15	127.69	5.20	
M5H	16.10×10^3	14.15	127.69	5.20	
M5I	16.10×10^3	14.16	127.69	5.20	initialisation au plus loin
M5J	16.09×10^3	14.17	127.69	5.22	
M5K	16.09×10^3	14.17	127.69	5.22	
M5L	16.09×10^3	14.17	127.69	5.22	

a. $\text{NN} = \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

TABLE 4.31: Résultats du critère SDNN pour la méthode Sphere-Exclusion (M5) - Jeu 1, k=1000 ; aléatoire, MinSum, MinMin, MaxMin

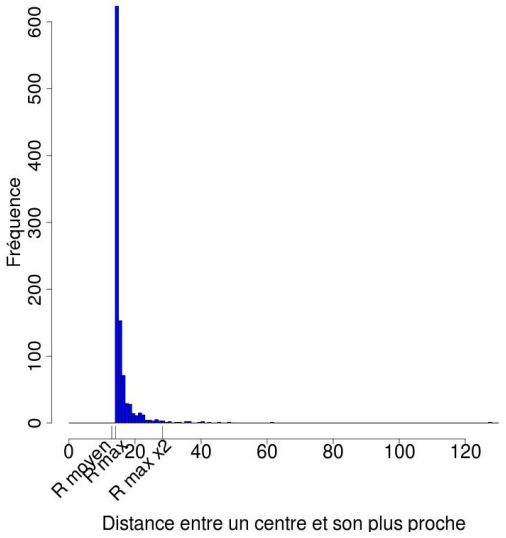


FIGURE 4.26: Pour M5 Sphere-Exclusion, Distribution des dissimilarités entre chaque centre et son centre le plus proche (SDNN)

Représentativité Tout d'abord nous étudions les résultats pour une même intialisation (par exemple : initialisation aléatoire pour M5A, M5B, M5C et M5D, cf. Tableau 4.32). La somme des dissimilarité entre chaque molécule du jeu de départ et son représentant dans l'échantillon (438598.69, 442101.59, 440838.19, 442557.75) reste du même ordre de grandeur quelque soit le paramètre de choix de la molécule suivante utilisé. Il en va de même pour les autres critères d'évaluation (SDDep Min, SDDep Max et SDDep EC). Cette

observation est également valable pour l'initialisation au centre et l'initialisation au plus éloigné.

Ensuite nous analysons l'impact d'une initialisation différente, par exemple pour le critère de choix de la molécule suivante par tirage aléatoire nous comparons : M5A, M5E et M5I (cf. Tableau 4.32, couleur :). La somme des dissimilarités entre chaque molécule du jeu de départ et son représentant le plus proche (SomDissimNN-Dep = 438598.69, 418589.47 et 440298.41) est meilleure pour l'initialisation au centre. Pour les autres critères de choix de la molécule suivante, l'initialisation au centre donne également de meilleurs résultats que les autres inititialisations.

Nous avons donc vu que le critère de choix de la molécule n'a pas d'impact sur la représentativité de l'échantillon, en revanche l'initialisation au centre donne un échantillon plus représentatif que les autres intialisations quelque soit le critère de choix de la molécule suivante. Enfin la distribution des distances entre chaque molécule et son représentant (cf. Figure 4.27) est concentrée avant le rayon maximum voire avant le rayon moyen qui sont faibles. Ceci est donc le signe d'une bonne représentativité de l'échantillon.

Dep [a]	Somme	Min	Max	EC	
M5A	438.59×10^3	0.01	14.15	2.08	initialisation aléatoire
M5B	442.10×10^3	0.01	14.15	2.08	
M5C	440.83×10^3	0.01	14.15	2.08	
M5D	442.55×10^3	0.01	14.15	2.08	
M5E	418.58×10^3	0.01	14.15	2.20	initialisation au centre
M5F	416.47×10^3	0.01	14.15	2.20	
M5G	416.47×10^3	0.01	14.15	2.20	
M5H	416.47×10^3	0.01	14.15	2.20	
M5I	440.29×10^3	0.01	14.16	2.08	initialisation au plus loin
M5J	445.88×10^3	0.01	14.17	2.07	
M5K	445.88×10^3	0.01	14.17	2.07	
M5L	445.88×10^3	0.01	14.17	2.07	

a. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

TABLE 4.32: Résultats du critère SDDep pour la méthode Sphere-Exclusion (M5) - Jeu 1, k=1000 ; aléatoire, MinSum, MinMin, MaxMin

Dissimilarité totale dans l'échantillon Nous voyons ici que quelque soit le paramètre de choix de la molécule suivante et quelque soit le type d'initialisation utilisé, les résultats sont similaires voire identiques pour tous les critères d'évaluation sur la dissimilarité totale de l'échantillon (cf. Tableau 4.33). En effet la somme des dissimilarités entre chaque centre deux à deux (SDTot) est comprise entre 18,99 millions et 19,14 millions, la dissimilarité minimum entre deux centres est de 14.16 et la dissimilarité maximum entre deux centres est de 201.77. Ce dernier critère étant élevé, il confirme que les extrêmes sont couverts par l'échantillon.

Enfin lorsqu'on étudie la ditribution des distances deux à deux dans cet échantillon (cf. Figure 4.28), on observe une concentration des distances autour du rayon maximum et de deux fois ce rayon, ce qui prouve une bonne couverture homogène de l'espace central des descriptions. Aucune ne se trouve sous le rayon moyen comme nous l'avons remarqué précédemment. Enfin il existe quelques distances extrêmes indiquant la couverture des

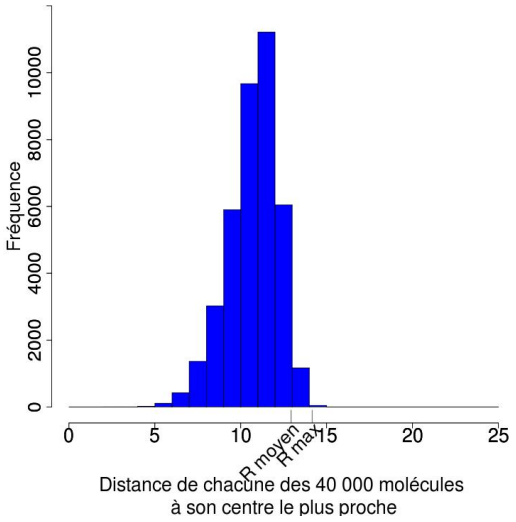


FIGURE 4.27: Pour M5 Sphere-Exclusion, Distribution des dissimilarités entre chaque molécule et son centre le plus proche (SDDep)

extrêmes dans une moindre mesure.

Tot [a]	Somme	Min	Max	EC	
M5A	18.99×10^6	14.16	201.77	16.03	initialisation aléatoire
M5B	19.07×10^6	14.16	201.77	16.02	
M5C	19.14×10^6	14.16	201.77	16.01	
M5D	19.15×10^6	14.16	201.77	16.01	
M5E	19.04×10^6	14.15	201.77	16.00	initialisation au centre
M5F	19.13×10^6	14.15	201.77	16.00	
M5G	19.13×10^6	14.15	201.77	16.00	
M5H	19.13×10^6	14.15	201.77	16.00	
M5I	19.07×10^6	14.16	201.77	16.02	initialisation au plus loin
M5J	19.03×10^6	14.17	201.77	16.02	
M5K	19.03×10^6	14.17	201.77	16.02	
M5L	19.03×10^6	14.17	201.77	16.02	

a. Tot $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.33: Résultats du critère SDTot pour la méthode Sphere-Exclusion (M5) - Jeu 1, k=1000 ; aléatoire, MinSum, MinMin, MaxMin

Singletons Pour le nombre de groupe ne contenant qu'un seul individu, on remarque également que ni le critère de choix de la molécule suivante, ni le type d'initialisation n'ont d'impact sur cette valeur. Nous avons donc un nombre élevé de singletons (cf. Tableau 4.34) quelque soit le paramèétrage de Sphere-Exclusion. Enfin en étudiant la distribution des distances entre chaque singletons et sa molécule la plus proche (cf. Figure 4.29), on observe que la plupart de ces distances se situent dans les distances extrêmes du jeu total. Et toutes les distances les plus extrêmes entre une molécule et sa plus proche sont couvertes par des singletons. La plupart d'entre eux sont donc de bons indicateurs d'outliers.

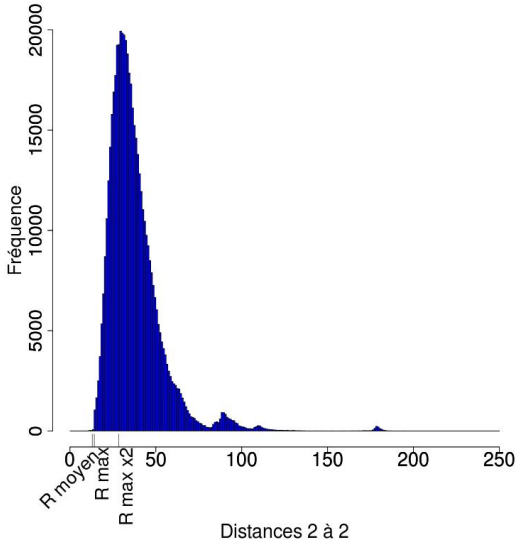


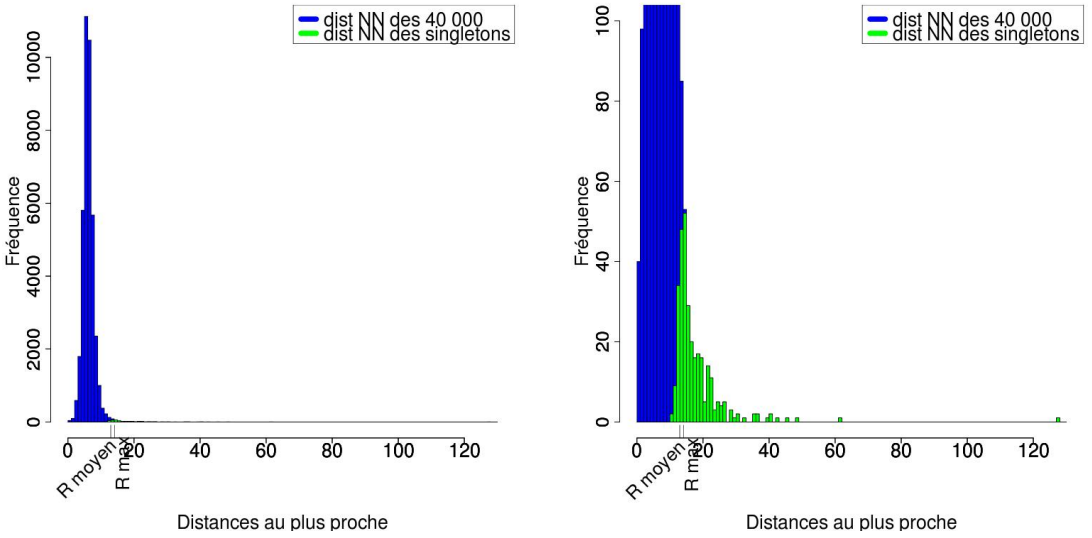
FIGURE 4.28: Distribution des dissimilarités entre tous les centres, SDTot, pour M5 Sphere-Exclusion

	Singletons	
M5A	305.50	initialisation aléatoire
M5B	305.10	
M5C	305.50	
M5D	305.00	
M5E	306.90	initialisation au centre
M5F	310.00	
M5G	310.00	
M5H	310.00	
M5I	308.60	initialisation au plus loin
M5J	300.00	
M5K	300.00	
M5L	300.00	

TABLE 4.34: Résultats du critère Singletons pour Sphere-Exclusion (M5) - Jeu 1, k=1000 ; aléatoire, MinSum, MinMin, MaxMin

Conclusion Nous avons vu que pour la méthode Sphere-Exclusion, le critère de choix de la molécule suivante (aléatoire, MinSum, MinMin, MaxMin) n'a aucune influence sur l'échantillon obtenu. Ils donnent tous de bons résultats. En revanche, une initialisation à l'individu central du jeu de départ peut donner de meilleurs résultats pour certains critères d'évaluations. Comme pour la méthode Maximum-Dissimilarity, cette initialisation a une influence sur le nombre d'individus dans les groupes et la représentativité des échantillons. Enfin il semble que pour une initialisation aléatoire, les critères de choix non aléatoires donnent exactement les mêmes résultats quelque soit le critère. En effet M5F = M5G = M5H et M5J = M5K =M5L. Or il s'avère que les échantillons produits par M5F, M5G et M5H sont identiques exactement. Il en est de même pour J, K et L.

De même que pour la méthode Maximum-Dissimilarity, la méthode Sphere-Exclusion a la même temps de calcul pour tous les paramétrages de cette méthode, soit 30 à 40



(a) M5 Singletons histogramme complet

(b) M5 Singletons échelle réduite

FIGURE 4.29: Pour M5 Sphere-Exclusion, Distribution des dissimilarités entre chaque molécule et sa molécule la plus proche et projection de ces distances pour les singletons

secondes.

4.4.6 Comparaison des méthodes

Au cours de notre analyse des différentes méthodes nous avons fait ressortir quelques paramétrages meilleurs que d'autres. Nous comparerons donc pour chaque méthode le ou les paramétrages les plus performants autres méthodes. Dans cette partie nous étudions donc les différences entre les méthodes suivantes :

- M1, tirage aléatoire uniforme sans remise.
- M2A, notre implémentation du k-center avec une initialisation aléatoire
- M3B, k-medoïds avec une intialisation aléatoire et des itérations avec les centres virtuels
- M3C, k-medoïds avec une initialisation avec FFT et des itérations avec les centres réels
- M4D, Maximum-Dissimilarity ou FFT avec initialisation à la molécule centrale
- M5F, Sphere-Exclusion avec une initialisation à la molécule centrale et le critère MaxSum (résultats égaux aux critères MinMin et MinMax)

Rayon et densité des groupes Pour le critère rayon maximum (cf.Tableau 4.35), la méthode k-center (M2A = 12.68) donne les meilleurs résultats, ainsi que pour le rayon moyen (M2A= 11.81). Les méthodes les plus proches de k-center sont les méthodes FFT et Sphere-Exclusion (avec respectivement le rayon maximum égal à 13.99 et 14.15). La méthode k-medoïds avec initialisation à FFT (M3C) reste dans le même ordre de grandeur. En revanche, on remarque que k-medoïds avec initialisation aléatoire et la méthode de tirage aléatoire donne les pires résultats en terme de rayon.

Enfin les valeurs de l'écart moyen du nombre d'individus par groupe sont faibles et donc mauvaises également pour ces deux dernières méthodes citées. En revanche EC Ind est plus

élevé pour M4D et M5F que pour M2A. L'étude des critères suivants nous indiquera donc laquelle de ses méthodes donne le quadrillage le plus homogène.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind
M1	159.25	14.41	9.58	31.31
M2A	12.68	11.81	1.02	177.76
M3B	109.18	12.91	6.32	22.99
M3C	16.67	12.19	1.18	162.93
M4D	13.99	12.97	1.03	585.84
M5F	14.15	12.91	1.08	559.71

TABLE 4.35: Résultats pour le critère Rayon pour les 6 méthodes comparées

Diversité/Recouvrement de l'espace Encore une fois, les méthodes de tirage aléatoire (M1) et k-medoïds avec initialisation aléatoire (M3B) donnent des résultats similaires et moins bons que les autres (SDNN = 8000 environ, cf. Tableau 4.36). Ensuite si l'on étudie la somme des dissimilarités entre centres plus proches, on remarque que les méthodes M4D et M5F (respectivement FFT et Sphere-Exclusion) donnent de meilleurs résultats que la méthode k-center (M2A) puis k-medoïds avec initialisation à FFT. En effet cette somme des dissimilarités SDNN est plus élevée pour M4D et M5F que pour M2A et M3C. De plus la plus petite distance entre deux centres proches (M4D = 13.99 et M5F =14.15) est plus élevée et égale au rayon maximum des méthodes concernées que pour les méthodes M2A et M3C (où SDNN Min= 6.97 et 6.15 respectivement). Il semble que les méthodes FFT et Sphere-Exclusion donnent des échantillons plus homogènement répartis dans l'espace que les méthodes k-center et k-medoïds.

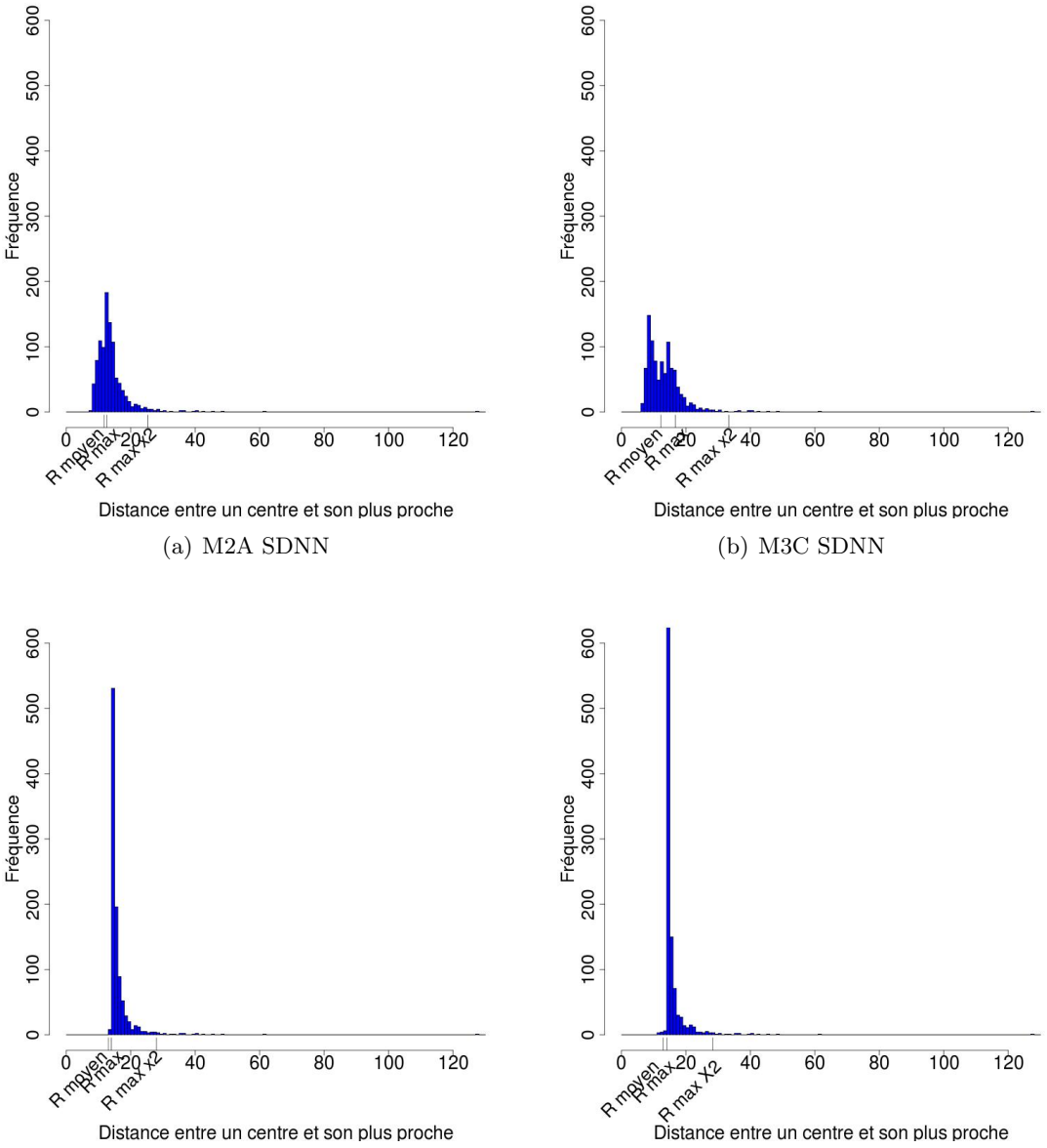
Cette observation se confirme avec la comparaison de la distribution de ces distances (cf. Figure 4.30). En effet on observe que les méthodes FFT (4.30(c)) et Sphere-Exclusion (4.30(d)) ont des distances extrêmement concentrées autour du rayon maximum et n'en ont aucune avant le rayon moyen. Or pour M2A (4.30(a)), les distances sont plus dispersées et certaines sont inférieures au rayon moyen. La distribution pour M3C (4.30(b)) est pire encore avec beaucoup de distances situées avant le rayon moyen.

Les méthodes FFT et Sphere-Exclusion semblent donc donner des échantillons quadrillant l'espace des descriptions de manière plus homogène que k-center puis k-medoïds.

NN [a]	Somme	Min	Max	EC
M1	8.61×10^3	3.71	50.00	2.86
M2A	14.05×10^3	6.97	127.69	5.96
M3B	8.27×10^3	4.09	106.17	4.70
M3C	13.35×10^3	6.15	127.69	6.52
M4D	16.17×10^3	13.99	127.69	5.22
M5F	16.10×10^3	14.15	127.69	5.20

a. $\text{NN} = \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

TABLE 4.36: Résultats pour le critère SDNN pour les 6 méthodes comparées



(a) M2A SDNN

(b) M3C SDNN

(c) M4D SDNN

(d) M5F SDNN

FIGURE 4.30: Pour M2A, M3C, M4D et M5F, Distribution des dissimilarités entre chaque centre et son centre le plus proche (SDNN)

Représentativité La somme des dissimilarités entre chaque molécule et son représentant est maximale pour les méthodes FFT et Sphere-Exclusion (M4D et M5F = 416 000 environ, cf. Tableau 4.37). Les meilleures valeurs pour SDDep sont données par la méthode k-medoïds avec initialisation aléatoire suivie de la méthode de tirage aléatoire (M3B = 315 000 environ et M1 = 335 000 environ). On pourrait donc penser que ces deux méthodes donnent des échantillons plus représentatifs que les autres. Or lorsqu'on observe la distance maximale existant entre une molécule et son représentant (également appelée rayon maximum), on s'aperçoit que celle-ci est très élevée pour ces méthodes (M3C = 109.18 et M1 = 159.25). Ceci signifie donc que l'échantillon produit par tirage aléatoire ou par

k-medoïds avec initialisation aléatoire n'est pas représentatif de l'ensemble des molécules de départ, or nous l'avons vu dans l'étude précédente, il est représentatif des zones les plus denses de l'espace.

Enfin lorsqu'on étudie la distribution des distances entre chaque molécule et son représentant nous confirmons les présomptions précédentes. Les méthodes M1 et M3B donnent trop de distances supérieures au rayon moyen et supérieures à 25 (cf. Figures 4.31(b) et 4.31(d)). Ensuite M3C , la méthode k-medoïds avec initialisation à FFT (cf. Figure 4.32(b)) donne une somme des dissimilarités plus faible que la méthode k-center , M2A (cf. Figure(4.32(a))). Cependant cette dernière donne des distances ne dépassant pas 13. Les distributions de distances pour les méthodes Maximum-Dissimilarity et Sphere-Exclusion (cf. Figures 4.32(c) et 4.32(d) respectivement) sont moins bonnes également que celle de la méthode k-center. Il en ressort donc que la méthode donnant un échantillon le plus représentatif de l'ensemble de départ dans sa globalité, est la méthode k-center.

Dep [a]	Somme	Min	Max	EC
M1	335.34×10^3	0.01	159.25	3.23
M2A	362.14×10^3	0.10	12.68	1.89
M3B	315.50×10^3	0.01	109.18	2.70
M3C	341.61×10^3	0.01	16.67	1.93
M4D	419.57×10^3	0.01	13.99	2.20
M5F	416.47×10^3	0.01	14.15	2.20

a. Dep $= \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

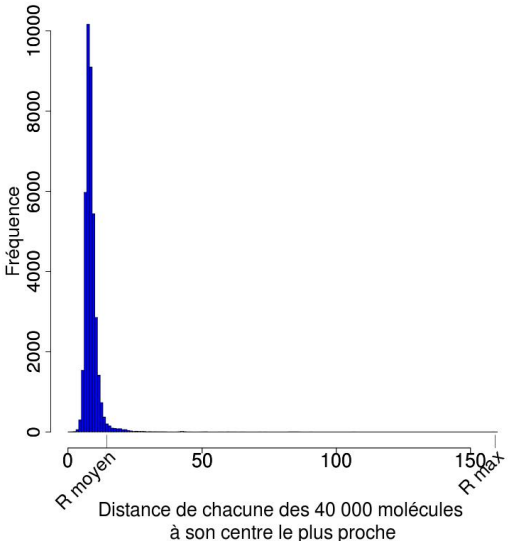
TABLE 4.37: Résultats pour le critère SDDep pour les 6 méthodes comparées

Dissimilarité totale de l'échantillon Les méthodes M1 (tirage aléatoire) et M3B (k-medoïds avec initialisation aléatoire) donne des sommes de dissimilarités entre centres deux à deux, deux fois inférieures à celles des autres méthodes (cf. Tableau 4.38). De plus la distance maximale entre deux centres est plus faible (77.95 et 148.57 respectivement) que pour M3C, M2A, M4D et M5F. Ces deux méthodes sont donc encore celles qui donnent les moins bons résultats. Les 4 autres méthodes donnent des résultats du même ordre de grandeur. La distribution des distances étant sensiblement la même pour ces quatre méthodes (cf. Figure 4.33), nous considérons que la méthode M5F (Sphere-Exclusion) ayant la plus forte somme des dissimilarités est la meilleure pour ce critère.

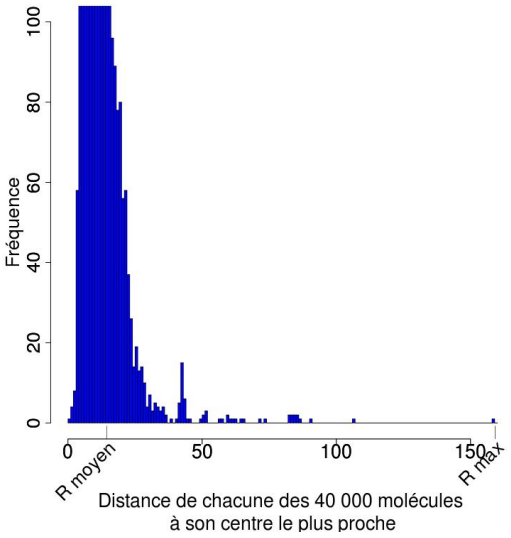
Tot [a]	Somme	Min	Max	EC
M1	9.10×10^6	3.71	77.95	6.21
M2A	18.52×10^6	6.97	201.77	16.72
M3B	9.23×10^6	4.09	148.57	9.11
M3C	17.73×10^6	6.15	201.77	16.89
M4D	18.73×10^6	13.99	201.77	16.04
M5F	19.13×10^6	14.15	201.77	16.00

a. Tot $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

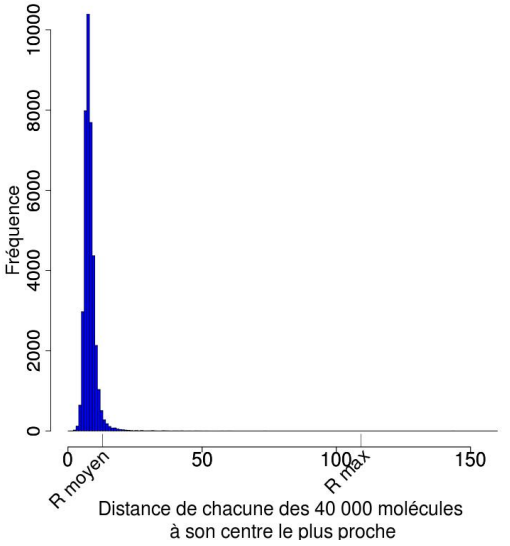
TABLE 4.38: Résultats pour le critère SDTot pour les 6 méthodes comparées



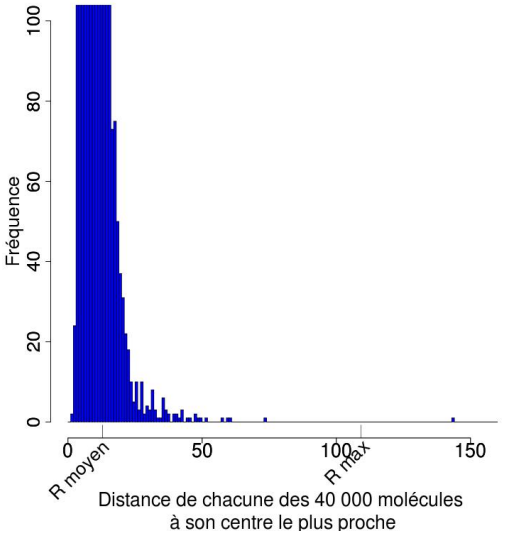
(a) M1 SDDep histogramme complet



(b) M1 SDDep échelle réduite

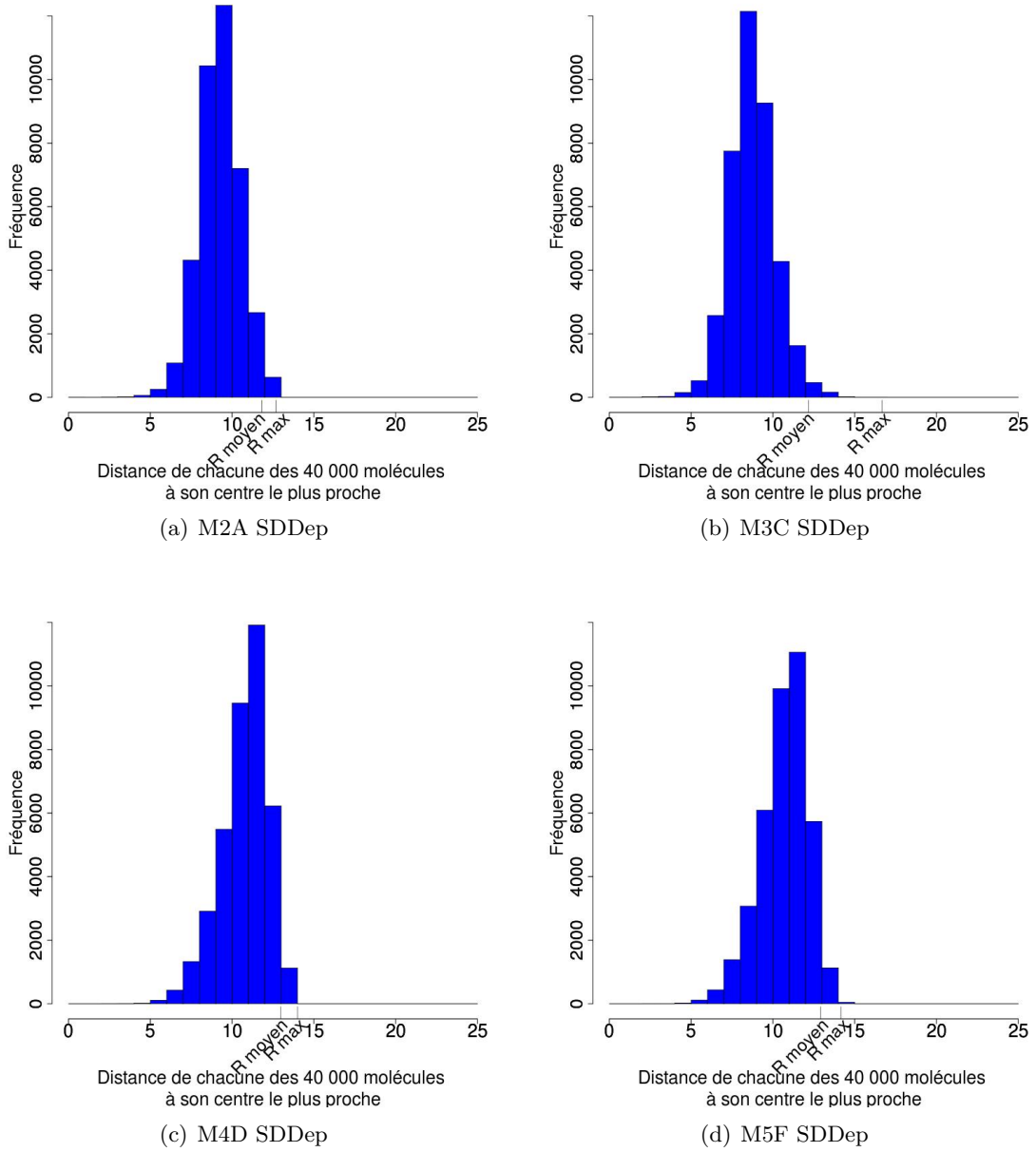


(c) M3B SDDep histogramme complet



(d) M3B SDDep échelle réduite

FIGURE 4.31: Pour M1 et M3B, Distribution des dissimilarités entre chaque molécule et son centre le plus proche (SDDep)



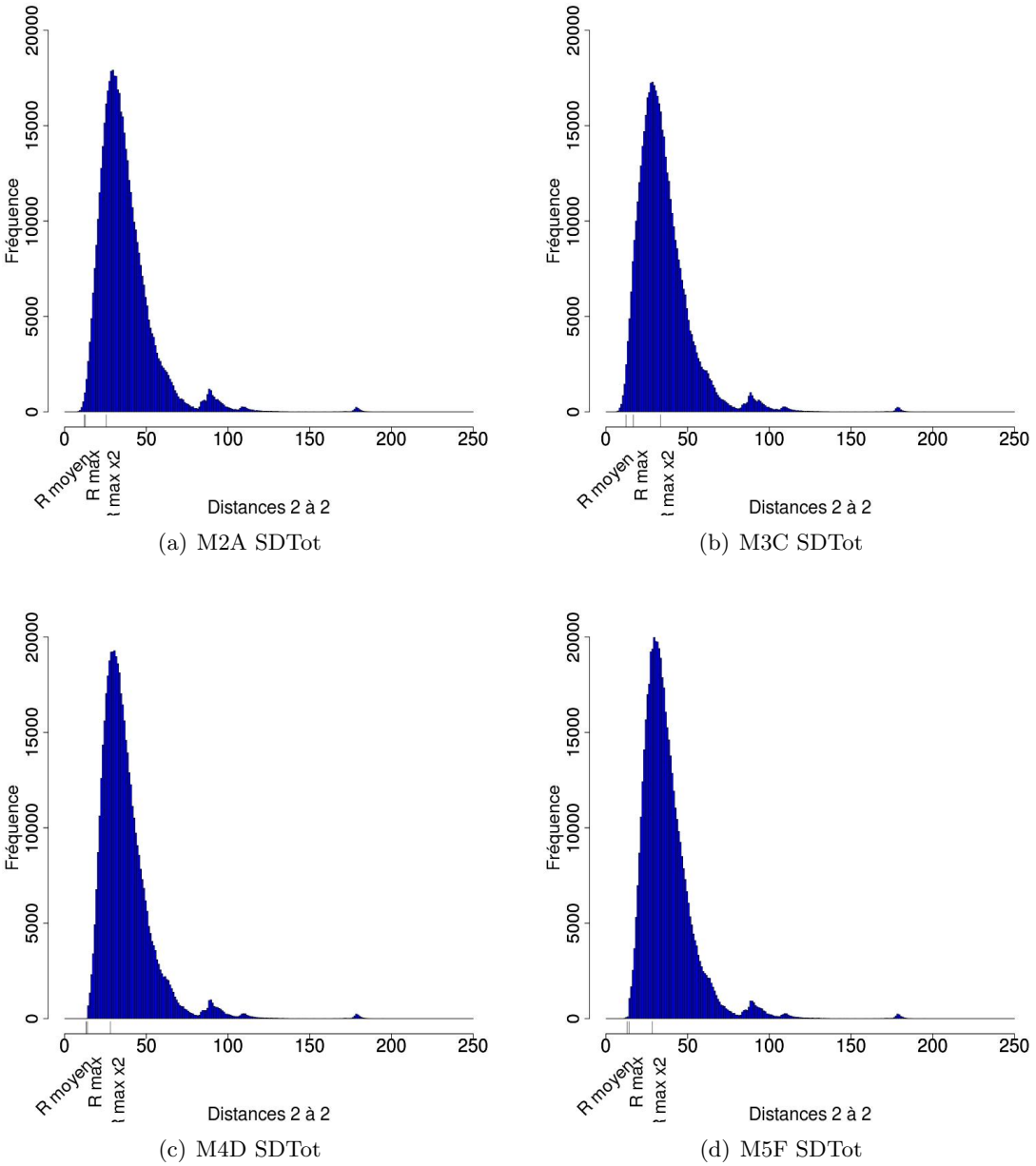
(a) M2A SDDep

(b) M3C SDDep

(c) M4D SDDep

(d) M5F SDDep

FIGURE 4.32: Pour M2A, M3C et M4D et M5F, Distribution des dissimilarités entre chaque molécule et son centre le plus proche (SDDep)



(a) M2A SDTot

(b) M3C SDTot

(c) M4D SDTot

(d) M5F SDTot

FIGURE 4.33: Pour M2A, M3C, M4D et M5F, Distribution des dissimilarités entre tous les centres deux à deux (SDTot)

Singletons Nous avons vu précédement que les méthodes M1 et M3B ne donnent que très peu de singletons (cf. Tableau 4.39) et que ceux-ci ne sont pas de bons indicateurs d'outliers. Pour les autres méthodes, on remarque que k-center (M2A) en donne le plus. Nous étudions ensuite la distribution de leur distance à la molécule la plus proche pour déterminer quelle méthode est la plus performante pour fournir de bons outliers (cf. Figure 4.34). Notons que pour une meilleure clarté nous ne donnons pas les histogrammes complets déjà vus dans les sections précédentes puisqu'ils n'apportent pas d'information supplémentaire. On remarque alors que M2A et M4D couvrent mieux les extrêmes que les deux autres méthodes. De plus les distances avant le rayon maximum sont moins couvertes par les singletons de M2A puis de M4D. La méthode M2A donne donc des singletons meilleurs pour indiquer les outliers.

	Singletons
M1	3.40
M2A	427.70
M3B	3.60
M3C	326.00
M4D	309.00
M5F	310.00

TABLE 4.39: Résultats pour le critère Singletons pour les 6 méthodes comparées

Méthodes	Rayon	SDNN	SDDep	SDTot	Singletons	Position Totale
M1	6	6	6	6	6	**6**
M2	1	3	1	2	1	**1.5**
M3B	5	5	5	5	5	**5**
M3C	4	4	2	2	4	**3.2**
M4	2	1	3	2	2	**2**
M5	2	1	3	1	2	**1.8**

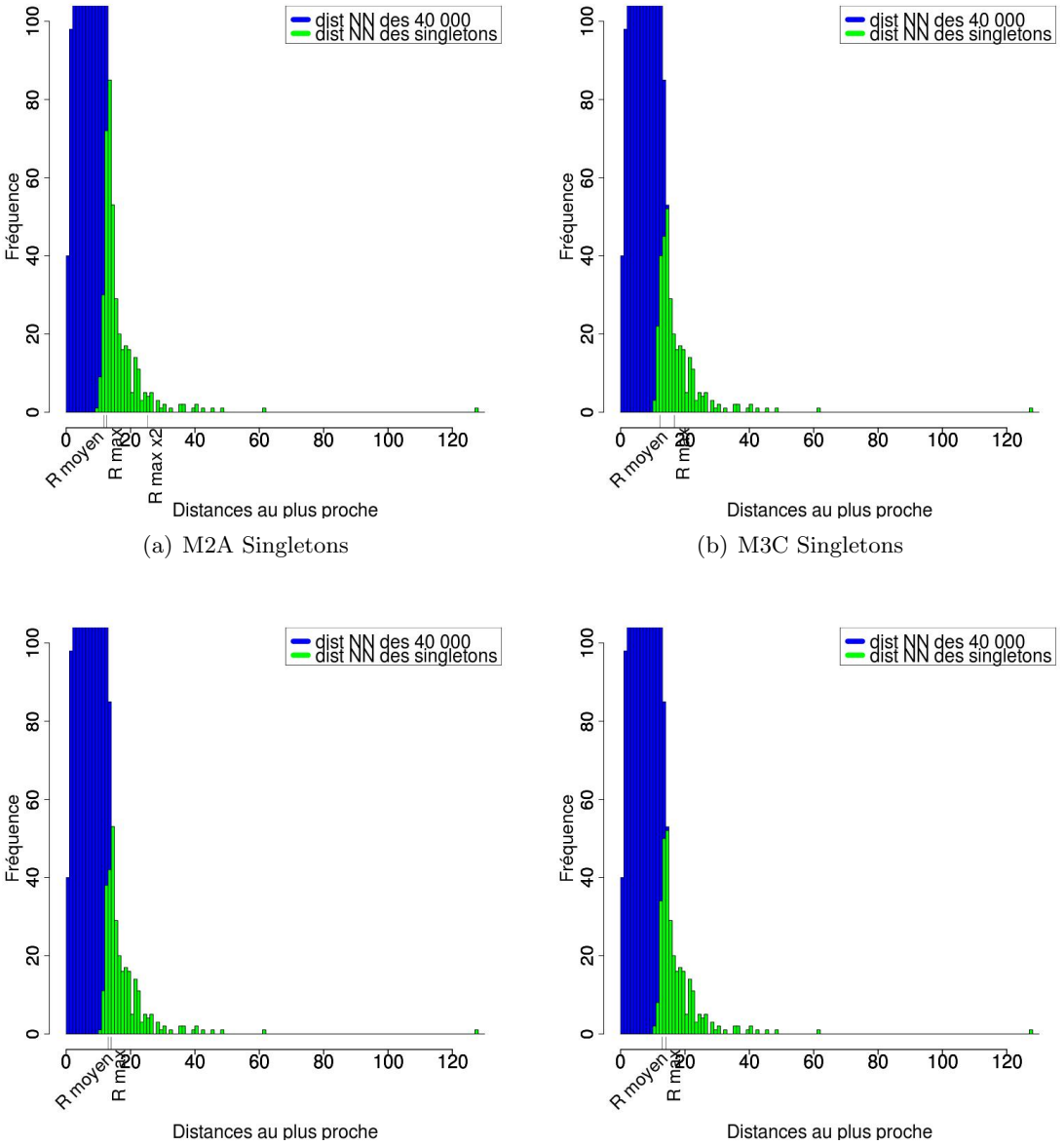
TABLE 4.40: Récapitulatif des positions de chaque méthode par critère

Conclusion En conclusion, les méthodes de tirage aléatoire uniforme (M1) et k-medoïds avec initialisation aléatoire (M3B) donnent les résultats les moins bons pour tous les critères d'évaluation (cf. Tableau 4.40). Ces deux méthodes ne produisent donc pas de bons échantillons divers et leur utilisation dans ce but est à proscrire.

Ensuite notre implémentation du problème k-center (M2A) donne un meilleur rayon maximum et moyen des groupes, un échantillon plus représentatif de l'ensemble de départ et des singletons qui sont de bons indicateurs d'outliers. En revanche pour un quadrillage homogène de l'espace il est préférable d'utiliser la méthode FFT avec initialisation au centre (M4D) et la méthode Sphere-Exclusion (M5) ainsi que pour maximiser la somme des dissimilarités totales de l'échantillon. Sachant que M5 requiert un rayon seuil comme paramètre en entrée, il est plus simple d'utiliser FFT.

Globalement, notre implémentation de k-center correspond au maximum du critère que l'on souhaite obtenir pour un échantillon divers. Mais si l'on veut un quadrillage plus homogène de l'espace des descriptions, on se tournera vers FFT.

Enfin en temps de calcul, les méthodes Maximum-Dissimilarity et Sphere-Exclusion sortent du lot avec des temps de sélection quasi instantannés (30 secondes). Arrive ensuite



(a) M2A Singletons

(b) M3C Singletons

(c) M4D Singletons

(d) M5F Singletons

FIGURE 4.34: Pour M2A, M3C, M4D et M5F, Distribution des distances entre chaque molécule et sa plus proche et projection de ces distances pour les singletons

la méthode k-medoïds avec un temps de réponse de 3 à 17 minutes. Et enfin la méthode la plus longue est k-center avec 54 minutes de temps de sélection. Cependant ce temps reste raisonnable.

4.5 Impact du jeu initial avec outliers sur l'échantillonnage (jeux : J1 versus J3 versus J5 - k = 1000)

Pour vérifier la robustesse de nos méthodes, nous les avons testées sur plusieurs jeux différents. Nous rappelons que les jeux J1, J3 et J5 sont trois jeux non chevauchants avec outliers tirés aléatoirement dans une base de plusieurs millions de composés.

Si nos méthodes sont robustes, on s'attend à ce que les conclusions de la section précédente valables pour un échantillonnage de 1000 molécules du jeu J1 soient également valables pour un échantillonnage de 1000 molécules du jeu J3 et du jeu J5. Comme nous l'avons dit précédemment, seuls certains paramétrages pour chaque méthode donnent les meilleurs résultats. Ainsi pour une meilleure lisibilité, bonne compréhension et afin de limiter la taille des tableaux, seuls les paramétrages suivants seront mis en avant pour comparer J1 à J3 et J5 :

- M1, tirage aléatoire uniforme sans remise.
- M2A, notre implémentation du k-center avec une initialisation aléatoire
- M3C, k-medoïds avec une initialisation avec FFT et des itérations avec les centres réels
- M4D, Maximum-Dissimilarity ou FFT avec initialisation à la molécule centrale
- M5F, Sphere-Exclusion avec une initialisation à la molécule centrale et le critère MaxSum (résultats égaux aux critères MinMin et MinMax)

Toutefois, les tableaux complets de chaque méthode pour les jeux J3 et J5 sont disponibles en annexe et seront commentés dans cette section.

4.5.1 Sélection par tirage aléatoire (M1)

En ce qui concerne le rayon maximum, les jeux J3 et J5 produisent un échantillon dont le rayon maximum est légèrement supérieur à celui produit par J1 (cf. Tableau 4.41). Cependant l'ordre de grandeur est respecté. Le rayon moyen ainsi que l'écart du nombre d'individus par groupe sont très similaires d'un jeu à l'autre. Il en est de même pour les autres critères (SDNN, SDDep et SDTot). Seuls les maximums diffèrent légèrement (SDNN max, SDDep max et SDTot max). Ceci s'explique par le fait que les jeux étant différents, certains outliers produisent des distances plus ou moins élevées en fonction du jeu.

Globalement, la méthode de sélection par tirage aléatoire ne semble pas influencée par le jeu initial puisque les résultats sont similaires quelque soit le jeu donné en entrée.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M1 J1	159.2	14.4	9.5	31.3	8.61×10^3	3.7	50.0	2.8
M1 J3	204.6	14.7	10.3	31.4	8.55×10^3	2.9	39.9	2.5
M1 J5	213.6	14.3	13.4	30.8	8.65×10^3	3.7	51.7	2.6

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M1 J1	335.34×10^3	0.01	159.2	3.2	9.10×10^6	3.7	77.9	6.2	3.4
M1 J3	335.93×10^3	0.01	204.6	3.2	9.06×10^6	2.9	74.0	6.0	3.6
M1 J5	338.69×10^3	0.01	213.6	3.4	9.07×10^6	3.7	87.7	5.8	4.0

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.41: Résultats pour le tirage aléatoire (M1), Jeux : J1, J3 et J5, k = 1000

4.5.2 Sélection par *k*-center (M2)

Tout d'abord, pour le jeu J3 et pour le jeu J5, les paramétrages M2A et M2B (respectivement initialisation aléatoire et initialisation avec FFT) donnent des résultats similaires comme vu également pour le jeu J1 (cf. Annexe C, Tableaux C.1 et C.2). Ensuite, quelque soit le jeu, les critères rayon maximum, rayon moyen, SDNN, SDDep et SDTot sont tous du même ordre de grandeur. On remarque que le jeu J5 produit un SDNN et un SDTot légèrement plus faible que les autres. Comme ceux-ci restent dans le même ordre de grandeur que pour les autres jeux, on en déduit que ce résultat est dû à la différence de composés contenus dans le jeu.

En général, les résultats étant très similaires, on en déduit que la méthode k-center n'est pas influencée par le jeu initial.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M2A J1	12.6	11.8	1.0	177.7	14.05×10^3	6.9	127.6	5.9
M2A J3	12.8	11.9	1.1	186.6	14.19×10^3	7.3	115.3	5.6
M2A J5	12.4	11.6	0.9	142.7	13.23×10^3	6.8	172.5	7.3

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M2A J1	362.14×10^3	0.1	12.6	1.8	18.52×10^6	6.9	201.7	16.7	427.7
M2A J3	365.62×10^3	0.1	12.8	1.9	18.31×10^6	7.3	241.7	16.2	429.0
M2A J5	361.16×10^3	0.1	12.4	1.8	16.96×10^6	6.8	280.6	21.4	333.0

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.42: Résultats pour la méthode k-center (M2A), Jeux : J1, J3 et J5, k = 1000

4.5.3 Sélection par *k*-medoïds (M3)

Tout d'abord, les rapports observés pour le jeu J1 entre les résultats des initialisations aléatoire (M3A,M3B) et avec FFT(M3C,M3D) sont également valables pour les jeux J3 et J5 (cf. Annexe C, Tableaux C.3 et C.4). En effet, tant pour J3 que pour J5, le rayon maximum est plus élevé pour M3A et B que pour M3C et D. Les critères SDNN et SDTot sont plus faibles pour l'initialisation aléatoire (A et B) que pour l'initialisation avec FFT (C et D). Que ce soit pour J1, J3 ou J5, l'initialisation avec FFT produit donc de meilleurs résultats pour la sélection par diversité que l'initialisation aléatoire.

Ensuite, comme pour les deux précédentes méthodes de sélection les résultats sont similaires d'un jeu à l'autre ou du même ordre de grandeur (cf. Tableau 4.43). on remarque encore une fois que le jeu J5 donne un rayon maximum, un SDNN et un SDTot plus faible que pour le autres jeux, néanmoins les résultats restent comparables. De même, les maximums (SDNN max, SDDep max notamment) diffèrent avec des valeurs plus faibles pour le jeu J3 en général.

Les résultats étant tout de même d'un ordre de grandeur comparable, on considère que la méthode k-medoïds permet une qualité de sélection comparable quelque soit le jeu donné en entrée.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M3C J1	16.6	12.1	1.1	162.9	13.35×10^3	6.1	127.6	6.5
M3C J3	15.6	12.2	1.3	127.5	13.49×10^3	5.7	115.3	6.2
M3C J5	14.5	11.8	1.1	109.4	12.38×10^3	4.4	172.5	7.8

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M3C J1	341.61×10^3	0.0	16.6	1.9	17.73×10^6	6.1	201.7	16.8	326.0
M3C J3	344.53×10^3	0.0	15.6	1.9	17.59×10^6	5.7	241.7	16.5	320.0
M3C J5	340.44×10^3	0.0	14.5	1.9	15.97×10^6	4.4	280.6	20.9	225.0

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.43: Résultats pour la méthode des k-médoïds (M3C), Jeux : J1, J3 et J5, k = 1000

4.5.4 Sélection par Maximum-Dissimilarity (M4D)

Nous avions vu avec le jeu J1, que le type de choix de la molécule suivante avait une influence sur les résultats. En effet, le critère MaxSum (M4A, C et E) produisait des échantillons de moins bonne qualité que le critère MaxMin (M4B, D et F). Nous remarquons que ces observations se vérifient également avec le jeu J3 et le jeu J5, notamment le rayon maximum et SDTot sont plus élevés pour le critère MaxSum, alors que SDNN est plus faible que pour le critère MaxMin (cf. Annexe C, Tableaux C.5 et C.6). Les rapports observés entre les résultats des différents paramétrages sont donc conservés quelque soit le jeu initial.

Ensuite, pour un même paramétrage (ici initialisation au centre, critère MaxMin soit M4D : cf. Tableau 4.44), les valeurs des différents critères sont comparables. Encore une fois le jeu J5 donne des résultats pour le rayon maximum, SDNN, SDDep et SDTot plus faibleque pour les autres jeux. Les maximums diffèrent également mais là aucun jeu ne ressort comme donnant globalement des maximums plus faibles que les autres jeux.

En conclusion, le jeu initial n'a pas d'impact sur la qualité des échantillons obtenus avec la méthode Maximum-Dissimilarity.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M4D J1	13.9	12.9	1.0	585.8	16.17×10^3	13.9	127.6	5.2
M4D J3	14.1	12.9	1.2	553.1	16.26×10^3	14.1	115.3	4.9
M4D J5	13.5	12.6	1.0	471.5	15.54×10^3	13.6	172.5	6.9

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M4D J1	419.57×10^3	0.0	13.9	2.2	18.73×10^6	13.9	201.7	16.0	309.0
M4D J3	427.86×10^3	0.0	14.1	2.1	18.57×10^6	14.1	241.7	15.7	296.0
M4D J5	416.96×10^3	0.0	13.5	2.0	17.25×10^6	13.6	280.6	20.1	212.0

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.44: Résultats pour la méthode Maximum-Dissimilarity (M4D), Jeux : J1, J3 et J5, k = 1000

4.5.5 Sélection par Sphere-Exclusion (M5)

Nous avions vu avec le jeu J1 que cette méthode produisait des résultats très similaires quelques soient les critères de choix de la molécule suivante et quelque soit le type d'initialisation. Il s'avérait cependant que l'initialisation à la molécule centrale du jeu initial tendait à donner des résultats légèrement meilleurs. Ces observations se confirment pour les jeux J3 et J5 avec notamment un rayon maximum et un SDNN meilleur pour l'initialisation au centre que pour les autres initialisations. De plus les paramétrages F, G et H et J, K et L qui donnaient des résultats strictement identiques pour J1, donnent le même type de résultats pour J3 et J5 (cf. Annexe C, Tableaux C.7 et C.8).

Enfin lorsque l'on compare une même paramétrage selon les différents jeux initiaux, on remarque des valeurs similaires pour J1, J3 et J5 (cf. Tableau 4.45).Le jeu J5 donne des valeurs légèrement plus faibles pour le rayon maximum, SDNN, et SDTot. Et encore une fois les maximums diffèrent selon les jeux.

Globalement, on peut conclure que la méthode Sphere-Exclusion est robuste quelque soit le jeu initial utilisé.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M5F J1	14.1	12.9	1.0	559.7	16.10×10^3	14.1	127.6	5.2
M5F J3	14.4	13.0	1.3	622.1	16.05×10^3	14.4	115.3	4.9
M5F J5	13.9	12.7	1.0	509.4	15.08×10^3	13.9	172.5	7.0

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M5F J1	416.47×10^3	0.0	14.1	2.2	19.13×10^6	14.1	201.7	16.0	310.0
M5F J3	431.70×10^3	0.0	14.4	2.1	18.50×10^6	14.4	241.7	15.7	299.0
M5F J5	421.19×10^3	0.0	13.9	2.0	16.54×10^6	13.9	280.6	20.3	213.0

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.45: Résultats pour la méthode Sphere-Exclusion (M5F), Jeux : J1, J3 et J5, k = 1000

4.5.6 Conclusion

Il semble que toutes les méthodes comparées soient robustes quel que soit le jeu de molécules donné en entrée. On remarque toutefois des différences dues aux particularités des jeux eux-mêmes. Notamment, le jeu J5 semble donner des valeurs plus faibles de rayon maximum, de SDNN et de SDTot. Et les maximums, propres aux outliers de chaque jeu, diffèrent donc d'un jeu à l'autre.

4.6 Impact de la taille de l'échantillonnage d'un jeu avec outliers (jeu : J1 - k = 1000, 500, 100)

Pour chaque méthode testée précédemment (cf. section 4.4.6) nous comparons les résultats pour des échantillons de 1000, 500 et 100 molécules parmi les 40 000 du jeu J1. D'une part nous verrons si les rapports entre critères, observés pour un échantillon de 1000 molécules, se retrouvent pour les échantillons de 500 et 100 molécules. D'autre part, nous verrons si la taille de l'échantillon a bien l'impact escompté sur les résultats.

En effet, on s'attend à ce que moins il y a de molécules sélectionnées, plus la taille des groupes est grande. Donc les rayons, les distances entre centres plus proches, les distances entre chaque molécule et son représentant et les distances deux à deux dans l'échantillon devraient augmenter à mesure que la taille de l'échantillon diminue.

Enfin, notons que les sommes telles que SDNN (Somme des Distances entre chaque centre et son centre le plus proche), SDDep (Somme des Distances entre chaque molécule du jeu de départ et son représentant le plus proche), SDTot (Somme des distances deux à deux dans l'échantillon) seront ramenées à des moyennes pour que les résultats soient comparables quelque soit la taille de l'échantillon sélectionné.

4.6.1 Sélection aléatoire (M1)

Rayon et densité des groupes Pour un échantillon de 1000 molécules, nous avions remarqué un rayon maximum très élevé par rapport au rayon moyen signifiant une grande disparité dans la taille des rayons des groupes. Pour les échantillons 500 et 100 (cf. Tableau 4.46) ces tendances se confirment. Le rayon maximum reste très élevé par rapport au rayon moyen. Pour 500, l'écart du nombre d'individu entre les groupes est également faible. En revanche pour 100, cet écart est relativement élevé. Ceci peut s'expliquer par le fait qu'avec un petit échantillon, les centres choisis peuvent être très éloignés les uns des autres et donc induire des groupes de densités très variées.

Enfin les valeurs sont plus élevées pour un échantillon de 100 que pour un échantillon de 500 puis de 1000 comme on peut s'y attendre. En revanche l'échantillon de 500 molécules produit un rayon maximum plus faible que l'échantillon de 1000 molécules contrairement à nos attentes. Cependant nous avons vu que cette méthode ne produisait pas de bons échantillons divers, ce résultat signifie donc simplement qu'une sélection par tirage aléatoire ne donne pas de résultats constants et en rapport avec la taille de l'échantillonnage.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind
M1(1000)	159.2	14.4	9.5	31.3
M1(500)	157.4	19.3	13.9	63.3
M1(100)	174.1	36.9	25.7	353.6

TABLE 4.46: Résultats pour le critère Rayon pour un tirage aléatoire (M1), Jeu 1, k= 1000, 500 et 100

Diversité/Recouvrement de l'espace Comme pour l'échantillon de 1000 molécules, ceux de 500 et 100 molécules produisent des centres très proches les uns des autres (cf. Tableau 4.47). En effet, les deux centres les plus proches ont une très faible distance entre eux (très inférieure au rayon moyen, 500 : 4.7 pour 19.3 de rayon et 100 : 6.1 pour 36.9 de rayon). Et les centres les plus éloignés de leur plus proche ont une distance peu élevée également (61.4 et 31.5 pour 500 et 100 respectivement).

Ensuite, comme attendu la moyenne des distances entre chaque centre et son plus proche augmente à mesure que la taille de l'échantilon diminue. Les résultats respectent donc la proportion existant entre les différentes tailles d'échantillonnage pour la moyenne et pour la distance minimum. En revanche pour la distance maximum entre deux centres proches, l'échantillonnage de 100 molécules semble produire un extrême plus faible que pour 500 et 1000. Ceci est encore une fois dû à la mauvaise qualité de l'échantillonnage par tirage aléatoire.

Enfin lorsqu'on observe la distribution des distances (cf. Figure 4.35), on remarque que quelle que soit la taille de l'échantillon, les centres sont pour la plupart très proches de leur voisin et à une distance inférieure au rayon moyen.

NN [a]	Moyenne	Min	Max	EC
M1(1000)	8.62	3.7	50.0	2.8
M1(500)	9.06	4.5	61.4	3.5
M1(100)	10.358	6.1	31.5	3.2

a. $\text{NN} = \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

TABLE 4.47: Résultats pour le critère SDNN pour un tirage aléatoire (M1), Jeu 1, k= 1000, 500 et 100

Représentativité Quelle que soit la taille de l'échantillon, la tendance selon laquelle la molécule la plus éloignée de son représentant en est à une distance très élevée, est respectée puisque celle-ci correspond au rayon maximum (cf. Tableau 4.48). De plus la moyenne des distances entre chaque molécule et son représentant augmente à mesure que la taille de l'échantillon diminue, les rapports entre critère d'évaluation et taille de l'échantillon sont donc respectés.

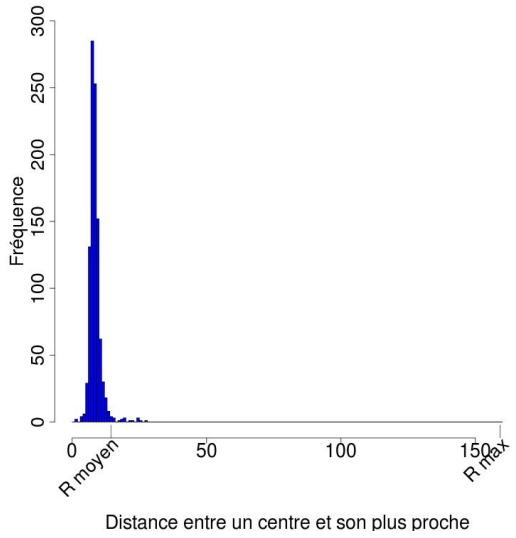
Enfin lorsque l'on étudie la distribution de ces distances (cf. Figure 4.36) on remarque que les tendances observées pour un échantillon de 1000 molécules sont les mêmes pour tout taille déchantillon.

Dep [a]	Moyenne	Min	Max	EC
M1(1000)	8.59	0.0	159.2	3.2
M1(500)	9.2	0.0	157.4	3.3
M1(100)	10.55	0.0	174.1	3.5

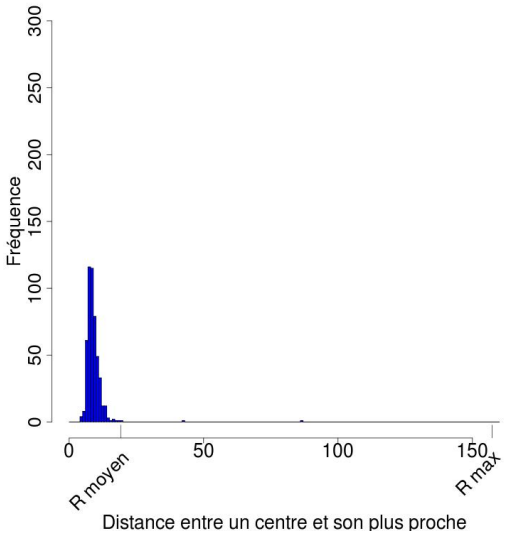
a. $\text{Dep} = \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

TABLE 4.48: Résultats pour le critère SDDep pour un tirage aléatoire (M1), Jeu 1, k= 1000, 500 et 100

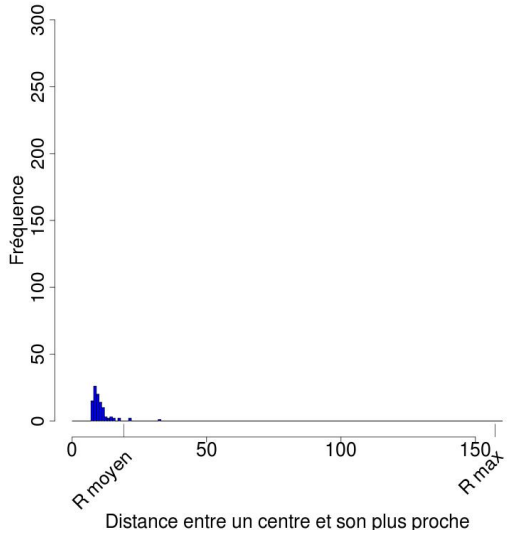
Dissimilarité totale dans l'échantillon Comme pour un échantillonnage de 1000, la distance maximum entre deux centres est faible au regard du rayon maximum que ce



(a) M1 SDNN, 1000

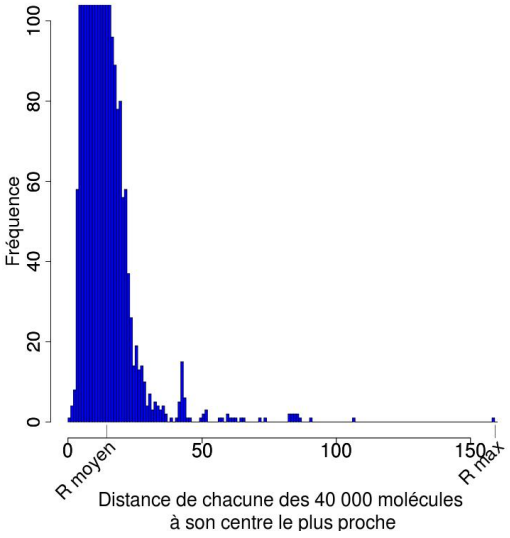


(b) M1 SDNN, 500

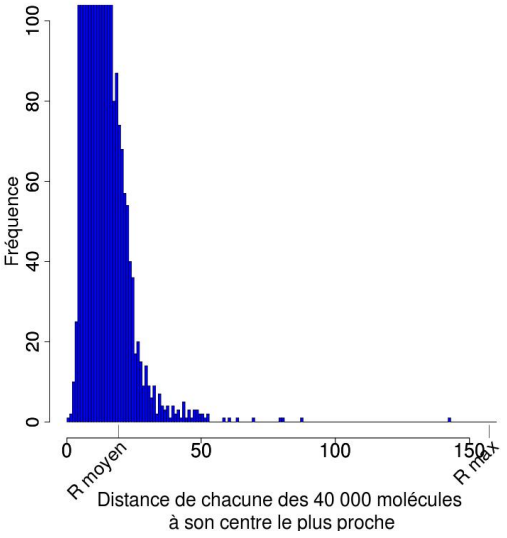


(c) M1 SDNN, 100

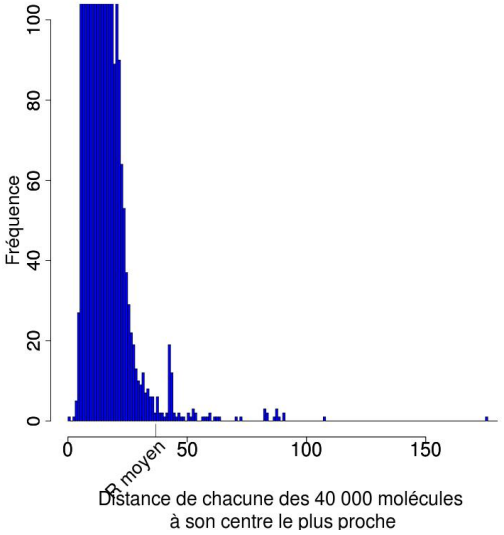
FIGURE 4.35: Pour M1 1000, 500 et 100, Distribution des distances entre chaque centre et son centre le plus proche (SDNN)



(a) M1 SDDep échelle réduite, 1000



(b) M1 SDDep échelle réduite, 500



(c) M1 SDDep échelle réduite, 100

FIGURE 4.36: Pour M1 1000, 500 et 100, Distribution des distances entre chaque molécule et son représentant (SDDep)

soit pour 500 et 100 molécules sélectionnées (cf. Tableau 4.49, exemple 500 : SDT Max= 84.7 pour un rayon maximum de 157.4). Ceci montre toujours une mauvaise dispersion de l'échantillon dans l'espace des descriptions quelle que soit sa taille.

En revanche, étonnament la moyenne des distances deux à deux dans l'échantillon ne varie pas avec la taille de l'échantillon (environ 18.2) alors qu'il devrait augmenter à mesure que la taille de l'échantillon diminue. Nous avons donc un échantillon de 1000 molécules aussi dispersé dans l'espace des descriptions que l'échantillon de 500 et 100. Cette observation confirme nos précédentes affirmations concernant la dispersion des échantillons sélectionnés par tirage aléatoire. En effet ces derniers résultats confirment que le tirage aléatoire produit des sélections peu diverses et très concentrées dans les zones les plus denses de l'espace.

Enfin lorsqu'on étudie la distribution des distances (cf. Figure 4.37), on remarque que les tendances sont conservées d'une taille d'échantillon à l'autre.

Tot [a]	Moyenne	Min	Max	EC
M1(1000)	18.24	3.7	77.9	6.2
M1(500)	18.18	4.5	84.7	6.7
M1(100)	18.28	6.1	54.2	6.3

a. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.49: Résultats pour le critère SDTot pour un tirage aléatoire (M1), Jeu 1, k= 1000, 500 et 100

Singletons Enfin contrairement à nos attentes, la proportion de singletons par rapport à la taille de l'échantillon est plus faible à mesure que la taille de l'échantillon diminue. Ceci s'explique encore une fois par le fait que les échantillons produits par tirage aléatoire sont plutôt concentrés dans les zones denses de l'espace.

	NB Singletons
M1(1000)	3.4
M1(500)	1.0
M1(100)	0.0

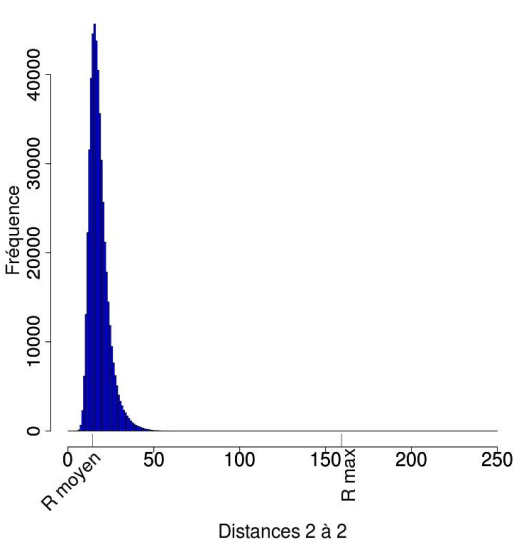
TABLE 4.50: Résultats pour le critère Singletons pour un tirage aléatoire (M1), Jeu 1, k= 1000, 500 et 100

Conclusion Globalement les tendances observées pour un échantillonnage de 1000 molécules se retrouvent pour les échantillonnages de 500 et 100 molécules, notamment par rapport à la dispersion peu diverse de l'échantillon dans l'espace. Cependant on observe quelques exceptions et de plus les proportions attendues sur les résultats en fonction de la taille des échantillons, ne sont pas toujours respectées. Ceci s'explique par le fait que le tirage aléatoire ne donne un échantillonnage ni très régulier ni très divers.

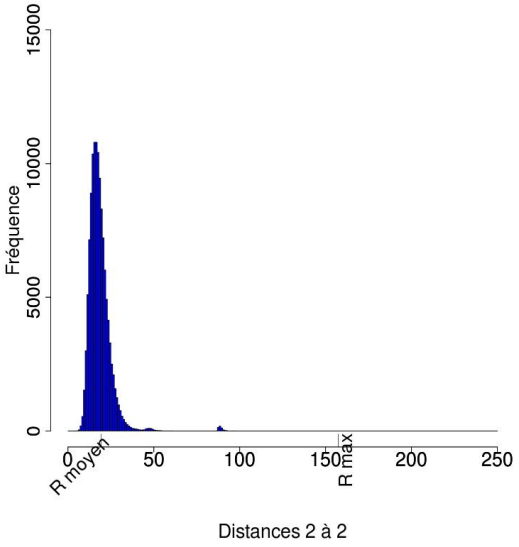
Pour toute taille de l'échantillon, on observe donc que la sélection par tirage aléatoire ne donne pas de jeux divers.

4.6.2 Sélection avec *k*-center (M2A)

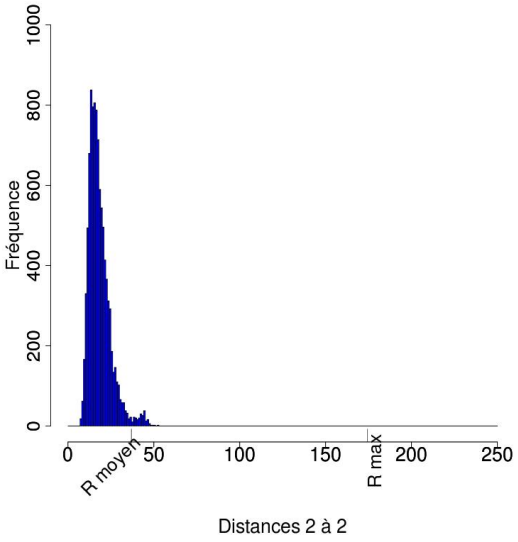
Rayon et densité des groupes Rappelons que pour un échantillonnage de 1000 molécules le rayon maximum est très proche du rayon moyen et que l'écart moyen du nombre



(a) M1 SDTot, 1000



(b) M1 SDTot, 500



(c) M1 SDTot, 100

FIGURE 4.37: Pour M1 1000, 500 et 100, Distribution des distances entre centres deux à deux dans l'échantillon (SDTot). Echelles différentes

d'individus entre les groupes est assez élevé pour indiquer une taille homogène des rayons groupes couvrant des zones plus ou moins denses de l'espace. Pour les échantillons de 500 et 100 molécules, le rayon maximum est également très proche du rayon moyen et l'écart du nombre d'individus entre les groupes est élevé (cf. Tableau 4.51). Enfin les valeurs augmentent à mesure que la taille de l'échantillon diminue.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind
M2A(1000)	12.6	11.8	1.0	177.7
M2A(500)	14.9	13.7	1.3	413.6
M2A(100)	23.6	21.2	2.9	2003.7

TABLE 4.51: Résultats pour le critère Rayon pour k-center (M2A), Jeu 1, k= 1000, 500 et 100

Diversité/Recouvrement de l'espace Comme pour l'échantillon de 1000 molécules, le centre le plus proche de son voisin est à une distance inférieure au rayon maximum et au rayon moyen (cf. Tableau 4.52). La distance maximale entre deux centres voisins est la même quelle que soit la taille de l'échantillon. Ceci signifie qu'au moins un centre situé dans les extrêmes de l'espace est sélectionné à chaque fois. Enfin les proportions des résultats respectent la taille des échantillons.

Lorsqu'on étudie la ditribution des distances (cf. Figure 4.38), on s'aperçoit que plus l'échantillon est petit et plus la majorité des distances entre centres voisins se situe après le rayon moyen. De plus elles sont plus dispersées vers les extrêmes, notamment pour l'échantillon de 100 (cf. Figure 4.38(b)). En effet moins il y a de molécules sélectionnées et plus les distances entre représentants proches sont élevées.

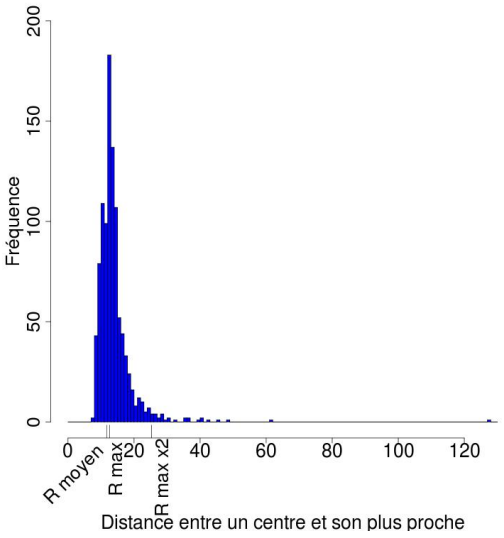
NN [a]	Moyenne	Min	Max	EC
M2A(1000)	14.05	6.9	127.6	5.9
M2A(500)	16.80	8.4	127.6	7.5
M2A(100)	27.77	14.5	127.6	12.8

a. $\text{NN} = \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

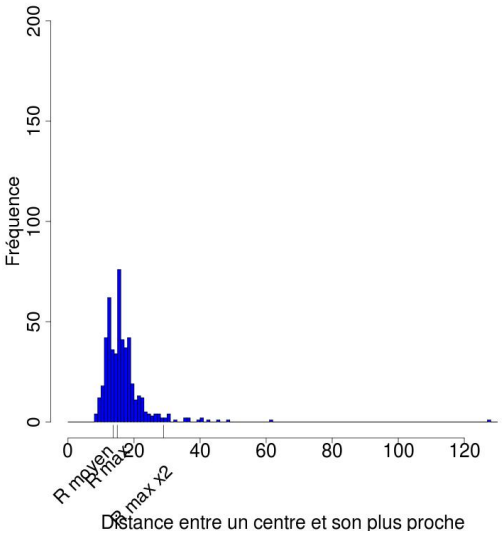
TABLE 4.52: Résultats pour le critère SDNN pour k-center (M2A), Jeu 1, k= 1000, 500 et 100

Représentativité Comme pour l'échantillon de 1000 molécules, les échantillons de 500 et 100 molécules produisent des représentants très proches des molécules du jeu de départ (cf. Tableau 4.53). En effet, toute molécule est plus proche de son représentant que le rayon maximum. La moyenne des distances entre chaque molécule et son représentant augmente à mesure que la taille de l'échantillon diminue. L'étude de la distribution des distances n'apporte pas d'information supplémentaire (histogrammes non montrés).

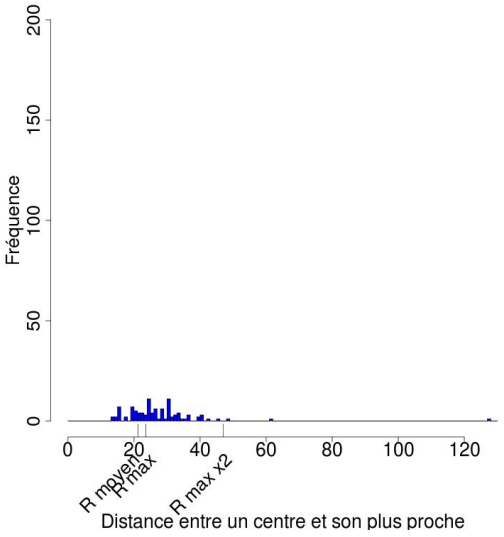
Dissimilarité totale dans l'échantillon Nous observons ici que la distance maximum entre deux centres est la même, confirmant qu'au moins un centre extrême est toujours sélectionné (cf. Tableau 4.54). De plus la moyenne des distances entre centres deux à deux augmente à mesure que la taille de l'échantillon diminue, ce qui indique une bonne conservation des proportions en fonction de la taille de l'échantillon. La distribution des distances n'apporte pas d'information supplémentaire (histogrammes non montrés).



(a) M2A SDNN, 1000



(b) M2A SDNN, 500



(c) M2A SDNN, 100

FIGURE 4.38: Pour M2A 1000, 500 et 100, Distribution des distances entre chaque centre et son centre le plus proche (SDNN)

Dep [a]	Moyenne	Min	Max	EC
M2A(1000)	9.28	0.1	12.6	1.8
M2A(500)	10.3	0.1	14.9	1.8
M2A(100)	14.15	0.1	23.6	2.0

a. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

TABLE 4.53: Résultats pour le critère SDDep pour k-center (M2A), Jeu 1, k= 1000, 500 et 100

Tot [a]	Moyenne	Min	Max	EC
M2A(1000)	37.09	6.9	201.7	16.7
M2A(500)	43.61	8.4	201.7	19.9
M2A(100)	64.71	14.5	201.7	28.4

a. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.54: Résultats pour le critère SDTot pour k-center (M2A), Jeu 1, k= 1000, 500 et 100

Singletons On remarque que plus l'échantillon est petit plus le nombre de singletons diminue (cf. Tableau 4.54). En revanche ramené à la taille de l'échantillon, nous obtenons 42.77 % de singletons pour 1000 molécules sélectionnées, 43.22 % pour 500 molécules et 40.9 % pour 100 molécules sélectionnées. La proportion de singletons reste donc sensiblement la même quelle que soit la taille de l'échantillon. On peut expliquer ce phénomène en observant la distribution des distances entre chaque singleton identifié pour un échantillon de 1000 molécules et sa molécule la plus proche (cf. Figure 4.39). Sur cet histogramme on a indiqué le rayon maximum de l'échantillon de 1000 molécules mais également ceux des échantillons de 500 et 100. Un singleton (c'est vrai pour la plupart d'entre eux) a une distance à sa molécule la plus proche supérieure au rayon maximum de la partition à laquelle il appartient. De ce fait il n'est pas inclus dans un groupe (sinon il augmenterait trop le rayon maximum de la partition) et est alors seul dans son groupe. Or on remarque sur la figure 4.39, que certains singletons, qui avaient une distance à leur molécule plus proche supérieure au rayon maximum de la partition pour 1000 molécules, ont cette même distance inférieure au rayon maximum de la partition pour 500 molécules ou 100 molécules. Donc dans la partition pour 500 molécules, ces singletons ont été inclus dans des groupes, et d'autres qui avaient une distance supérieure avec leur molécule plus proche supérieure à ce rayon sont restés singletons. C'est pourquoi d'une taille d'échantillon à l'autre le nombre de singletons diminue tout en conservant une proportion équivalente. Cette démonstration sera valable pour les autres méthodes également.

Enfin lorsqu'on observe la distribution des distances entre chaque singleton et sa molécule la plus proche (cf. Figure 4.40), les distances les plus extrêmes sont couvertes par les singletons pour toutes les tailles d'échantillon. Les singletons de l'échantillon 100 sont plus nombreux à être à une distance de leur plus proche supérieure au rayon maximum que les singletons des échantillons 500 et 1000. Ils sont donc de meilleurs indicateurs d'outliers.

Conclusion Les observations effectuées pour la méthode k-center sur un échantillon de 1000 sont également valables pour des échantillons de 500 et 100 molécules. De plus les proportions des résultats sont conservées en fonction de la taille des échantillons. Cette méthode est donc robuste au changement de taille de la sélection souhaitée.

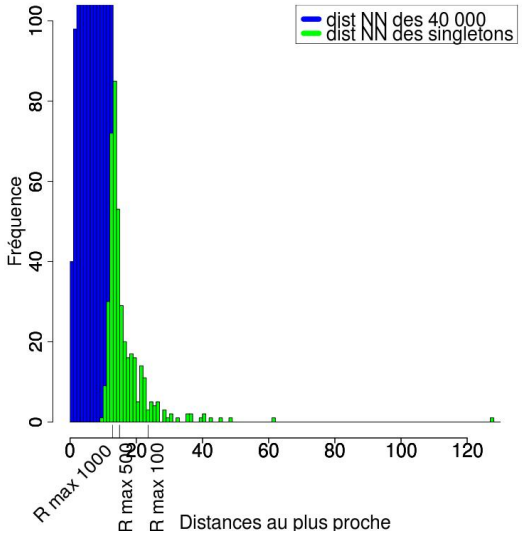


FIGURE 4.39: Pour M2A 1000, Distribution des distances entre chaque molécule et sa molécule la plus proche et projection de ces distances pour les singletons. Projection des rayons max pour 1000, 500 et 100

	NB Singletons
M2A(1000)	427.7
M2A(500)	216.1
M2A(100)	40.9

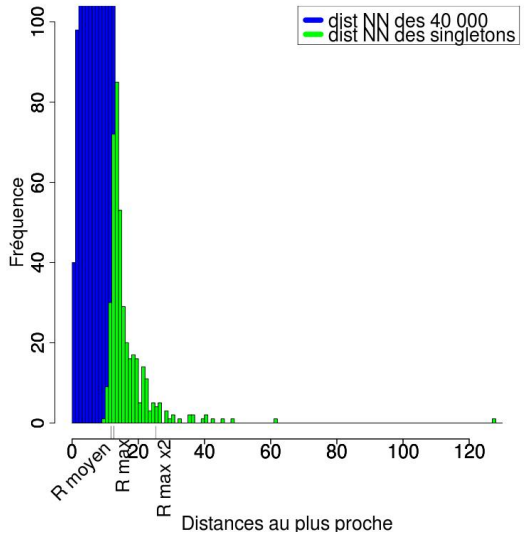
TABLE 4.55: Résultats pour le critère Singleton pour k-center (M2A), Jeu 1, k= 1000, 500 et 100

Enfin en temps de calcul on observe pour M2A (initialisation aléatoire) un temps de sélection de 54 minutes pour 1000 molécules, 42 minutes pour 500 molécules et 42 minutes pour 100 molécules. En revanche pour M2B (initialisation avec FFT), on observe un temps de calcul de 54 minutes pour 1000 molécules, 42 minutes pour 500 molécules et 16 minutes pour 100 molécules. Cela montre bien que la complexité de l'algorithme est dépendante de la taille de l'échantillon final. En revanche elle n'est pas linéaire avec cette taille.

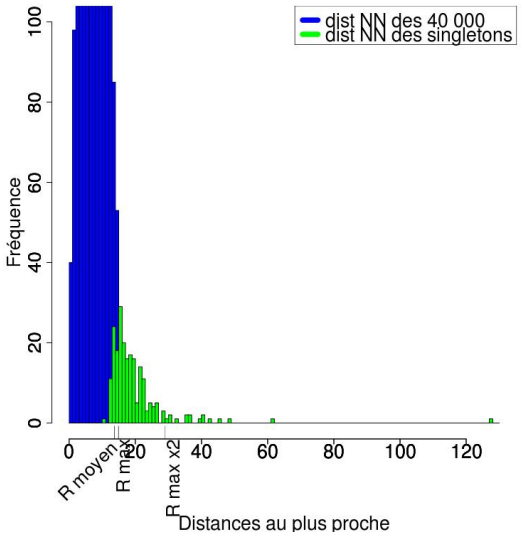
4.6.3 Sélection avec *k*-medoïds (M3)

Nous comparons ici les résultats des échantillons 1000, 500 et 100 obtenus à partir de M3B (initialisation aléatoire, centres virtuels) et de M3C (initialisation avec FFT, centres réels).

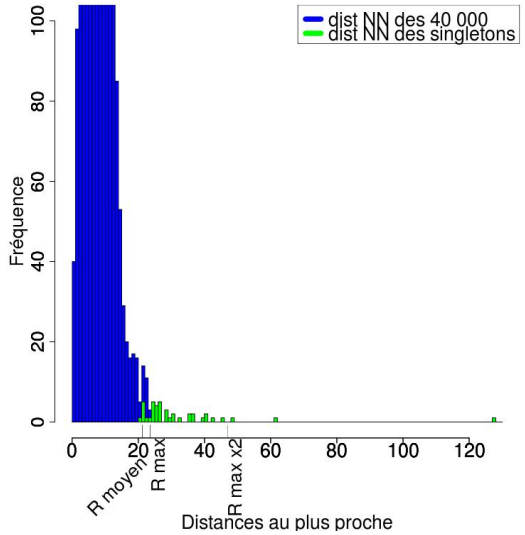
Rayon et densité des groupes Pour chaque type d'initialisation (aléatoire et avec FFT), le rayon maximum et le rayon moyen restent du même ordre de grandeur pour tout échantillon (cf. Tableau 4.56). Leur valeur augmente à mesure que la taille de l'échantillon diminue. On observe donc toujours de meilleurs résultats pour l'initialisation avec FFT (M3C) que pour l'intialisation aléatoire (M3B). De même pour l'écart du nombre d'individus par groupe, quelle que soit la taille de l'échantillon, l'intialisation avec FFT donne de meilleurs résultats.



(a) M2A Singletons échelle réduite, 1000



(b) M2A Singletons échelle réduite, 500



(c) M2A Singletons échelle réduite, 100

FIGURE 4.40: Pour M2A 1000, 500 et 100, Distribution des distances entre chaque molécule et sa molécule la plus proche et projection de ces distances pour les singletons

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind
M3B(1000)	109.1	12.9	6.3	22.9
M3B(500)	124.3	15.8	9.7	44.5
M3B(100)	145.8	28.6	18.4	209.4
M3C(1000)	16.6	12.1	1.1	162.9
M3C(500)	19.1	14.3	1.5	314.6
M3C(100)	32.4	23.0	3.1	1609.7

TABLE 4.56: Résultats pour le critère Rayon pour k-medoïds (M3B et M3C), Jeu 1, k= 1000, 500 et 100

Diversité/Recouvrement de l'espace On remarque pour l'initialisation aléatoire (M3B, cf. Tableau 4.57) que la moyenne des distances entre chaque centre et son plus proche ne varie que très peu entre l'échantillon de 1000 et l'échantillon de 500 molécules, alors qu'elle augmente pour l'échantillon de 100 molécules. Ceci peut s'expliquer comme pour la méthode par tirage aléatoire par un mauvais échantillonnage des 1000 molécules plus éclaté que l'échantillonnage des 500 molécules, ou inversement un mauvais échantillonnage des 500 molécules trop concentré dans une zone de l'espace des descriptions.

Ensuite pour l'intialisation avec FFT, les résultats respectent les proportions des échantillons. Et cette dernière donne de meilleurs résultats que l'initialisation aléatoire avec une moyenne des distances plus élevée et une distance maximum entre deux centres plus élevée également.

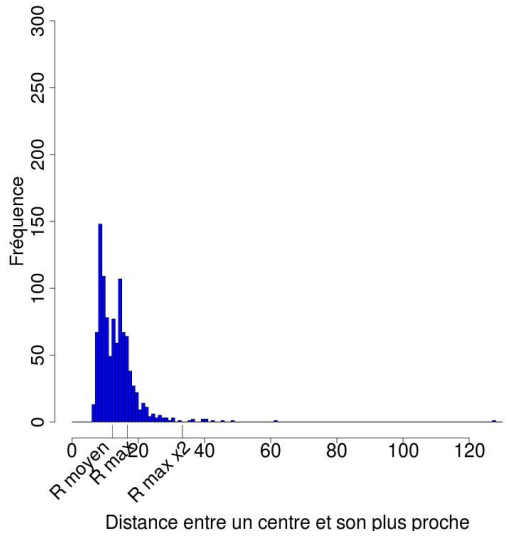
Si l'on observe la distribution des distances, on remarque que toute proportion gardée, les histogrammes des échantillons de 500 et 100 molécules suivent les mêmes tendances que ceux des échantillons de 1000, et ce quelque soit le type d'intialisation (cf. Figure 4.41, M3B non montré).

NN [a]	Moyenne	Min	Max	EC
M3B(1000)	8.277	4.0	106.1	4.7
M3B(500)	8.59	4.2	88.3	5.3
M3B(100)	10.27	5.8	81.3	8.9
M3C(1000)	13.36	6.1	127.6	6.5
M3C(500)	15.85	6.1	127.6	8.3
M3C(100)	25.35	7.4	127.6	15.0

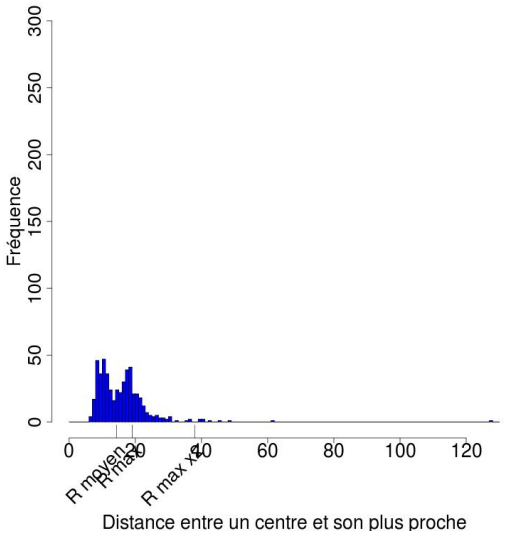
a. $\text{NN} = \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

TABLE 4.57: Résultats pour le critère SDNN pour k-medoïds (M3B et M3C), Jeu 1, k= 1000, 500 et 100

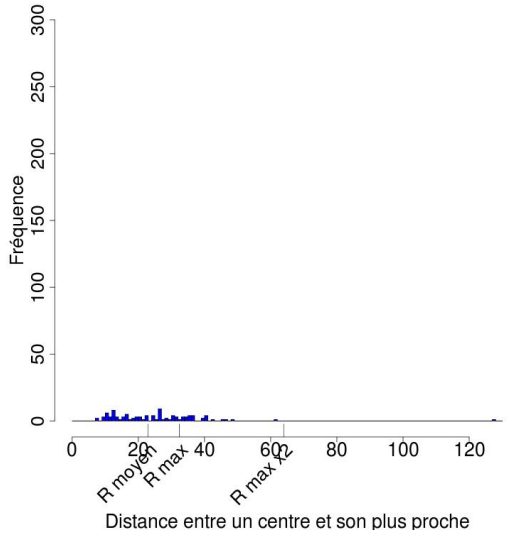
Représentativité Comme pour un échantillon de 1000 molécules, la moyenne des distances entre chaque molécule et son représentant est plus faible pour l'intialisation aléatoire (M3B) que pour l'intialisation avec FFT (cf. Tableau 4.58). Et cette moyenne augmente à mesure que la taille de l'échantillon diminue, respectant ainsi les proportions des échantillons. Sur ce seul critère, quel que soit l'échantillonnage, l'intialisation aléatoire semble donner de meilleurs résultats que l'initialisation avec FFT. Mais comme nous l'avions observé précédemment (cf. section 4.4.3), la distance maximum séparant une molécule de son représentant est bien meilleure pour une initialisation avec FFT et ce quel que soit la taille



(a) M3C SDNN, 1000



(b) M3C SDNN, 500



(c) M3C SDNN, 100

FIGURE 4.41: Pour M3C 1000, 500 et 100, Distribution des distances entre chaque centre et son plus proche (SDNN)

de l'échantillon. La distribution des distances n'apporte pas d'information supplémentaire (non montré) sur la différence de résultats en fonction de la taille de la sélection.

Dep [a]	Moyenne	Min	Max	EC
M3B(1000)	8.09	0.0	109.1	2.7
M3B(500)	8.59	0.0	124.3	2.8
M3B(100)	9.65	0.0	145.8	3.0
M3C(1000)	8.75	0.0	16.6	1.9
M3C(500)	9.39	0.0	19.1	1.9
M3C(100)	10.87	0.0	32.4	2.3

a. $\text{Dep} = \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

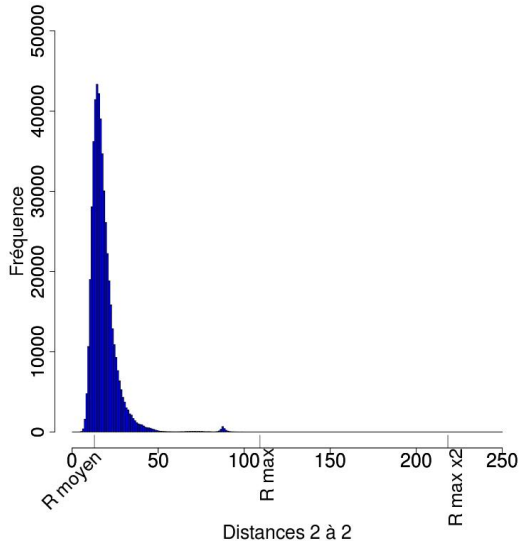
TABLE 4.58: Résultats pour le critère SDDep pour k-medoïds (M3B et M3C), Jeu 1, k= 1000, 500 et 100

Dissimilarité totale dans l'échantillon Encore une fois, pour toute taille d'échantillon, l'initialisation aléatoire (M3B) donne une distance maximale entre deux centres inférieure à celle obtenue avec l'initialisation avec FFT (M3C). Comme nous pouvons le constater dans le tableau 4.59, cette distance maximale reste la même pour une initialisation avec FFT et pour toute taille de la sélection. Ceci indique qu'au moins un centre extrême est toujours sélectionné.

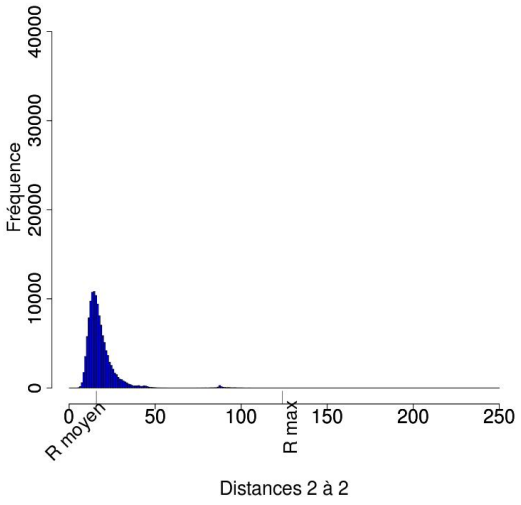
Ensuite, pour une initialisation aléatoire on retrouve le même type d'anomalie que pour le critère SDNN, à savoir que l'échantillon de 500 molécules ne produit pas de moyenne des distances deux à deux plus grande que celle produite par l'échantillon de 1000 molécules. Encore une fois cela indique que la méthode k-medoïds ne permet pas une régularité des résultats pour toute taille d'échantillon.

En effet lorsque l'on observe la distribution des distances, l'échantillon de 100 molécules (cf. Figure 4.42(c)), proportionnellement au nombre de distances total, produit plus de distances extrêmes (proches de 100 et proches de 50) que les autres échantillons. Ceci s'explique par le fait qu'il existe un certain nombre de points très éloignés dans l'espace qui sont toujours sélectionnés pour 100 molécules, il reste donc moins de possibilités de sélection de molécules non extrêmes que pour les autres échantillons. La moyenne des distances deux à deux pour l'échantillon de 100 est donc logiquement plus élevée. En revanche on remarque que la ditribution des distances pour l'échantillon de 500 (cf. Figure 4.42(b)) est la même que celle pour l'échantillon de 1000 molécules (cf. Figure 4.42(a)), toute proportion gardée. Ceci signifie certainement que l'échantillon de 1000 molécules est plus concentré dans une zone de l'espace englobant ainsi moins de points extrêmes, sa moyenne des distances deux à deux se rapproche alors de celle de l'échantillon de 500 molécules.

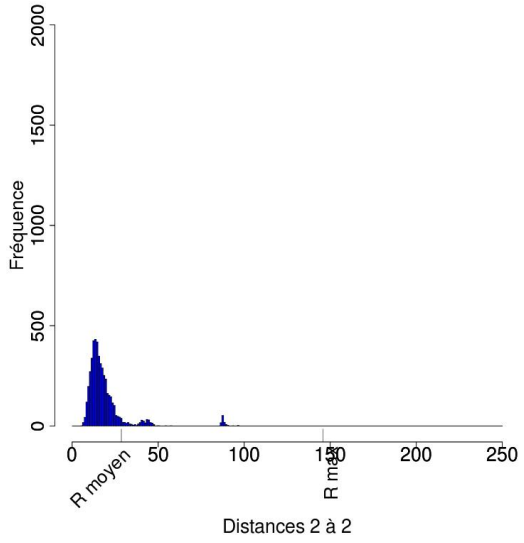
Singletons Tout d'abord on remarque que pour l'intialisation aléatoire (M3B, cf. Tableau 4.60), le nombre de singletons qui était déjà très faible (et rappelons-le ceux-ci n'étaient pas de bons indicateurs d'outliers) devient nul à mesure que la taille de l'échantillon diminue. Pour l'initialisation avec FFT (M3C), si l'on ramène le nombre de singletons à sa proportion par rapport à la taille de l'échantillon, on obtient : 32.6% de singletons pour l'échantillon de 1000 molécules, 34.2 % de singletons pour l'échantillons de 500 molécules et 31 % de singletons pour l'échantillons de 100 molécules. La proportion de singletons reste donc sensiblement la même quelque soit la taille de l'échantillon.



(a) M3B SDTot, 1000



(b) M3B SDTot, 500



(c) M3B SDTot, 100

FIGURE 4.42: Pour M3B 1000, 500 et 100, Distribution des distances deux à deux dans l'échantillon. Echelles différentes

Tot [a]	Moyenne	Min	Max	EC
M3B(1000)	18.48	4.0	148.5	9.1
M3B(500)	18.34	4.2	124.4	9.3
M3B(100)	19.45	5.8	102.6	12.8
M3C(1000)	35.5	6.1	201.7	16.8
M3C(500)	41.81	6.1	201.7	20.4
M3C(100)	61.42	7.4	201.7	30.1

a. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.59: Résultats pour le critère SDTot pour k-medoïds (M3B et M3C), Jeu 1, k= 1000, 500 et 100

Enfin en observant la distribution des distances entre chaque singleton et sa molécule la plus proche (cf. Figure 4.43), on remarque que, comme pour la méthode k-center, les singletons de l'échantillon 100 sont en majorité de meilleurs indicateurs d'outliers que ceux des échantillons 500 et 1000. En effet pour ces derniers, beaucoup de leurs singletons se situent avant le rayon maximum voire même avant le rayon moyen.

	NB Singletons
M3B(1000)	3.6
M3B(500)	0.9
M3B(100)	0.0
M3C(1000)	326.0
M3C(500)	171.0
M3C(100)	31.0

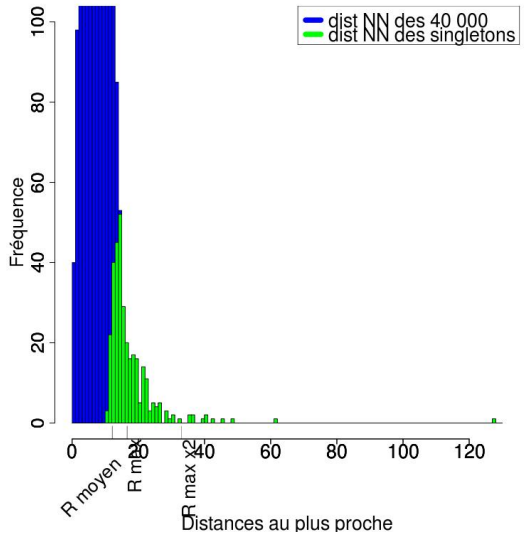
TABLE 4.60: Résultats pour le critère Singletons pour k-medoïds (M3B et M3C), Jeu 1, k= 1000, 500 et 100

Conclusion Pour l'initialisation avec FFT (M3C) les résultats suivent les mêmes tendances et gardent les proportions quelle que soit la taille de l'échantillon. Pour l'intialisation aléatoire (M3B), l'échantillonnage n'est pas toujours régulier, puisqu'on a remarqué que les résultats de l'échantillon de 500 ne gardaient pas toujours les proportions des résultats. Donc pour la méthode k-medoïds, il est préférable d'utiliser l'initialisation avec FFT pour obtenir des résultats robustes pour toute taille sélection.

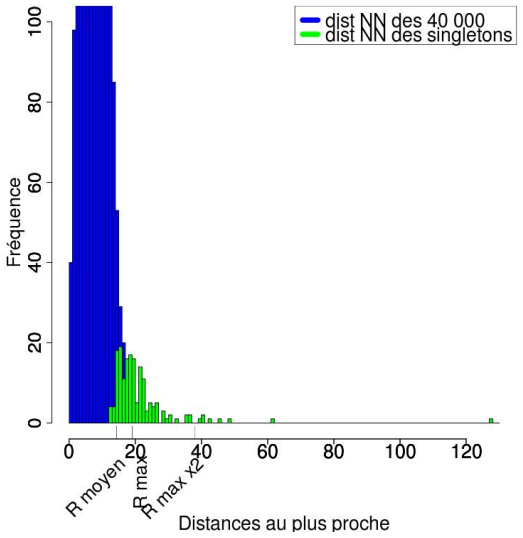
Le temps de calcul est de 3 minutes 30 pour une sélection de 1000 molécules, 2 minutes pour 500 molécules et 30 secondes pour 100 molécules. Contrairement à la méthode k-center, on remarque ici la dépendance linéaire de la complexité de l'algorithme avec la taille de l'échantillon.

4.6.4 Sélection avec Maximum-Dissimilarity (M4)

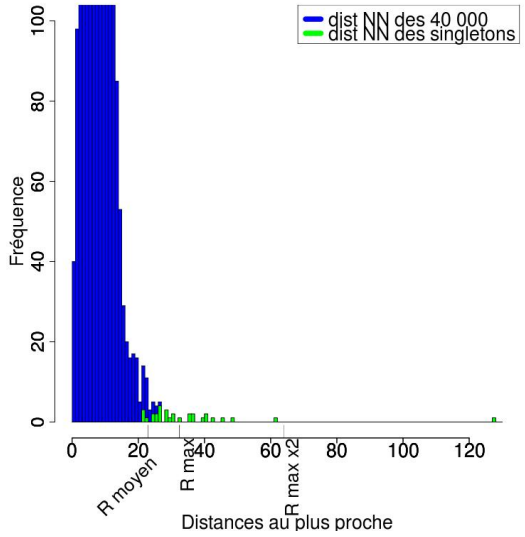
Rayon et densité des groupes Comme pour l'échantillon de 1000 molécules, les échantillons de 500 et 100 molécules produisent des rayons maximums proches des rayons moyens (cf. Tableau 4.61). De plus l'écart du nombre d'individus par groupe est toujours élevé. Enfin les valeurs augmentent à mesure que la taille de la sélection diminue, les proportions sont donc respectées.



(a) M3C Singletons échelle réduite, 1000



(b) M3C Singletons échelle réduite, 500



(c) M3C Singletons échelle réduite, 100

FIGURE 4.43: Pour M3C 1000, 500 et 100, Distribution des distances entre chaque molécule et sa molécule la plus proche et projection de ces distances pour les singletons

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind
M4D(1000)	13.9	12.9	1.0	585.8
M4D(500)	16.7	15.2	1.3	1309.1
M4D(100)	27.5	23.7	2.7	4583.8

TABLE 4.61: Résultats pour le critère Rayon pour Maximum-Dissimilarity (M4), Jeu 1, k= 1000, 500 et 100

Diversité/Recouvrement de l'espace Comme précédemment, les proportions sont respectées en fonction de la taille des échantillons. Les centres sont toujours à une distance inférieure ou égale au rayon moyen de leur centre le plus proche (cf. Tableau 4.62). La distance maximum entre deux centres voisins est toujours la même quelle que soit la taille de l'échantillon, indiquant qu'au moins un centre extrême est sélectionné à chaque fois. La distribution des distances n'apporte pas d'information supplémentaire (non montré ici).

NN [a]	Moyenne	Min	Max	EC
M4D(1000)	16.178	13.9	127.6	5.2
M4D(500)	19.59	16.7	127.6	6.5
M4D(100)	32.79	27.6	127.6	10.7

a. $\text{NN} = \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

TABLE 4.62: Résultats pour le critère SDNN pour Maximum-Dissimilarity (M4), Jeu 1, k= 1000, 500 et 100

Représentativité De même que pour les autres critères, la moyenne des distances entre chaque molécule et son représentant respecte les proportions des différents échantillons (cf. Tableau 4.63).

Dep [a]	Moyenne	Min	Max	EC
M4D(1000)	10.76	0.0	13.9	2.2
M4D(500)	11.99	0.0	16.7	2.4
M4D(100)	13.51	0.0	27.5	3.6

a. $\text{Dep} = \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

TABLE 4.63: Résultats pour le critère SDDep pour Maximum-Dissimilarity (M4), Jeu 1, k= 1000, 500 et 100

Dissimilarité totale dans l'échantillon Comme nous l'avons vu pour le critère SDNN, la distance maximum entre deux centres est la même pour toutes les tailles d'échantillon (cf. Tableau 4.64), indiquant encore une fois que les extrêmes sont toujours sélectionnés. Les valeurs des autres critères augmentent à mesure que la taille de l'échantillon diminue.

Singletons Le nombre de singletons diminue avec la taille de la sélection (cf. Tableau 4.65). Or si on rapporte ce nombre à la taille des échantillons, on trouve : 30.9 % de singletons dans la sélection de 1000 molécules, 31 % de singletons dans la sélection de 500 molécules et 34 % de singletons dans l'échantillon de 100 molécules. Lorsqu'on étudie la

Tot [a]	Moyenne	Min	Max	EC
M4D(1000)	37.51	13.9	201.7	16.0
M4D(500)	44.71	16.7	201.7	19.1
M4D(100)	67.56	27.6	201.7	26.8

a. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.64: Résultats pour le critère SDTot pour Maximum-Dissimilarity (M4), Jeu 1, k= 1000, 500 et 100

distribution des distances entre chaque singleton et sa molécule la plus proche (cf. Figure 4.44), on remarque comme pour les autres méthodes que les singletons de l'échantillon de 100 molécules sont de meilleurs indicateurs d'outliers que ceux des autres échantillons.

	NB Singletons
M4D(1000)	309.0
M4D(500)	155.0
M4D(100)	34.0

TABLE 4.65: Résultats pour le critère Singletons pour Maximum-Dissimilarity (M4), Jeu 1, k= 1000, 500 et 100

Conclusion Quelle que soit la taille de l'échantillon, les résultats suivent les mêmes tendances pour la méthode Maximum-Dissimilarity et respectent les proportions. Cette méthode est donc robuste aux différentes tailles de sélection.

Enfin les temps de calcul pour la sélection de molécules sont dépendants de la taille de l'échantillon. En effet ils diminuent avec la taille de la sélection soit : 30 secondes pour sélectionner 1000 molécules, 20 secondes pour 500 molécules et 10 secondes pour 100 molécules.

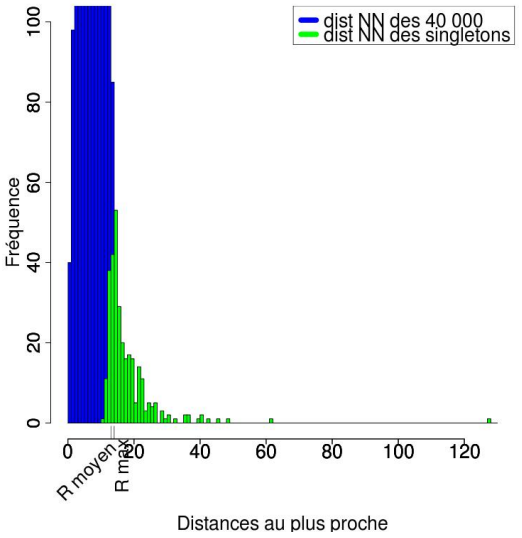
4.6.5 Sélection avec Sphere-Exclusion (M5)

Pour cette méthode nous obtenons exactement les mêmes résultats que pour la méthode Maximum-Dissimilarity (cf. Annexe D). Nous ne détaillerons donc pas les résultats ici.

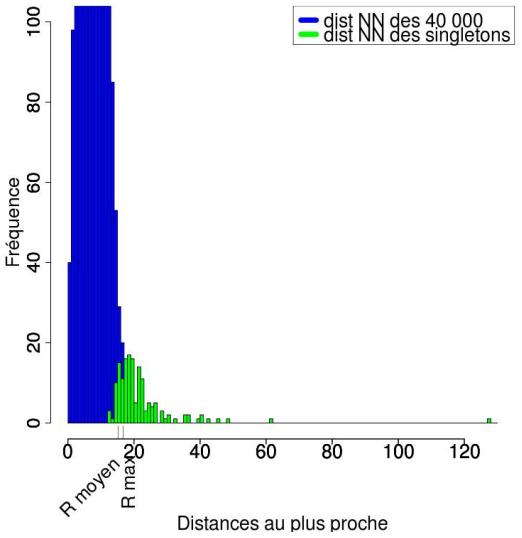
4.6.6 Conclusion sur l'impact de la taille de l'échantillon

Nous avons pu voir que la taille de l'échantillon n'a aucune incidence sur les tendances que suivent les résultats pour les méthodes k-center, Maximum-Dissimilarity et Sphere-Exclusion. De plus les résultats obtenus avec ces méthodes respectent les proportions des échantillons. Pour la méthode k-medoïds, seule l'initialisation avec FFT permet d'obtenir la même robustesse des résultats.

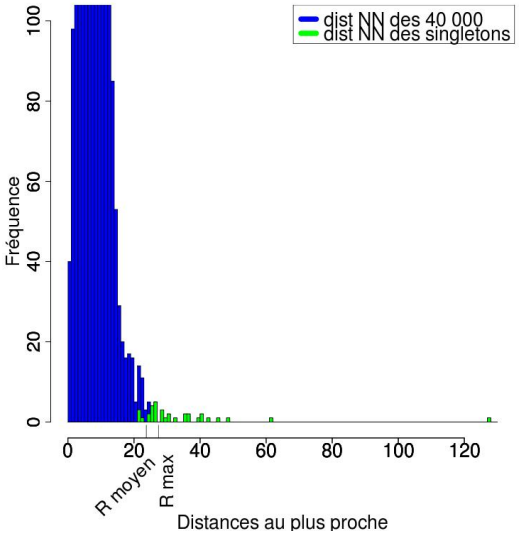
Enfin la méthode k-medoïds avec une initialisation aléatoire et la méthode de sélection par tirage aléatoire ne permettent pas d'obtenir des échantillons respectant les proportions attendues. En effet, les conclusions obtenues pour un échantillon de 1000 molécules dans la section précédente (cf. section 4.4) ne sont pas forcément valables pour d'autres tailles d'échantillon. Comme nous avons pu le voir, ces deux méthodes ne permettent donc pas d'obtenir des résultats reproductibles quelle que soit la taille de la sélection.



(a) M4D Singletons échelle réduite, 1000



(b) M4D Singletons échelle réduite, 500



(c) M4D Singletons échelle réduite, 100

FIGURE 4.44: Pour M4D 1000, 500 et 100, Distribution des distances entre chaque molécule et sa molécule la plus proche et projection de ces distances pour les singletons

4.7 Impact de l'absence d'outliers dans le jeu initial sur l'échantillonnage (jeux : J1 versus J2 - k = 1000)

Nous souhaitons étudier l'impact des outliers sur la sélection par diversité. L'identification des outliers est difficile et n'a pas de solution absolue. Or comme nous l'avons suggéré précédemment (cf. section 4.4), les singletons obtenus par certaines méthodes peuvent être de bons indicateurs d'outliers. Nous avons estimé que supprimer quelques molécules non outliers au jeu de départ n'aurait pas beaucoup d'impact. Par précaution nous avons donc retiré tous les singletons de toutes les méthodes pour obtenir le jeu J2 à partir de J1, J4 à partir de J3 et J6 à partir de J5. Nous comparons donc ici les résultats d'une sélection par diversité sur un jeu avec outliers J1 et sur un jeu sans outlier J2.

On s'attend à ce que les résultats soient plus faibles pour le jeu sans outlier que pour le jeu avec outliers car le premier est plus concentré dans l'espace que le deuxième.

4.7.1 Sélection par tirage aléatoire M1

Lorsque l'on étudie le rayon maximum (cf. Tableau 4.66), on remarque qu'il est beaucoup plus faible pour le jeu sans outlier (J2). Ceci s'explique par le fait que ce rayon maximum était régi par un outlier pour le jeu J1. En revanche le rayon moyen reste faible quelque soit le jeu, ainsi que l'écart du nombre d'individus par groupe. De plus les critères SDNN, SDDep et SDTot ont sensiblement les mêmes valeurs. Ceci signifie que les échantillons produits à partir de J1 et ceux produits à partir de J2 sont aussi divers l'un que l'autre. A quelques outliers prêts, il semble que la sélection par tirage aléatoire produise bien (comme vu précédemment, cf. section 4.4) un échantillon peu divers et concentré dans les zones les plus denses de l'espace.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M1 J1	159.25	14.41	9.58	31.31	8.61×10^3	3.71	50.00	2.86
M1 J2	37.61	12.91	3.67	30.95	8.69×10^3	3.25	34.08	2.09

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M1 J1	335.34×10^3	0.01	159.25	3.23	9.10×10^6	3.71	77.95	6.21	3.40
M1 J2	331.87×10^3	0.01	37.61	2.46	9.16×10^6	3.25	56.89	5.44	2.20

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.66: Résultats pour le tirage aléatoire (M1), Jeu 1 et Jeu 2

4.7.2 Sélection par *k*-center (M2)

Les rayons maximum et moyen sont sensiblement les mêmes pour le jeu avec outliers et le jeu sans outlier. En revanche comme attendu, l'écart du nombre d'individus par groupe, SDNN, SDNN Max, SDDep et SDTot sont plus faibles pour l'échantillon issu du jeu sans outlier (J2) que pour le jeu avec outliers (J1). Ceci montre qu'au delà de la particularité qu'induit un jeu sans outlier (concentration dans une zone de l'espace induisant des distances plus faibles), les résultats de la méthode k-center ne sont pas affectés par les outliers. C'est à dire que malgré des valeurs plus faibles, les tendances sont conservées.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M2A J1	12.68	11.81	1.02	177.76	14.05×10^3	6.97	127.69	5.96
M2A J2	12.03	11.28	0.87	98.79	11.34×10^3	6.46	22.02	1.82

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M2A J1	362.14×10^3	0.10	12.68	1.89	18.52×10^6	6.97	201.77	16.72	427.70
M2A J2	344.42×10^3	0.10	12.03	1.85	13.69×10^6	6.46	65.66	8.18	163.20

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.67: Résultats pour la méthode du k-center (M2A), Jeu 1 et Jeu 2

4.7.3 Sélection par *k*-medoïds (M3)

Comme pour la méthode k-center, la méthode k-medoïds produit un échantillon à partir du jeu sans outlier qui donne des résultats plus faibles que l'échantillon issu du jeu avec outliers (cf. Tableau 4.68). Cependant les rapports observés entre les critères sont conservés. Cette méthode produit donc le même genre de résultat quelque soit le type du jeu de molécules de départ.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M3C J1	16.67	12.19	1.18	162.93	13.35×10^3	6.15	127.69	6.52
M3C J2	14.57	11.62	0.93	99.48	10.49×10^3	3.95	22.56	2.54

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M3C J1	341.61×10^3	0.01	16.67	1.93	17.73×10^6	6.15	201.77	16.89	326.00
M3C J2	331.30×10^3	0.01	14.57	1.90	13.06×10^6	3.95	65.85	8.41	78.00

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.68: Résultats pour la méthode des k-medoïds (M3C), Jeu 1 et Jeu 2

4.7.4 Sélection par Maximum-Dissimilarity (M4)

Comme pour la méthode k-center, la méthode Maximum-Dissimilarity produit des échantillons issus de jeu avec ou sans outlier qui ont des rayons maximum et moyen similaires (cf. Tableau 4.69). Pour les autres critères, l'échantillon issu du jeu sans outlier donne des résultats plus faibles que l'échantillon issu du jeu avec outliers. Malgré ces valeurs plus faibles, les rapports entre critères sont respectés quelque soit le type de jeu de départ.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M4D J1	13.99	12.97	1.03	585.84	16.17	13.99	127.69	5.22
M4D J2	13.22	12.29	0.87	368.56	13.96	13.22	23.64	0.77

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M4D J1	419.57×10^3	0.01	13.99	2.20	18.73×10^6	13.99	201.77	16.04	309.00
M4D J2	402.55×10^3	0.01	13.22	2.07	14.45×10^6	13.22	65.85	7.80	49.00

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.69: Résultats pour la méthode Maximum-Dissimilarity (M4D), Jeu 1 et Jeu 2

4.7.5 Sélection par Sphere-Exclusion (M5)

Enfin les observations précédentes pour la méthode Maximum-Dissimilarity sont également valables pour la méthode Sphere-Exclusion (cf. Tableau 4.70). En effet les valeurs pour l'échantillon issu du jeu sans outlier sont inférieures à celles de l'échantillon issu du jeu avec outliers tout en conservant les tendances.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M5F J1	14.15	12.91	1.08	559.71	16.10×10^3	14.15	127.69	5.20
M5F J2	13.35	12.28	0.97	367.56	13.93×10^3	13.36	21.07	0.63

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M5F J1	416.47×10^3	0.01	14.15	2.20	19.13×10^6	14.15	201.77	16.00	310.00
M5F J2	401.86×10^3	0.01	13.35	2.07	14.80×10^6	13.36	65.85	7.91	56.00

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.70: Résultats pour la méthode Sphere-Exclusion (M5F), Jeu 1 et Jeu 2

4.7.6 Conclusion

Seule la méthode de sélection par tirage aléatoire ne produit pas les résultats attendus lors d'une sélection sur un jeu sans outlier. Ceci confirme nos observations précédentes selon lesquelles, ce tirage a tendance à sélectionner des molécules dans les zones les plus denses de l'espace. Il est donc logique que la sélection sur le jeu sans outlier donne des résultats similaires à la sélection sur le jeu avec outliers.

Ensuite toutes les autres méthodes ne sont pas affectées par le type de jeu de départ, en effet qu'il y ait ou non des outliers, les tendances précédemment observées sont respectées.

Enfin nous remarquons que même les jeux sans outlier produisent des singletons dans une moindre mesure.

4.8 Impact du jeu initial sans outlier sur l'échantillonnage (Jeux : J2 versus J4 versus J6 - k = 1000)

Comme pour les jeux avec outliers, nous avons créé 3 jeux sans outlier afin de tester la robustesse des méthodes sur différents jeux de données en entrée.

Comme nous l'avons dit précédemment, seuls certains paramétrages pour chaque méthode donnent les meilleurs résultats. Ainsi, seuls les paramétrages suivants seront mis en avant pour comparer J2 à J4 et J6 :

- M1, tirage aléatoire uniforme sans remise.
- M2A, notre implémentation du k-center avec une initialisation aléatoire
- M3C, k-medoïds avec une initialisation avec FFT et des itérations avec les centres réels
- M4D, Maximum-Dissimilarity ou FFT avec initialisation à la molécule centrale
- M5F, Sphere-Exclusion avec une initialisation à la molécule centrale et le critère MaxSum (résultats égaux aux critères MinMin et MinMax)

Toutefois, les tableaux complets de chaque méthode pour les jeux J4 et J6 sont disponibles en annexe et seront commentés dans cette section.

4.8.1 Sélection par tirage aléatoire (M1)

Quelque soit le jeu sans outlier utilisé en entrée, les valeurs des critères sont très similaires (cf. Tableau 4.71). Seules les valeurs des maxima (Rayon maximum, SDNN max, SDDep Max et SDTot max) diffèrent plus fortement entre les différents jeux que les autres critères. Cependant, ces écarts restent raisonnables. La méthode de sélection par tirage aléatoire n'est donc pas influencée par le jeu initial lorsque celui-ci ne comporte pas d'outlier.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M1 J2	37.6	12.9	3.6	30.9	8.69×10^3	3.2	34.0	2.0
M1 J4	49.1	12.9	3.7	30.5	8.64×10^3	3.8	38.3	2.2
M1 J6	49.3	12.8	4.4	30.4	8.74×10^3	3.6	26.9	1.8

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M1 J2	331.87×10^3	0.0	37.6	2.4	9.16×10^6	3.2	56.8	5.4	2.2
M1 J4	332.35×10^3	0.0	49.1	2.5	9.11×10^6	3.8	60.0	5.3	2.0
M1 J6	334.50×10^3	0.0	49.3	2.3	9.14×10^6	3.6	56.3	5.0	2.1

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.71: Résultats pour le tirage aléatoire (M1), Jeux : J2, J4 et J6, k = 1000

4.8.2 Sélection par *k*-center (M2)

Comme nous l'avons observés dans la section 4.4, les résultats sont similaires quelle que soit l'initialisation de la méthode (cf. Annexe E, Tableau E.1). Ensuite pour les trois jeux sans outlier, les critères ont des valeurs similaires. Le jeu J6 issu du jeu J5 donne des résultats légèrement plus faibles pour SDNN et SDTot (cf. Tableau 4.72). La méthode k-center n'est donc pas impactée par le jeu initial lorsqu'il n'y a pas d'outlier.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M2A J2	12.0	11.2	0.8	98.7	11.34×10^3	6.4	22.0	1.8
M2A J4	12.2	11.4	0.9	106.1	11.61×10^3	6.7	25.6	1.9
M2A J6	12.0	11.2	0.8	95.6	11.20×10^3	6.6	17.6	1.7

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M2A J2	344.42×10^3	0.1	12.0	1.8	13.69×10^6	6.4	65.6	8.1	163.2
M2A J4	347.75×10^3	0.1	12.2	1.8	13.85×10^6	6.7	84.1	8.2	190.4
M2A J6	346.69×10^3	0.1	12.0	1.8	12.71×10^6	6.6	72.2	8.0	140.5

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.72: Résultats pour le tirage aléatoire (M1), Jeux : J2, J4 et J6, k = 1000

4.8.3 Sélection par *k*-medoïds (M3)

Les rapports observés en section 4.4 entre les valeurs obtenues pour les différents paramétrages de la méthode sont respectés également pour les trois jeux sans outlier (cf. Annexe E, Tableau E.2). Ensuite, les trois jeux donnent des valeurs similaires pour les critères (cf. Tableau 4.73). Seuls les valeurs maximales diffèrent selon les jeux. Cette méthode donne donc des échantillons de qualité équivalente quelque soit le jeu sans outlier proposé en entrée.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M3C J2	14.5	11.6	0.9	99.4	10.49×10^3	3.9	22.5	2.5
M3C J4	14.7	11.7	1.1	78.1	10.70×10^3	5.3	25.3	2.6
M3C J6	14.7	11.5	1.0	80.3	10.25×10^3	3.9	17.8	2.2

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M3C J2	331.30×10^3	0.0	14.5	1.9	13.06×10^6	3.9	65.8	8.4	78.0
M3C J4	330.89×10^3	0.0	14.7	1.9	13.17×10^6	5.3	84.1	8.5	88.0
M3C J6	333.25×10^3	0.0	14.7	1.9	11.76×10^6	3.9	72.2	7.2	55.0

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.73: Résultats pour la méthode des k-médoïds (M3C), Jeux : J2, J4 et J6, k = 1000

4.8.4 Sélection par Maximum-Dissimilarity (M4D)

Comme vu précédemment le critère de choix de la molécule suivante MaxMin donne des échantillons divers de meilleure qualité que le critère MaxSum (cf. Annexe E, Tableau E.3). De plus les valeurs sont similaires quelque soit le jeu proposé en entrée (cf. Tableau 4.74). Cette méthode n'est donc pas influencée par le jeu initial sans outlier.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M4D J2	13.2	12.2	0.8	368.5	13.96×10^3	13.2	23.6	0.7
M4D J4	13.3	12.3	0.9	325.4	14.15×10^3	13.3	25.3	0.8
M4D J6	13.1	12.3	0.8	346.4	13.81×10^3	13.1	18.4	0.6

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M4D J2	402.55×10^3	0.0	13.2	2.0	14.45×10^6	13.2	65.8	7.8	49.0
M4D J4	409.44×10^3	0.0	13.3	2.0	14.51×10^6	13.3	84.1	7.8	60.0
M4D J6	404.44×10^3	0.0	13.1	2.0	13.35×10^6	13.1	72.2	6.8	37.0

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.74: Résultats pour la méthode Maximum-Dissimilarity (M4D), Jeux : J2, J4 et J6, k = 1000

4.8.5 Sélection par Sphere-Exclusion (M5)

Les conclusions obtenues avec les jeux comportant des outliers (cf. section 4.4) sont également valables pour les jeux sans outlier. Et notamment, les paramétrages F, G et H et J, K et L sont identiques entre eux car donnant des échantillons identiques. Cependant, l'intialisation au centre ne donne pas de meilleurs résultats que les autres initialisations comme c'était le cas pour les jeux avec outliers (cf. Annexe E, Tableau E.4). En ce qui concerne la comparaison des trois jeux sans outlier pour un même paramétrage (soit M5F : initialisation au centre, choix de la molécule suivante avec le critère MaxSum), les résultats sont similaires quel que soit le jeu de départ (cf. Tableau 4.75). Le jeu J6 donne des valeurs plus faibles que les autres pour le critère rayon maximum, SDNN et SDTot.

Globalement, le jeu initial sans outlier n'a pas d'impact sur les échantillons produits par la méthode Sphere-Exclusion.

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M5F J2	13.3	12.2	0.9	367.5	13.93×10^3	13.3	21.0	0.6
M5F J4	13.5	12.4	1.0	366.8	14.00×10^3	13.5	25.9	0.7
M5F J6	13.2	12.2	0.9	337.7	13.72×10^3	13.2	19.5	0.5

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M5F J2	401.86×10^3	0.0	13.3	2.0	14.80×10^6	13.3	65.8	7.9	56.0
M5F J4	410.46×10^3	0.0	13.5	2.0	14.74×10^6	13.5	84.1	7.9	68.0
M5F J6	403.16×10^3	0.0	13.2	2.0	13.68×10^6	13.2	72.2	7.2	49.0

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE 4.75: Résultats pour la méthode Sphere-Exclusion (M5F), Jeux : J2, J4 et J6, k = 1000

4.8.6 Conclusion

Les méthodes comparées ne sont pas influencées par le jeu initial sans oultier comme nous l'avons déjà vu pour les jeux avec outliers.

4.9 Conclusion de la sélection par diversité

Dans un premier temps, nous avons analysé la robustesse des méthodes par différents moyens :

- par la répétition de sélection lorsqu'un paramètre aléatoire intervient
- en faisant varier la taille des échantillons
- en faisant varier le contenu (J1, J3, J5) et le type (avec ou sans outliers) des jeux initiaux.

En effet, lorsque des initialisations ou des choix de molécules suivantes devaient être effectués aléatoirement, 10 lancements différents des méthodes concernées ont été effectués. L'étude des résultats ainsi obtenus a montré que les méthodes donnent des résultats comparables quelque soit le tirage utilisé. Les résultats étudiés sont donc reproductibles pour tout tirage aléatoire uniforme sans remise.

Ensuite, différents jeux initiaux ont été donnés en entrée pour s'assurer que les résultats obtenus pour un jeu étaient comparables avec tout autre jeu ayant les mêmes propriétés ou des propriétés différentes. Tout d'abord, les méthodes ont été testées avec trois jeux différents contenant des outliers, puis avec trois jeux différents sans outlier. Dans les deux cas les résultats sont restés comparables. Ensuite, la qualité des échantillons divers obtenus à partir d'un jeu avec outliers a été comparée à la qualité des échantillons obtenus à partir d'un jeu sans outlier. Ces deux jeux initiaux (ayant des propriétés différentes) ont donné des échantillons de qualité comparables par rapport à leur jeu de départ.

Enfin, nous avons produit des échantillons de différentes tailles. Quelle que soit la taille de l'échantillon, celui-ci possédait les mêmes propriétés pour chaque méthode.

Les méthodes comparées montrent donc toutes une très bonne robustesse quels que soient le type de données en entrée et la taille de l'échantillon souhaitée.

Ensuite, nos différentes analyses montrent que certains critères sont plus importants que d'autres pour conclure sur la diversité d'un échantillon et qu'ils ne doivent pas être étudiés séparément. Les critères les plus importants pour la diversité de l'échantillon s'avèrent être le rayon maximum et le rayon moyen, la somme des dissimilarités entre chaque centre de groupe (ou individus de l'échantillon) et leur centre plus proche (SDNN), et la représentativité (SDDep). Ces critères ne peuvent être étudiés seuls.

En effet, le rayon maximum doit être le plus faible possible mais il est également intéressant qu'il soit proche du rayon moyen pour s'assurer d'une faible dispersion de la taille des rayons et ainsi conclure à un quadrillage homogène de l'espace. Cette conclusion se confirme d'ailleurs avec une bonne valeur du critère SDNN. Celui-ci étant une somme, il convient de l'analyser avec précaution en s'appuyant sur l'histogramme de distribution des distances. Ce dernier peut permettre de mettre en lumière une grande dispersion des distances. En effet, la somme de très grandes et très faibles distances entre centres pourrait donner un critère SDNN de bonne qualité mais la conclusion que l'on pourrait en tirer, à savoir un échantillon divers, serait alors erronnée.

De même, nous avons montré avec l'analyse de la méthode k-medoïds qu'une somme élevée des dissimilarités entre chaque point du jeu initial et son représentant ne signifiait pas forcément une bonne représentativité. Ceci a été démontré par l'étude de l'histogramme

de distribution des distances qui montrait une grande dispersion dans celles-ci.

Ensuite, le critère SDTot (somme des dissimilarités entre les centres deux à deux) a certes été faible pour des méthodes donnant des échantillons peu divers, mais ne permet pas une bonne discrimination des méthodes donnant des échantillons divers.

Le dernier critère étudié, la distance entre chaque singleton et son individu le plus proche, a permis de montrer que certaines méthodes de sélections pouvaient être utilisées pour identifier des outliers. C'est d'ailleurs grâce à cette identification que nos jeux sans outlier ont été construits.

Tout d'abord, la méthode de sélection par tirage aléatoire uniforme sans remise produit des échantillons peu divers, ils sont d'ailleurs comme nous l'avions défini dans le chapitre 1, plutôt représentatifs au sens statistique du terme du jeu initial.

Ensuite notre implémentation du problème k-center semble donner des échantillons produisant une couverture homogène de l'espace. Ils ont un rayon maximum faible et une faible dispersion de la taille des rayons. De plus ils sont bien représentatifs au sens d'une couverture homogène de l'espace occupé par le jeu initial. Aucune différence significative n'a été observée quant aux différentes initialisations utilisées (aléatoire ou avec les résultats de FFT).

La méthode k-medoïds en revanche est très sensible au type d'initialisation utilisée. En effet, l'initialisation avec les résultats de FFT donne des réusltats nettement meilleurs que l'initialisation aléatoire. De plus le travail avec des centres virtuels semble donner des échantillons un peu plus divers que le travail en centres réels.

Maximum-Dissimilarity produit des échantillons légèrement meilleurs pour une initialisation à la molécule centrale du jeu initial et notamment pour les critères rayon maximum et SDDep (représentativité). De plus, le critère de choix de la molécule suivante a une grande influence sur la qualité des échantillons. En effet, le critère MaxMin donne des échantillons bien plus divers que le critère MaxSum. Le critère MaxMin correspond à la méthode dite FFT qui est ensuite utilisée pour initialiser d'autres méthodes et qui est également considérée comme l'une des plus efficaces parmi les méthodes de diversité existantes.

Les résultats de la méthode Sphere-Exclusion quant à eux ne sont aucunement influencés par le critère de choix de la molécule suivante. En revanche, l'initialisation à la molécule centrale semble donner de meilleurs résultats, notamment pour le critère de représentativité. De plus pour une initialisation non aléatoire, les critères de choix de la molécule suivante (sauf l'aléatoire) donnent tous exactement les mêmes échantillons.

Enfin, nous pouvons conclure sur la qualité des méthodes les unes par rapport aux autres. En effet, la comparaison des méthodes montrent que le tirage aléatoire et la méthode k-medoïds initialisée aléatoirement ne donnent pas de bons échantillons divers. Cette dernière est un peu meilleure avec une initialisation avec FFT. Cependant, étant donné le temps de calcul de cette méthode et la qualité relative de la diversité de ses échantillons, il n'est pas intéressant de l'utiliser à la place des autres méthodes décrites ci-dessous.

En effet, les méthodes FFT et Sphere-Exclusion donnent de bons échantillons divers de qualité très proches et souvent équivalentes. Elles sont notamment meilleurs que les autres pour le critère SDNN.

Ensuite, notre implémentation du problème k-center ressort comme la méthode donnant les échantillons les plus divers, tout en ayant un temps de calcul raisonnable. Les critères de rayon et de représentativité sont les meilleurs pour les échantillons produits par notre implémentation. De plus leur critère SDNN est presqu'aussi bon que celui des

échantillons produits par les méthodes Maximum-Dissimilarity et Sphere-Exclusion.

Conclusion

Ce travail a porté sur l'utilisation de la chémoinformatique et de l'apprentissage artificiel pour la sélection par diversité. Les résultats que l'on peut dégager de cette thèse sont les suivants :

Dans un premier temps, nous avons formalisé un nouveau critère de diversité. Celui-ci prend en compte plusieurs contraintes : la diversité mais également la représentativité de l'espace. Il en existe d'ailleurs plusieurs définitions, celle que nous donnons ici est formalisée par un critère mathématique et induit implicitement une sorte de quadrillage de l'espace chimique.

La création de ce nouveau critère nous a permis de construire une méthode de sélection par diversité efficace et adaptée à de grands jeux de données. En se basant sur des techniques d'apprentissage déjà existantes, une méthode différente de celles utilisées dans le domaine de la chémoinformatique, a été proposée. Celle-ci nous permet d'obtenir de meilleurs résultats qu'avec les méthodes connues dans un temps d'exécution raisonnable et acceptable.

Ensuite, ce critère nous a également permis de définir un nouveau type d'évaluation de la diversité : le rayon maximum d'une partition. Celui-ci est applicable à toutes les méthodes même celles qui ne se basent pas sur la partition de l'espace en groupes de molécules. De plus l'étude de nos résultats a mis en lumière les atouts et les faiblesses de plusieurs autres critères d'évaluation couramment utilisés en chémoinformatique. Nous avons d'ailleurs vu qu'il est nécessaire d'en utiliser plusieurs en fonction de l'objectif de la sélection.

L'état de l'art en chémoinformatique sur la collecte, le traitement des molécules et leur sélection a fait l'objet d'une publication :

Dubois, J. ; Bourg, S. ; Vrain, C. ; Morin-Allory, L. Collections of Compounds - How to Deal with them ? *Current Computer-Aided Drug Design.* **2008**, 156-168.

Ensuite mes travaux préliminaires de collecte de données, ainsi que mes recherches sur la diversité m'ont permis de participer à une publication ainsi qu'à un poster :

Le Guilloux, V. ; Colliandre, L. ; Bourg, S. ; Guénegou, G. ; Dubois-Chevalier, J. ; Morin-Allory, L. Visual Characterization and Diversity Quantification of Chemical Libraries : 1. Creation of Delimited Reference Chemical Subspaces. *Journal of Chemical Information and Modeling.* **2011**, 1762-1774.

Le Guilloux, V. ; Guenegou, G. ; Bourg, S. ; Dubois, J. ; Morin-Allory, L. ; Colliandre,

L. Implementation of a reference “chemical space” for HTS commercial compounds : application to the comparison of chemicals libraries **2010**. *Seconde école d'été de Strasbourg sur la chemoinformatique - Obernai et Ecole Thématique de Criblage - Marseille.*

En conclusion, la thèse a permis de créer un nouveau critère de diversité et d'apporter une nouvelle méthode de sélection de molécules. Elle sera implémentée dans l'outil de l'équipe de chémoinformatique : Screening Assistant. Dans l'équipe d'apprentissage, de nouveaux travaux sur le cloud-computing laissent entrevoir une réduction considérable du temps de calcul de cette méthode. La recherche de plusieurs milliers de composés divers parmi une base de 7 millions de composés, comme celle constituée dans l'équipe de chémoinformatique, ne serait alors qu'une question de jours.

Enfin il reste à étudier l'impact sur la diversité des différents types de descripteurs (réels pour les physico-chimiques, binaires pour les fingerprints, graphes...) pour tenter de découvrir une nouvelle mesure de similarité entre molécules prenant en compte tous les aspects faisant l'activité d'une molécule : sa structure, ses intéractions potentielles...

Bibliographie

[1] D. Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28 :31–36, 1988.

[2] InChI. http ://iupac.org/web/ins/2004-039-1-800.

[3] CAS Registry website. http ://www.cas.org/expertise/cascontent/registry/regsys.html.

[4] M. M. Hann and T. I. Oprea. Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.*, (8) :255–263, 2004.

[5] R. S. Bohacek, C. Martin, and W. C. Guida. The art and practice of structure-based drug design : a molecular modeling perspective. *Med. Res. Rev.*, 16 :3–50, 1996.

[6] P. Ertl. Cheminformatics analysis of organic substituents : Identification of the most common. *J. Chem. Inf. Comput. Sci.*, 43 :374–380, 2003.

[7] T. Scior, P. Bernard, J. L. Medina-Franco, and G. M. Maggiora. Large compound databases for structure-activity relationships studies in drug discovery. *Mini Rev. Med. Chem.*, 7 :851–860, 2007.

[8] T. Engel. Basic overview of chemoinformatics. *J. Chem. Inf. Model.*, 46(6) :2267–2277, 2006.

[9] J. Gasteiger. *Handbook of Chemoinformatics : From Data to Knowledge in 4 Volumes*. WILEY-VCH Verlag GmbH & Co, 2003.

[10] Chemical Abstract Services. http ://www.cas.org/casdb.html.

[11] Beilstein Database. http ://www.beilstein.com/.

[12] ChemNavigator. http ://www.chemnavigator.com/.

[13] J. J. Irwin and B. K. Shoichet. Zinc - a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, 45(1) :177–182, 2005.

[14] French National Chemical Library. http ://chimiotheque-nationale.enscm.fr/.

[15] J. Chen, S. J. Swamidass, Y. Dou, J. Bruand, and P. Baldi. Chemdb : a public database of small molecules and related chemoinformatics resources. *Bioinformatics*, 21(22) :4133–4139, 2005.

[16] ChemIDPlus. http ://chem.sis.nlm.nih.gov/chemidplus/.

[17] e molecule. http ://www.emolecules.com/.

[18] Available Chemicals Directory. http ://www.daylight.com/products/acd.html.

[19] CHEMCATS. http ://www.cas.org/expertise/cascontent/chemcats.html.

[20] ChemSource. http ://www.chemsources.com/chemonline.html.

[21] C. P. Austin, L. S. Brady, T. R. Insel, and F. S. Collins. Pubchem. *Science*, 306 :1138, 2004.

[22] I. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. Palyulin, E. Radchenko, N. Zefirov, A. Makarenko, V. Tanchuk, and V. Prokopenko. Virtual computational chemistry laboratory–design and description. *J. Comput-Aided. Mol. Des.*, 19(6) :453–463, 2005.

[23] K. P. Seiler, G. A. George, M. P. Happ, N. E. Bodycombe, H. A. Carrinski, S. Norton, S. Brudz, J. P. Sullivan, J. Muhlich, M. Serrano, P. Ferraiolo, N. J. Tolliday, S. L. Schreiber, and P. A. Clemons. Chembank : a small-molecule screening and chemoinformatics resource database. *Nucl. Ac. Res.*, pages 1–9, 2007.

[24] T. Girke, L. C. Cheng, and N. Raikhel. Chemmine. *Plant Physiol.*, 138 :573, 2005.

[25] C. Brooksbank, G. Cameron, and J. Thornton. Chebi : Chemical entities of biological interest. *Nucl. Ac. Res.*, 33, 2005.

[26] Drug Bank. http ://www.drugbank.ca/.

[27] Prestwick Chemical Library. http ://www.prestwickchemical.fr.

[28] W. Lutz. Current status of virtual combinatorial library design. *QSAR Comb. Sci.*, 24(7) :809–823, 2005.

[29] D. K. Agrafiotis and E. J. Martin. Advances in combinatorial library design. *J. Mol. Graph. Model.*, 18(4/5), 2000.

[30] P. Sharma, S. Salapaka, and C. Beck. A scalable approach to combinatorial library design for drug discovery. *J. Chem. Inf. Model.*, 48(1) :27–41, 2008.

[31] T. Fink and J. L. Reymond. Virtual exploration of the chemical universe up to 11 atoms of c, n, o, f : Assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.*, 47(2) :342–353, 2007.

[32] L. Arve, T. Voigt, and H. Waldmann. Charting biological and chemical space : Pssc and sconp as guiding principles for the development of compound collections based on natural product scaffolds. *QSAR Comb. Sci.*, 25(5-6) :449–456, 2006.

[33] Q. Liao, J. Yao, and S. Yuan. Svm approach for predicting logp. *Mol. Div.*, 10(3) :301–309, 2006.

[34] J. Gola, O. Obrezanova, E. Champness, and M. Segall. Admet property prediction : The state of the art and current challenges. *QSAR Comb. Sci.*, 25(12) :1172–1180, 2006.

[35] QSARIS. http ://www.scivision.com/qsaris.html.

[36] Cerius2. http ://www.accelrys.com/products/cerius2/.

[37] Volsurf. http ://www.moldiscovery.com/soft_volsurf.php.

[38] DRAGON. http ://www.talete.mi.it/products/dragon_description.htm.

[39] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The chemistry development kit (cdk) : An open-source java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.*, 43(2) :493–500, 2003.

[40] R. Todeschini and V. Consonni. *Handbook of Molecular Descriptors.* Wiley-VCH, Weinheim, 2000.

[41] J. A. Haigh, B. T. Pickup, J. A. Grant, and A. Nicholls. Small molecule shape-fingerprints. *J. Chem. Inf. Model.*, 45(3) :673–684, 2005.

[42] Daylight. http ://www.daylight.com.

[43] MOLPRINT2D. http ://www.molprint.com.

[44] Fingerprint BCI. http ://www.digitalchemistry.co.uk.

[45] Tripos. http ://www.tripos.com.

[46] Scitegic. http ://www.scitegic.com.

[47] U. Fechner, J. Paetz, and G. Schneider. Comparison of three holographic fingerprint descriptors and their binary counterparts. *QSAR Comb. Sci.*, 24(8) :961–967, 2005.

[48] J. W. Godden, F. L. Stahura, and J. Bajorath. Anatomy of fingerprint search calculations on structurally diverse sets of active compounds. *J. Chem. Inf. Model.*, 45(6) :1812–1819, 2005.

[49] http ://www.ra.cs.uni tuebingen.de/forschung/molsim/welcome_e.html.

[50] D. C. Whitley, M. G. Ford, and D. J. Livingstone. Unsupervised forward selection : A method for eliminating redundant variables. *J. Chem. Inf. Comput. Sci.*, pages 1160–1168, 2000.

[51] P. Baldi, R. W. Benz, D. S. Hirschberg, and S. J. Swamidass. Lossless compression of chemical fingerprints using integer entropy codes improves storage and retrieval. *J. Chem. Inf. Model.*, 47(6) :2098–2109, 2007.

[52] A. T. Balaban, A. Beteringhe, T. Constantinescu, P. A. Filip, and O. Ivanciuc. Four new topological indices based on the molecular path code. *J. Chem. Inf. Model.*, 47(3) :716–731, 2007.

[53] E. Gregori-Puigjané and J. Mestres. Shed : Shannon entropy descriptors from topological feature distributions. *J. Chem. Inf. Model.*, 46(4) :1615–1622, 2006.

[54] J. Devillers and A. T. Balaban. *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach, The Netherlands, 1999.

[55] D. Bonchev. *Information Theoritic Indices for Characterization of Chemical Structures*. Wiley Research Studies Press, New York, 1983.

[56] M. Dash and H. Liu. Feature selection for clustering. In *PAKDD*, pages 110–121, 2000.

[57] I. Guyon and A. Elisseeff. An introduction into variable and feature selection. *Journal of Machine Learning Research*, 3 :1157–1182, 2003.

[58] E. Amaldi and V. Kann. On the approximation of minimizing non zero variables or unsatisfied relations in linear systems. *Theor. Comput. Sci.*, pages 237–260, 1998.

[59] R. Kohavi and G. John. Wrappers for feature selection. *Artif. Intell.*, pages 273–324, 1997.

[60] J. G. Dy and C. E. Brodley. Feature selection for unsupervised clustering. *J. Mach. Learn. Res.*, pages 845–889, 2004.

[61] C. A. Lipinski. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods*, 44(1) :235–249, 2000.

[62] C. A. Lipinski, F. Lombardo, B. W. Dominy, and Paul J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, 23(1-3) :3–25, 1997.

[63] W. P. Walters and M. A. Murcko. Prediction of 'drug-likeness'. *Adv. Drug Deliv. Rev.*, 54(3) :255–271, 2002.

[64] I. Muegge. Selection criteria for drug-like compounds. *Med. Res. Rev.*, 23(3) :302–321, 2003.

[65] M. Vieth, M. G. Siegel, R. E. Higgs, I. A. Watson, D. H. Robertson, K. A. Savin, G. L. Durst, and P. A. Hipskind. Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem.*, 47(1) :224–232, 2004.

[66] C. A. Lipinski. Lead- and drug-like compounds : the rule-of-five revolution. *Drug Discov. Today Technol.*, 1(4) :337–341, 2004.

[67] P. Charifson and W. Walters. Filtering databases and chemical libraries. *J. Comput-Aided Mol. Des.*, 16(5) :311–323, 2002.

[68] D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward, and K. D. Kopple. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.*, 45(12) :2615–2623, 2002.

[69] T. M. Frimurer, R. Bywater, L. Narum, L. N. Lauritsen, and S. Brunak. Improving the odds in discriminating "drug-like" from "non drug-like" compounds. *J. Chem. Inf. Comput. Sci.*, 40(6) :1315–1324, 2000.

[70] C. A. S. Bergstrom, M. Strafford, L. Lazorova, A. Avdeef, K. Luthman, and P. Artursson. Absorption classification of oral drugs based on molecular surface properties. *J. Med. Chem.*, 46(4) :558–570, 2003.

[71] G. Vistoli, A. Pedretti, and B. Testa. Assessing drug-likeness - what are we missing ? *Drug Discov. Today*, 13 :285–294, 2008.

[72] Absolv. Sirius analytical instruments ltd.

[73] iDEA. idea pk express,lion bioscience ag.

[74] Oraspotter. zyxbio : Cleveland oh.

[75] QikProp. Schrödinger, inc.

[76] QMPRPLUS. Simulations plus, inc : Lancaster, ca 93534-2902.

[77] M. P. Gleeson. Generation of a set of simple, interpretable admet rules of thumb. *J. Med. Chem.*, 51 :817–834, 2008.

[78] Y. C. Martin. A bioavailability score. *J. Med. Chem.*, 48(9) :3164–3170, 2005.

[79] A. Monge, A. Arrault, C. Marot, and L. Morin-Allory. Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers. *Mol. Div.*, 10(3) :389–403, 2006.

[80] M. C. Hutter. Separating drugs from nondrugs : A statistical approach using atom pair distributions. *J. Chem. Inf. Model.*, 47(1) :186–194, 2007.

[81] G. M. Rishton. Reactive compounds and in vitro false positives in hts. *Drug Discov. Today*, 2(9) :382–384, 1997.

[82] S. L. McGovern, E. Caselli, N. Grigorieff, and B. K. Shoichet. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.*, 45(8) :1712–1722, 2002.

[83] J. Seidler, S. L. McGovern, T. N. Doman, and B. K. Shoichet. Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J. Med. Chem.*, 46(21) :4477–4486, 2003.

[84] C. A. Lipinski and A. Hopkins. Navigating chemical space for biology and medecine. *Nature*, 432 :855–861, 2004.

[85] A. S. Raghavendra and G. M. Maggiora. Molecular basis sets-a general similarity-based approach for representing chemical spaces. *J. Chem. Inf. Model.*, 47(4) :1328–1340, 2007.

[86] C. M. Dobson. Chemical space and biology. *Nature*, 432 :824–828, 2004.

[87] G. M. Maggiora and V. Shanmugasundaram. *Chemoinformatics*, volume 275, chapter Molecular Similarity Measures, pages 1–50. Human Press, 2004.

[88] D. M. Maniyar, I. T. Nabney, B. S. Williams, and A. Sewing. Data visualization during the early stages of drug discovery. *J. Chem. Inf. Model.*, 46(4) :1806–1818, 2006.

[89] C. M. Bishop. *Neural Networks for Pattern Recognition, 1st Ed.* Oxford University Press, New York, 1995.

[90] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.*, C-18 :401–409, 1969.

[91] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Berlin, 1995.

[92] C. M. Bishop, M. Svensén, and C. K. I. Williams. Gtm : The generative topographic mapping. *Neural Comput.*, 10 :215–234, 1998.

[93] D. Lowe and M. E. Tipping. Neuroscale : Novel topographic feature extraction with radial basis function networks. *Adv. Neural Inf. Proc. Syst.*, 9 :543–549, 1997.

[94] Spotfire. www.spotfire.com.

[95] T. I. Oprea and J. Gottfries. Chemography : The art of navigating in chemical space. *J. Comb. Chem.*, 3 :157–166, 2001.

[96] W. S. Torgerson. *Theory and Methods of Scaling*. Wiley, New-York, 1958.

[97] J. W. Godden and J. Bajorath. A distance function for retrieval of active molecules from complex chemical space representations. *J. Chem. Inf. Model.*, 46(3) :1094–1097, 2006.

[98] G. M. Maggiora and M. A. Johnson. *Concepts and Applications of Molecular Similarity*. John Wiley & Sons, New York, 1990.

[99] R. W. Spencer. Diversity analysis in high throughput screening. *J. Biomol. Screening*, 2 :69–70, 1997.

[100] T. Potter and H. Matter. Random or rational design ? evaluation of diverse compound subsets from chemical structure databases. *J. Med. Chem.*, 41 :478–488, 1998.

[101] Y. C. Martin, J. L. Kofron, and L. M. Traphagen. Do structurally similar molecules have similar biological activity ? *J. Med. Chem.*, 45 :4350–4358, 2002.

[102] Guillaume Cleuziou. *Une méthode de classification non supervisée pour l'apprentissage de règles et la recherche d'informations*. PhD thesis, Université d'Orléans, France, 2004.

[103] W. M. Rand. Objective criteria for the evaluation of clustering methods. *J.Am. Stat. Assoc.*, 66 :846–850, 1971.

[104] P. Jaccard. Etude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37 :547–579, 1901.

[105] Tanimoto T. T. *IBM internal report*, 17th Nov. 1957.

[106] S. H. Cha, S. Choi, and Tappert C. C. Anomaly between jaccard and tanimoto coefficients. *Proceedings of Student-Faculty Research Day,CSIS, Pace university*, May 8th 2009.

[107] A. Lipkus. A proof of the triangle inequality for the tanimoto distance. *J. Math. Chem.*, 26 :263–265, 1999.

[108] A. Monge. *Création et utilisation de chimiothèques optimisées pour la recherche "in silico" de nouveaux composés bioactifs.* PhD thesis, Université d'Orléans, France, 2006.

[109] A. G. Maldonado, J. P. Doucet, M. Petitjean, and B. T. Fan. Molecular similarity and diversity in chemoinformatics : From theory to applications. *Mol. Div.*, 10(1) :39–79, 2006.

[110] P. Willett, J. M. Barnard, and G. M. Downs. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.*, 38 :983–996, 1998.

[111] G. W. Bemis and M. A. Murcko. Properties of known drugs. 1. molecular frameworks. *J. Med. Chem.*, 39 :2887–2893, 1996.

[112] G.W Bemis and MA Murcko. Properties of known drugs. 2. side chains. *J. Med. Chem.*, 42 :5095–5099, 1999.

[113] S. H. Fitzgerald, M. Sabat, and H. M. Geysen. Diversity space and its application to library selection and design. *J. Chem. Inf. Model.*, 46(4) :1588–1597, 2006.

[114] J. Batista, J. W. Godden, and J. Bajorath. Assessment of molecular similarity from the analysis of randomly generated structural fragment populations. *J. Chem. Inf. Model.*, 46(5) :1937–1944, 2006.

[115] M. Rupp, E. Proschak, and G. Schneider. Kernel approach to molecular similarity based on iterative graph similarity. *J. Chem. Inf. Model.*, 47(6) :2280–2286, 2007.

[116] A. Papp, A. Gulyas-Forro, Z. Gulyas, G. Dorman, L. Urge, and F. Darvas. Explicit diversity index (edi) : A novel measure for assessing the diversity of compound databases. *J. Chem. Inf. Model.*, 46(5) :1898–1904, 2006.

[117] R. D. Clark and W. J. Langton. Balancing representativeness agaisnt diversity using optimizable k-dissimilarity and hierarchical clustering. *J. Chem. Inf. Comput. Sci.*, 38 :1079–1086, 1998.

[118] D. M. Bayada, H. Hamersma, and V. J. Van Geerestein. Molecular diversity and representativity in chemical databases. *J. Chem. Inf. Comput. Sci.*, 39 :1–10, 1999.

[119] M. Hassan, J. P. Bielawski, J. C. Hempel, and M. Waldman. Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Div.*, 2 :64–74, 1996.

[120] A. M. Ferguson, D. E. Patterson, C. D. Garn, and T. L. Underiner. Designing chemical libraries for lead discovery. *J. Biomol. Screening*, 1 :65–73, 1996.

[121] B. R. Stockwell. Exploring biology with small organic molecules. *Nature*, 432 :846–854, 2004.

[122] C. H. Reynolds, A. Tropsha, B. L. Pfahler, R. Druker, S. Chakravorty, G. Ethiraj, and W. Zheng. Diversity and coverage of structural sublibraries selected using the sage and sca algorithms. *J. Chem. Inf. Comput. Sci.*, 41 :1470–1477, 2001.

[123] P. Willett. Chemoinformatics-similarity and diversity in chemical libraries. *Curr. Opin. Biotechnol.*, 11(1) :85–88, 2000.

[124] P. H. A. Sneath and R. R. Sokal. *Numerical taxonomy : The principles and practice of numerical classification.* W.H. Freeman, San Francisco, 1973.

[125] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. of the Fifth Berkeley Symp. Math. Stat. Proba.*, pages 281–297, 1967.

[126] M. Ester, H. P. Kriegel, J. Sander, M. Wimmer, and X. Xu. Density-connected sets and their application for trend detection in spatial databases. In *Sec. Inter. Conf. Know. Discov. Data Mining*, pages 226–231, 1996.

[127] M. Ankerst, M. Breunig, H. P. Kriegel, and J. Sander. Optics : Ordering points to identify the clustering structure. In *Proc. Inter. Conf. Manag. Data ACM-SIGMOD*, pages 49–60, 1999.

[128] O. Rabal, R. Pascual, J. I. Borrell, and J. Teixido. Cell-integral-diversity criterion : A proposal for minimizing cluster artifact in cell-based selections. *J. Chem. Inf. Model.*, 47(5) :1886–1896, 2007.

[129] M. Snarey, N. K. Terrett, P. Willett, and D. J. Wilton. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graph. Model.*, 15(6) :372 – 385, 1997.

[130] D. K. Agrafiotis and V.S. Lobanov. An efficient implementation of distance-based diversity measures based on k-d trees. *J. Chem. Inf. Comput. Sci.*, 39 :51–58, 1999.

[131] S. V. Trepalin, V. A. Gerasimenko, A. V. Kozyukov, N. P. Savchuk, and A. A. Ivaschenko. New diversity calculations algorithms used for compound selection. *J. Chem. Inf. Comput. Sci.*, 42(2) :249–258, 2002.

[132] S. D. Pickett, C. Luttman, V. Guerin, A. Laoui, and E. James. Divsel and complib-startegies for the design and comparison of combinatorial libraries using pharmacophore descriptors. *J. Chem. Inf. Comput. Sci.*, 38 :144–150, 1998.

[133] J. Mount, J. Ruppert, W. Welch, and A. N. Jain. Icepick : a flexible surface based system for molecular diversity. *J. Med. Chem.*, 42 :60–66, 1999.

[134] F. E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11 :1–21, 1969.

[135] D. Hawkins. *Identification of Outliers*. Chapman and Hall.

[136] R. B. Dean and W. J. Dixon. Simplified statistics for small numbers of observations. *Anal. Chem.*, 23(4) :636–638, 1951.

[137] B. Peirce. Criterion for the rejection of doubtful observations. *Astron. J.*, II(45) :161–163, 1852.

[138] P. R. Menard, J. S. Mason, I. Morize, and S. Bauerschmidt. Chemistry space metrics in diversity analysis, library design, and compound selection. *J. Chem. Inf. Comput. Sci.*, 38(6) :1204–1213, 1998.

[139] M. Casalegno, G. Sello, and E. Benfenati. Definition and detection of outliers in chemical space. *J. Chem. Inf. Model.*, 48(8) :1592–1601, 2008.

[140] R. Guha, D. Dutta, P. C. Jurs, and C. Ting. R-nn curves : An intuitive approach to outlier detection using a distance based method. *J. Chem. Inf. Model.*, 46(4) :1713–1722, 2006.

[141] P. R. Menard, R. A. Lewis, and J. S. Mason. Rational screening set design and compound selection : cascaded clustering. *J. Chem. Inf. Comput. Sci.*, 38(3) :497–505, 1998.

[142] Z. Drezner. The p-centre problem-heuristic and optimal algorithms. *J. Oper. Res. Soc.*, 35(8) :741–748, 1984.

[143] R. Durier. The general one center location problem. *Math. Oper. Res.*, 20(2) :400–414, 1995.

[144] J. Mihelic and B. Robic. Solving the k-center problem efficiently with a dominating set algorithm. *CIT*, 13(3) :225–234, 2005.

[145] N. Mladenovic, J. Brimberg, P. Hansen, and J. A. Moreno-Pérez. The p-median problem : A survey of metaheuristic approaches. *Europ. J. Oper. Res.*, 179(3) :927–939, 2007.

[146] M. Badoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. In *STOC*, pages 250–257, 2002.

[147] D. S. Hochbaum and D. B. Shmoys. A best possible heuristic for the k-center problem. *Math. Oper. Res.*, 10(2) :180–184, 1985.

[148] B. D. Hudson, R. M. Hyde, E. Rahr, J. Wood, and J. Osman. Parameter based methods for compound selection from chemical databases. *Quant. Struct-Act. Rel.*, 15(4) :285–289, 1996.

[149] B. Gärtner. Fast and robust smallest enclosing balls. In *ESA*, pages 325–338, 1999.

[150] E. Feldman, F. A. Lehrer, and T. L. Ray. Warehouse location under continuous economies of scale. *Manag. Sci.*, 12(9) :670–684, 1966.

[151] S. Salhi and R. A. Atkinson. Subdrop : A modified drop heuristic for location problems. *Location Science*, 3(4) :267 – 273, 1995.

[152] F. E. Maranzana. On the location of supply points to minimize transport costs. *Oper. Res. Quarterly*, (15), 1964.

[153] M. B. Teitz and P. Bart. Heuristic methods for estimating the generalized vertex median of a weighted graph. *Oper. Res.*, (16) :955–961, 1968.

[154] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data : An Introduction to Cluster Analysis*. John Wiley, 1990.

[155] R. T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, pages 144–155. Morgan Kaufmann, 1994.

[156] N. Mladenovic, M. Labbé, and P. Hansen. Solving the p-center problem with tabu search and variable neighborhood search. *Networks*, 42(1) :48–64, 2003.

Annexes

Annexe A

Liste des descripteurs discrets (compteurs)

Nom Desc	Moyenne	Min	Max	ecart-type
a_acc	3.62	0	22	1.40
a_acid	0.0008	0	6	0.04
a_aro	13.34	0	48	5.23
a_base	0.006	0	4	0.1005
a_count	47.64	4	192	11.53
a_don	1.15	0	10	0.96
a_heavy	27.05	4	96	5.98
a_hyd	18.63	0	70	4.97
a_nB	0.0007	0	2	0.028
a_nBr	0.06	0	4	0.27
a_nC	19.90	0	72	5.04
a_nCl	0.26	0	8	0.57
a_nF	0.32	0	39	0.90
a_nH	20.58	0	116	6.62
a_nI	0.00845	0	4	0.1002
a_nN	2.96	0	14	1.44
a_nO	2.92	0	34	1.61
a_nP	0.0027	0	4	0.05
a_nS	0.59	0	7	0.70
b_1rotN	5.53	0	46	2.38
b_ar	13.52	0	48	5.39
b_count	49.94	3	193	12.22
b_double	1.96	0	15	1.29
b_heavy	29.36	3	101	6.83
b_rotN	6.46	0	52	2.65
b_single	34.39	3	180	10.92
b_triple	0.05	0	4	0.24
chiral_u	0.39	0	20	0.86
lip_acc	5.89	0	34	1.95

Suite page suivante ...

Nom Desc	Moyenne	Min	Max	ecart-type
lip_don	1.15	0	12	0.95
lip_violation	0.28	0	4	0.56
opr_brigid	18.45	0	64	5.87
opr_nring	3.30	0	12	1.11
opr_nrot	5.98	0	47	2.46
opr_violation	0.78	0	5	1.038
Max.ringsize	5.99	0	24	0.49
rings	3.30	0	12	1.11
SA_Nb.SO2	0.02	0	2	0.15
SA_Nb.NOS	6.48	0	34	2.13
SA_Nb.NO2	0.06	0	6	0.27
SA_Nb.CF3halogens	0.56	0	33	0.85
SA_Nb.CF3	0.05	0	4	0.23
SA_Nb.badatom	0.001	0	2	0.03

Annexe B

Liste des descripteurs réels

Nom Desc	Moyenne	Min	Max	ecart-type
a_IC	78.21	3.24511	286.561	17.83
a_ICM	1.65	0.591673	2.51221	0.17
apol	57.11	2.771	207.276	13.123
BCUT_PEOE_0	-2.54	-3.06136	-1.67088	0.17
BCUT_PEOE_1	-0.56	-0.991162	0.0430293	0.08
BCUT_PEOE_2	0.57	-0.164806	0.929349	0.11
BCUT_PEOE_3	2.55	1.77248	3.24538	0.15
BCUT_SLOGP_0	-2.60	-3.4458	-1.32505	0.18
BCUT_SLOGP_1	-0.47	-1.09893	0.4202	0.09
BCUT_SLOGP_2	0.67	0.123	1.39269	0.12
BCUT_SLOGP_3	2.59	1.70933	3.25448	0.14
BCUT_SMR_0	-2.26	-2.80691	-1.13806	0.18
BCUT_SMR_1	-0.37	-0.758601	0.1108	0.10
BCUT_SMR_2	0.76	0.1057	1.45536	0.11
BCUT_SMR_3	2.83	1.87979	3.6256	0.14
b_1rotR	0.18	0	0.931035	0.07
b_rotR	0.22	0	0.931035	0.08
balabanJ	1.51	0.665796	9.12744	0.33
bpol	31.58	1.629	132.56	9.004
chi0	19.27	3.41421	71.4205	4.15
chi0_C	13.57	0	52.4551	3.46
chi0v	15.95	1.58111	56.7078	3.49
chi0v_C	12.06	0	49.8126	3.11
chi1	13.01	1.73205	44.9122	2.91
chi1_C	7.61	0	33.6779	2.27
chi1v	9.52	0.507093	31.3254	2.26
chi1v_C	5.86	0	30.0375	1.77
density	0.78	0.544493	2.11855	0.07
diameter	14.08	2	59	3.13
FCharge	0.0051	-1	3	0.075
GCUT_PEOE_0	-0.82	-1.21488	-0.517957	0.028

Suite page suivante ...

Nom Desc	Moyenne	Min	Max	ecart-type
GCUT_PEOE_1	-0.40	-0.587861	-0.116971	0.03004
GCUT_PEOE_2	0.04	-0.324806	0.303147	0.04
GCUT_PEOE_3	2.60	0.996524	3.90934	0.19
GCUT_SLOGP_0	-1.05	-2.26398	-0.191403	0.13
GCUT_SLOGP_1	-0.30	-0.642233	0.2602	0.04
GCUT_SLOGP_2	0.14	-0.037	0.862638	0.05
GCUT_SLOGP_3	2.67	1.14505	3.89639	0.18
GCUT_SMR_0	-0.506	-0.6391	-0.101977	0.02
GCUT_SMR_1	-0.19	-0.299172	0.189575	0.02
GCUT_SMR_2	0.19	-0.0543	0.914855	0.05
GCUT_SMR_3	2.88	1.08699	4.1169	0.18
Kier1	21.35	3.9375	84.9328	4.76
Kier2	9.79	0.666667	47.0389	2.45
Kier3	5.47	0	38.2639	1.62
KierA1	16.67	2.44715	78.4127	3.91
KierA2	7.56	0.723774	42.311	1.96
KierA3	4.21	0	48	1.36
KierFlex	4.73	0.462453	83.3333	1.57
logP	3.46	-5.785	22.386	1.72
logS	-5.25	-30.2166	3.3703	1.85
mr	10.54	0.486678	33.0103	2.32
PEOE_PC+	2.15	0.198693	10.9462	0.53
PEOE_PC-	-2.14	-10.9462	-0.172092	0.53
PEOE_RPC+	0.13	0.01	1	0.03
PEOE_RPC-	0.17	0.0264692	1	0.04
PEOE_VSA+0	80.49	0	493.961	44.70
PEOE_VSA+1	64.41	0	350.076	29.16
PEOE_VSA+2	14.03	0	131.079	14.61
PEOE_VSA+3	15.35	0	112.523	12.83
PEOE_VSA+4	9.28	0	101.376	10.36
PEOE_VSA+5	10.68	0	176.502	10.56
PEOE_VSA+6	3.31	0	74.1075	7.19
PEOE_VSA-0	71.38	0	418.296	31.81
PEOE_VSA-1	61.45	0	518.392	39.72
PEOE_VSA-2	4.74	0	135.724	9.56
PEOE_VSA-3	5.07	0	464.437	11.12
PEOE_VSA-4	9.22	0	128.066	12.72
PEOE_VSA-5	20.62	0	203.586	14.62
PEOE_VSA-6	4.57	0	82.4266	6.71
PEOE_VSA_HYD	316.97	34.4642	1198.59	74.84
PEOE_VSA_FHYD	0.84	0.22489	1	0.075
PEOE_VSA_FNEG	0.47	0	0.952907	0.08
PEOE_VSA_FPNEG	0.09	0	0.567468	0.048
PEOE_VSA_FPOL	0.15	0	0.77511	0.075

Suite page suivante ...

Nom Desc	Moyenne	Min	Max	ecart-type
PEOE_VSA_FPOS	0.52	0.0470929	1	0.08
PEOE_VSA_FPPOS	0.06	0	0.427868	0.03
PEOE_VSA_NEG	177.09	0	708.735	47.45
PEOE_VSA_PNEG	34.42	0	250.54	17.90
PEOE_VSA_POL	57.70	0	464.607	27.95
PEOE_VSA_POS	197.58	5.48143	889.589	55.83
PEOE_VSA_PPOS	23.27	0	214.067	13.97
PC+	2.15	0.199	10.948	0.53
PC-	-2.14	-10.947	-0.172	0.53
petitjean	0.48	0	0.5	0.02
petitjeanSC	0.92	0	1	0.07
Q_PC+	2.15	0.199	10.948	0.53
Q_VSA_FHYD	0.84	0.22489	1	0.075
Q_VSA_FNEG	0.47	0	0.952907	0.08
Q_VSA_FPNEG	0.09	0	0.567468	0.04
Q_VSA_FPOL	0.15	0	0.77511	0.075
Q_VSA_FPOS	0.52	0.0470929	1	0.08
Q_VSA_FPPOS	0.06	0	0.427868	0.038
Q_VSA_HYD	317.03	34.4642	1198.59	74.82
Q_VSA_NEG	176.77	0	708.735	47.44
Q_VSA_PNEG	34.39	0	250.54	17.89
Q_VSA_POL	57.63	0	464.607	27.93
Q_VSA_POS	197.89	5.48143	889.589	55.91
Q_VSA_PPOS	23.24	0	214.067	13.97
radius	7.29	1	30	1.58
RPC+	0.13	0.0164034	1	0.03
RPC-	0.17	0.0264609	1	0.049
SA_CFMS	0.42	-0.00745021	7.8893	0.59
SA_PDL	0.36	-0.00745021	4	0.46
SA_PLL	1.38	0	5.35367	0.95
SlogP	3.76	-5.3972	22.4677	1.52
SlogP_VSA0	18.40	0	261.467	18.17
SlogP_VSA1	38.33	0	209.534	25.48
SlogP_VSA2	26.52	0	616.976	21.89
SlogP_VSA3	36.96	0	368.792	36.45
SlogP_VSA4	18.89	0	231.634	17.71
SlogP_VSA5	26.77	0	592.149	30.11
SlogP_VSA6	1.39	0	29.5215	2.65
SlogP_VSA7	130.14	0	576.221	60.35
SlogP_VSA8	23.43	0	641.526	34.43
SlogP_VSA9	86.25	0	583.444	50.45
SMR	10.46	0.5163	33.2472	2.32
SMR_VSA0	45.24	0	521.069	27.88
SMR_VSA1	31.95	0	586.569	24.89

Suite page suivante ...

Nom Desc	Moyenne	Min	Max	ecart-type
SMR_VSA2	16.77	0	305.61	18.92
SMR_VSA3	13.64	0	138.977	9.04
SMR_VSA4	13.66	0	131.589	13.82
SMR_VSA5	160.05	0	849.568	62.15
SMR_VSA6	39.44	0	368.792	33.95
SMR_VSA7	86.33	0	634.761	52.40
TPSA	74.69	0	445.58	26.84
VAdjEq	0.41	0.15208	0.954434	0.06
VAdjMa	5.83	2.58496	7.65821	0.37
VDistEq	3.54	0.863121	5.52664	0.33
VDistMa	9.13	3.44644	12.9594	0.70
vdw_area	374.67	56.1507	1319.49	79.25
vdw_vol	498.107	37.5673	1703.77	111.072
vsa_acc	33.94	0	204.959	16.056
vsa_acid	0.012	0	87	0.68
vsa_base	0.01	0	60.213	0.43
vsa_don	6.25	0	80.5405	6.47
vsa_hyd	276.45	0	1114.68	71.49
vsa_other	37.01	0	267.59	19.68
vsa_pol	43.30	0	232.093	18.72
Weight	389.39	56.064	1487.86	84.22
weinerPol	41.86	0	165	12.17
weinerPath	2237.25	9	54448	1345.39
zagreb	140.88	10	496	34.03

Annexe C

Résultats des échantillons de 1000 molécules sur J3 et J5 pour tous les paramétrages des méthodes

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M2A J3	12.8	11.9	1.1	186.6	14193.5	7.3	115.3	5.6
M2B J3	12.9	11.9	1.1	186.5	14460.3	7.6	115.3	5.5

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M2A J3	365621.3	0.1	12.8	1.9	18311692.0	7.3	241.7	16.2	429.0
M2B J3	370427.1	0.1	12.9	1.9	18329722.0	7.6	241.7	16.1	411.0

a. NN $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep $= \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE C.1: Résultats pour la méthode k-center (M2), Jeu : J3, k = 1000

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M2A J5	12.4	11.6	0.9	142.7	13234.5	6.8	172.5	7.3
M2B J5	12.5	11.6	0.9	143.6	13456.3	6.5	172.5	7.8

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M2A J5	361169.8	0.1	12.4	1.8	16960874.0	6.8	280.6	21.4	333.0
M2B J5	370125.6	0.1	12.5	1.8	16981253.0	6.5	280.6	21.3	312.0

a. NN $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep $= \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE C.2: Résultats pour la méthode k-center (M2), Jeu : J5, k = 1000

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M3A J3	202.3	13.8	9.2	30.9	8532.5	3.6	57.0	2.9
M3B J3	97.0	12.8	6.1	21.8	8324.7	3.6	158.0	5.9
M3C J3	15.6	12.2	1.3	127.5	13495.5	5.7	115.3	6.2
M3D J3	15.5	12.2	1.3	87.5	12932.5	4.6	115.3	6.5

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M3A J3	330578.0	0.0	202.3	3.0	9011789.0	3.6	83.7	6.2	4.1
M3B J3	314058.5	0.0	97.0	2.6	9315830.0	3.6	200.6	10.6	5.1
M3C J3	344530.6	0.0	15.6	1.9	17598204.0	5.7	241.7	16.5	320.0
M3D J3	330667.8	0.0	15.5	1.9	17243582.0	4.6	241.7	16.8	314.0

a. NN $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep $= \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE C.3: Résultats pour la méthode k-medoïds (M3), Jeu : J3, k = 1000

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M3A J5	200.9	14.3	12.3	30.4	8681.9	3.4	107.2	4.1
M3B J5	179.4	12.8	7.6	24.5	8439.5	4.2	164.6	6.1
M3C J5	14.5	11.8	1.1	109.4	12381.0	4.4	172.5	7.8
M3D J5	14.9	11.9	1.1	67.3	11979.1	5.0	172.5	8.0

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M3A J5	334411.1	0.0	200.9	3.2	9143718.0	3.4	154.6	8.1	3.4
M3B J5	321869.1	0.0	179.4	2.7	9094755.0	4.2	193.3	9.8	3.1
M3C J5	340443.1	0.0	14.5	1.9	15977555.0	4.4	280.6	20.9	225.0
M3D J5	330549.0	0.0	14.9	1.9	15693214.0	5.0	280.6	21.1	217.0

a. NN $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep $= \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE C.4: Résultats pour la méthode k-medoïds (M3), Jeu : J5, k = 1000

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M4A J3	24.5	16.8	2.9	1089.7	12485.9	0.0	115.3	6.3
M4B J3	14.2	13.0	1.2	242.2	16298.5	14.2	115.3	4.9
M4C J3	22.9	15.7	2.3	1494.6	12487.7	0.0	115.3	6.3
M4D J3	14.1	12.9	1.2	553.1	16266.2	14.1	115.3	4.9
M4E J3	27.0	19.3	4.2	294.0	8752.6	0.0	32.3	6.2
M4F J3	14.2	13.0	1.2	170.1	16282.6	14.2	115.3	4.9

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M4A J3	595234.2	0.0	24.5	3.2	21441876.0	0.0	241.7	16.2	553.7
M4B J3	445365.1	0.0	14.2	2.0	18549046.0	14.2	241.7	15.7	293.5
M4C J3	482094.3	0.0	22.9	2.9	21440228.0	0.0	241.7	16.2	564.0
M4D J3	427868.5	0.0	14.1	2.1	18576612.0	14.1	241.7	15.7	296.0
M4E J3	738335.8	0.0	27.0	3.9	27628676.0	0.0	241.7	33.8	445.0
M4F J3	447869.7	0.0	14.2	2.0	18569972.0	14.2	241.7	15.7	297.0

a. NN $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep $= \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE C.5: Résultats pour la méthode Maximum-Dissimilarity (M4), Jeu : J3, k = 1000

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M4A J5	21.0	16.1	2.7	755.5	11776.2	2.3	172.5	7.7
M4B J5	13.6	12.6	0.9	161.9	15545.6	13.6	172.5	6.9
M4C J5	20.5	15.1	2.1	1205.1	11778.4	2.3	172.5	7.7
M4D J5	13.5	12.6	1.0	471.5	15545.5	13.6	172.5	6.9
M4E J5	22.7	18.0	3.2	192.4	8546.5	0.0	24.0	5.4
M4F J5	13.6	12.6	0.9	143.6	15533.7	13.6	172.5	6.9

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M4A J5	563504.8	0.0	21.0	3.1	20124792.0	2.3	280.6	21.2	455.1
M4B J5	431294.9	0.0	13.6	2.0	17242424.0	13.6	280.6	20.1	216.2
M4C J5	468102.3	0.0	20.5	2.7	20122654.0	2.3	280.6	21.2	457.0
M4D J5	416962.6	0.0	13.5	2.0	17251108.0	13.6	280.6	20.1	212.0
M4E J5	653380.5	0.0	22.7	3.5	30477586.0	0.0	280.6	48.1	390.0
M4F J5	431990.5	0.0	13.6	2.0	17302728.0	13.6	280.6	20.1	217.0

a. NN $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep $= \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE C.6: Résultats pour la méthode Maximum-Dissimilarity (M4), Jeu : J5, k = 1000

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M5A J3	16.5	13.1	1.4	327.4	16114.3	14.3	115.3	4.9
M5B J3	16.6	13.1	1.4	271.3	16111.2	14.3	115.3	4.9
M5C J3	16.9	13.1	1.4	240.5	16112.6	14.3	115.3	4.9
M5D J3	16.6	13.1	1.4	243.2	16109.7	14.3	115.3	4.9
M5E J3	15.8	13.0	1.3	614.8	16115.7	14.3	115.3	4.9
M5F J3	15.5	13.0	1.3	603.8	16112.7	14.3	115.3	4.9
M5G J3	15.5	13.0	1.3	603.8	16112.7	14.3	115.3	4.9
M5H J3	15.5	13.0	1.3	603.8	16112.7	14.3	115.3	4.9
M5I J3	16.9	13.1	1.4	254.1	16110.8	14.3	115.3	4.9
M5J J3	16.7	13.1	1.4	162.7	16115.7	14.3	115.3	4.9
M5K J3	16.7	13.1	1.4	162.7	16115.7	14.3	115.3	4.9
M5L J3	16.7	13.1	1.4	162.7	16115.7	14.3	115.3	4.9

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M5A J3	451945.6	0.0	16.5	2.1	18872142.0	14.3	241.7	15.7	305.7
M5B J3	458457.1	0.0	16.6	2.1	18881122.0	14.3	241.7	15.7	304.8
M5C J3	459383.5	0.0	16.9	2.2	18879036.0	14.3	241.7	15.7	305.1
M5D J3	458046.1	0.0	16.6	2.1	18876538.0	14.3	241.7	15.7	304.8
M5E J3	430495.5	0.0	15.8	2.1	18874262.0	14.3	241.7	15.7	306.0
M5F J3	429890.6	0.0	15.5	2.1	18880198.0	14.3	241.7	15.7	308.0
M5G J3	429890.6	0.0	15.5	2.1	18880198.0	14.3	241.7	15.7	308.0
M5H J3	429890.6	0.0	15.5	2.1	18880198.0	14.3	241.7	15.7	308.0
M5I J3	458459.1	0.0	16.9	2.2	18878798.0	14.3	241.7	15.7	304.6
M5J J3	465303.0	0.0	16.7	2.2	18883512.0	14.3	241.7	15.7	306.0
M5K J3	465303.0	0.0	16.7	2.2	18883512.0	14.3	241.7	15.7	306.0
M5L J3	465303.0	0.0	16.7	2.2	18883512.0	14.3	241.7	15.7	306.0

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE C.7: Résultats pour la méthode Sphere-Exclusion (M5), Jeu : J3, k = 1000

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M5A J5	14.1	12.7	1.0	227.2	15472.0	13.8	172.5	6.9
M5B J5	14.3	12.7	1.0	149.2	15479.1	13.8	172.5	6.9
M5C J5	14.3	12.7	1.0	169.1	15471.8	13.8	172.5	6.9
M5D J5	14.4	12.7	1.0	149.0	15480.7	13.8	172.5	6.9
M5E	13.9	12.6	1.0	488.9	15430.0	13.8	172.5	6.9
M5F	13.8	12.6	0.9	475.9	15481.3	13.8	172.5	6.9
M5G	13.8	12.6	0.9	475.9	15481.3	13.8	172.5	6.9
M5H J5	13.8	12.6	0.9	475.9	15481.3	13.8	172.5	6.9
M5I J5	14.1	12.7	0.9	164.2	15478.5	13.8	172.5	6.9
M5J J5	14.4	12.7	1.0	140.2	15478.4	13.8	172.5	6.9
M5K J5	14.4	12.7	1.0	140.2	15478.4	13.8	172.5	6.9
M5L J5	14.4	12.7	1.0	140.2	15478.4	13.8	172.5	6.9

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M5A J5	431222.4	0.0	14.1	2.0	17530972.0	13.8	280.6	20.1	220.4
M5B J5	437019.4	0.0	14.3	2.0	17553748.0	13.8	280.6	20.1	219.4
M5C J5	436049.7	0.0	14.3	2.0	17539964.0	13.8	280.6	20.1	216.7
M5D J5	437005.0	0.0	14.4	2.0	17556412.0	13.8	280.6	20.1	218.7
M5E J5	418200.4	0.0	13.9	2.1	17451200.0	13.8	280.6	20.1	219.8
M5F J5	417645.4	0.0	13.8	2.0	17557098.0	13.8	280.6	20.1	217.0
M5G J5	417645.4	0.0	13.8	2.0	17557098.0	13.8	280.6	20.1	217.0
M5H J5	417645.4	0.0	13.8	2.0	17557098.0	13.8	280.6	20.1	217.0
M5I J5	435576.0	0.0	14.1	2.0	17499332.0	13.8	280.6	20.1	218.3
M5J J5	436982.8	0.0	14.4	2.0	17562508.0	13.8	280.6	20.1	215.0
M5K J5	436982.8	0.0	14.4	2.0	17562508.0	13.8	280.6	20.1	215.0
M5L J5	436982.8	0.0	14.4	2.0	17562508.0	13.8	280.6	20.1	215.0

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE C.8: Résultats pour la méthode Sphere-Exclusion (M5), Jeu : J5, k = 1000

Annexe D

Résultats pour les différentes tailles d'échantillons pour la méthode M5

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M5F(1000)	14.1	12.9	1.0	559.7	16.105	14.1	127.6	5.2
M5F(500)	17.0	15.3	1.5	1383.2	19.46	17.1	127.6	6.5
M5F(100)	27.9	24.3	3.1	4526.8	32.77	28.0	127.6	10.8

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
	10.68	0.0	14.1	2.2	38.31	14.1	201.7	16.0	310.0
	12.06	0.0	17.0	2.4	44.6	17.1	201.7	19.0	162.0
	13.58	0.0	27.9	3.7	68.49	28.0	201.7	26.5	30.0

a. NN $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep $= \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE D.1: Résultats pour la méthode Sphere-Exclusion (M5), Jeu : J1, k = 1000, 500, 100

Annexe E

Résultats pour des échantillons de 1000 molécules sur J2 pour tous les apramétrages des méthodes

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M2A J2	12.0	11.2	0.8	98.7	11348.9	6.4	22.0	1.8
M2B J2	12.1	11.3	0.8	112.1	11880.8	8.1	22.5	1.5

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M2A J2	344429.3	0.1	12.0	1.8	13697315.0	6.4	65.6	8.1	163.2
M2B J2	354572.2	0.1	12.1	1.8	13890967.0	8.1	65.6	8.1	155.0

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE E.1: Résultats pour la méthode k-center (M2), Jeu : J2, k = 1000

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M3A J2	32.8	12.3	2.7	30.4	8527.1	3.6	27.7	1.8
M3B J2	24.1	11.7	2.1	21.7	8004.8	3.9	35.2	2.0
M3C J2	14.5	11.6	0.9	99.4	10497.0	3.9	22.5	2.5
M3D J2	14.5	11.6	0.9	78.0	10415.6	3.9	22.5	2.6

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M3A J2	326272.2	0.0	32.8	2.3	9083197.0	3.6	57.3	5.4	2.5
M3B J2	311632.3	0.0	24.1	2.1	8941394.0	3.9	57.9	5.7	2.3
M3C J2	331303.5	0.0	14.5	1.9	13068049.0	3.9	65.8	8.4	78.0
M3D J2	327737.3	0.0	14.5	1.9	13023182.0	3.9	65.8	8.4	79.0

a. NN $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep $= \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE E.2: Résultats pour la méthode k-medoïds (M3), Jeu : J2, k = 1000

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M4A J2	21.3	15.9	2.8	754.5	9691.0	0.0	19.1	2.2
M4B J2	13.2	12.3	0.8	131.6	13961.6	13.2	22.9	0.7
M4C J2	20.4	14.9	2.2	1132.2	9696.4	0.0	20.4	2.2
M4D J2	13.2	12.2	0.8	368.5	13960.2	13.2	23.6	0.7
M4E J2	23.0	17.6	3.9	161.4	7465.1	0.0	17.8	4.6
M4F J2	13.2	12.3	0.7	112.5	13942.6	13.2	23.6	0.7

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M4A J2	554410.3	0.0	21.3	3.2	18330044.0	0.0	65.8	10.7	417.8
M4B J2	413592.4	0.0	13.2	1.9	14396978.0	13.2	65.7	7.7	51.9
M4C J2	462139.2	0.0	20.4	2.8	18328124.0	0.0	65.8	10.7	418.0
M4D J2	402550.1	0.0	13.2	2.0	14453424.0	13.2	65.8	7.8	49.0
M4E J2	649608.0	0.0	23.0	3.7	19609396.0	0.0	65.8	11.2	357.0
M4F J2	414610.0	0.0	13.2	1.9	14418527.0	13.2	65.8	7.7	49.0

a. NN $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep $= \{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot $= \{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE E.3: Résultats pour la méthode Maximum-Dissimilarity (M4), Jeu : J2, k = 1000

	Rayon Max	Rayon Moyen	EC Rayon	EC Ind	SDNN [a]	SDNN Min	SDNN Max	SDNN EC
M5A J2	13.3	12.2	0.9	155.9	14001.6	13.3	21.8	0.6
M5B J2	13.3	12.2	0.9	125.8	14053.4	13.3	22.0	0.6
M5C J2	13.3	12.2	0.9	132.9	14029.5	13.3	21.8	0.6
M5D J2	13.3	12.2	0.9	112.4	14020.7	13.3	21.4	0.6
M5E J2	13.3	12.2	0.9	369.2	13904.5	13.3	21.2	0.6
M5F J2	13.3	12.2	0.9	367.5	13931.0	13.3	21.0	0.6
M5G J2	13.3	12.2	0.9	367.5	13931.0	13.3	21.0	0.6
M5H J2	13.3	12.2	0.9	367.5	13931.0	13.3	21.0	0.6
M5I J2	13.3	12.2	0.9	145.8	14027.0	13.3	21.8	0.6
M5J J2	13.3	12.2	0.9	109.3	14130.3	13.3	21.0	0.6
M5K J2	13.3	12.2	0.9	109.3	14130.3	13.3	21.0	0.6
M5L J2	13.3	12.2	0.9	109.3	14130.3	13.3	21.0	0.6

	SDDep [b]	SDDep Min	SDDep Max	SDDep EC	SDTot [c]	SDT Min	SDT Max	SDT EC	NB Singletons
M5A J2	412753.4	0.0	13.3	1.9	14937027.0	13.3	65.8	7.9	54.6
M5B J2	414078.6	0.0	13.3	1.9	15050115.0	13.3	65.8	7.9	54.2
M5C J2	413617.8	0.0	13.3	1.9	14995651.0	13.3	65.8	7.9	54.8
M5D J2	415519.0	0.0	13.3	1.9	14964794.0	13.3	65.8	7.9	53.8
M5E J2	401462.9	0.0	13.3	2.0	14759773.0	13.3	65.8	7.9	56.3
M5F J2	401860.0	0.0	13.3	2.0	14800904.0	13.3	65.8	7.9	56.0
M5G J2	401860.0	0.0	13.3	2.0	14800904.0	13.3	65.8	7.9	56.0
M5H J2	401860.0	0.0	13.3	2.0	14800904.0	13.3	65.8	7.9	56.0
M5I J2	412884.5	0.0	13.3	1.9	15010773.0	13.3	65.7	7.9	56.9
M5J J2	414964.2	0.0	13.3	1.9	15171909.0	13.3	65.8	7.9	58.0
M5K J2	414964.2	0.0	13.3	1.9	15171909.0	13.3	65.8	7.9	58.0
M5L J2	414964.2	0.0	13.3	1.9	15171909.0	13.3	65.8	7.9	58.0

a. NN = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(c_i^*, c_j^*)\}$

b. Dep = $\{(m_i, c_j^*) \mid m_i \in \mathcal{M},\ c_j^* \in C^*,\ c_j^* = \underset{j=1\ldots k}{ArgMin}\ d(m_i, c_j^*)\}$

c. Tot = $\{(c_i^*, c_j^*) \mid c_i^*, c_j^* \in C^*,\ c_i^* \neq c_j^*\}$

TABLE E.4: Résultats pour la méthode Sphere-Exclusion (M5), Jeu : J2, k = 1000

Julie DUBOIS-CHEVALIER

Chimiothèques ; vers une approche rationnelle pour la sélection de sous-chimiothèques

Résumé :

La sélection de sous-ensembles de molécules diverses est un enjeu très important de la recherche pharmaceutique. En effet, de la qualité de cette sélection, dépendra la découverte efficace d'un médicament.
De nombreuses méthodes existent pour répondre à cette demande. Certaines sont basées sur la création de groupes de molécules, d'autres sur le principe de dissimilarité inter-moléculaire. Nous proposons dans ce travail, une nouvelle technique à la croisée de ces méthodes, qui permet d'obtenir des sous-ensembles à la fois divers dans l'espace et représentatifs de l'ensemble initial duquel ils sont extraits. Pour créer cette méthode de sélection, nous avons tout d'abord défini et formalisé mathématiquement un critère de diversité, puis nous nous sommes appuyés sur des heuristiques connues en apprentissage artificiel pour concevoir l'algorithme.
Celui-ci a été comparé à d'autres types de sélections de diversité couramment utilisées en chémoinformatique telles que les *k*-medoïds, Maximum-Dissimilarity, Sphere-Exclusion. La formalisation du critère de diversité nous a enfin permis de proposer un nouveau critère d'évaluation de la qualité des sélections.
La méthode et le critère présentés dans ce travail donnent des échantillons divers et représentatifs d'un espace chimique.

Mots clés : Chémoinformatique, apprentissage, diversité, sélection

Chemical librairies ; towards a rational approach for the selection of sub-libraries

Abstract :

The selection of diverse molecules'subsets is a very important stake in the pharmaceutical research. Indeed, the effective discovery of a drug will depend of the quality of this selection. Several methods exist to address this problem. Some of them are based on the creation of groups of molecules, the others on the principle of dissimilarity between chemical compounds. In this work, we propose a new technique, between these two concepts, which allows to obtain subsets, at the same time, diverse in the space and representative from the initial set which they are extracted. First of all, to create this selection method, we defined and formalized mathematically a diversity criterion, then we used heuristics known in machine learning to conceive our algorithm.
This one was compared with the other types of diversity selections usually used in chemoinformatic such as *k*-medoïds, Maximum-Dissimilarity, Sphere-Exclusion. The formalization of the diversity criterion finally allowed us to propose a new criterion of evaluation of the quality of the selections.
The algorithm and the criterion presented in this work give diverse and representative samples of a chemical space.

Keywords : Chemoinformatic, machine learning, diversity, sampling

ICOA, UMR CNRS 6005 et LIFO, EA 4022 - Université d'Orléans - BP 6759 - 45067 Orléans CEDEX 2