Clustering en chémoinformatique pour le raffinement de l'activité des molécules

Maroua Lejmi, Ilef Ben Slima, Bertrand Cuissart, Nida Meddouri, Ronan Bureau, Alban Lepailleur, Jean-Luc Lamotte, Amel Borgi

Clustering en chémoinformatique pour le raffinement de l'activité des molécules

Maroua Lejmi
LIPAH / Universite Tunis El Manar
GREYC / Université de Caen Normandie
Tunis, Tunisie
lejmi.maroua@gmail.com

Ilef Ben Slima
ISMAIK, University of Kairaoun
SM@RTS Sfax, Tunisie
ilef.benslima@crns.rnrt.tn

Bertrand Cuissart
Normandie Univ, UNICAEN
ENSICAEN, CNRS, GREYC
Caen, France
bertrand.cuissart@unicaen.fr

Nida Meddouri
Laboratoire de Recherche de l'EPITA (LRE)
Le Kremlin-Bicetre, France.
nida.meddouri@epita.fr

Ronan Bureau
Normandie Univ, UNICAEN
CERMN
Caen, France
ronan.bureau@unicaen.fr

Alban Lepailleur
Normandie Univ, UNICAEN
CERMN
Caen, France
alban.lepailleur@unicaen.fr

Jean-Luc Lamotte
Normandie Univ, UNICAEN
ENSICAEN, CNRS, GREYC
Caen, France
jean-luc.lamotte@unicaen.fr

Amel Borgi
ISI-LIPAH / Université de Tunis El Manar
Tunis, Tunisie
Amel.Borgi@insat.rnu.tn

***Abstract*—Dans le domaine de la conception des médicaments, la chémoinformatique utilise des méthodes informatiques et mathématiques pour analyser des données chimiques et biologiques et essayer de trouver très en amont des molécules intéressantes. Dans notre contexte, nous transformons les molécules pour ne conserver que leurs caractéristiques pharmacophoriques (partie active de la molécule). L'objectif de ce travail est de raffiner l'activité des molécules qui seront utilisées dans le processus de conception des médicaments en des classes d'activité. Cela permettra aux chimistes et pharmaciens une meilleure visualisation et compréhension de l'activité des molécules, et fournira des données plus fines pour le développement ultérieur d'un modèle de prédiction des molécules d'intérêt thérapeutique.**

Mots clés : Chémoinformatique, conception de médicaments, pharmacophores, activité moléculaire, clustering.

I. Introduction

Ce travail s'inscrit dans le domaine de la chémoinformatique. C'est une discipline qui utilise des méthodes informatiques et mathématiques pour aider à la conception et à la découverte de nouveaux médicaments [1]. Elle est basée sur l'analyse de données chimiques et biologiques, et elle utilise des outils informatiques pour simuler les interactions entre les molécules et les cibles. Le processus de conception des médicaments commence par l'identification de cibles chez le patient (protéines-acides nucléiques classiquement) dont on souhaite moduler l'activité pour traiter une maladie. Il s'agit de déterminer un ensemble de molécules qui interagissent suffisamment avec ces cibles : les molécules candidates qui maximisent l'effet thérapeutique, tout en limitant les effets secondaires et les effets indésirables.

Ainsi, l'objectif est de déterminer automatiquement les assemblages de motifs chimiques responsables d'une activité biologique. Nous souhaitons développer un modèle de prédiction qui contribue à définir des molécules d'intérêt thérapeutique. Plus précisément, il s'agira d'analyser et de prédire l'activité de petites molécules (ligands) envers des kinases (les cibles). Les kinases sont des protéines qui ont un grand rôle dans les voies de signalisation qui régissent le fonctionnement des cellules. A partir de données dont une étiquette est connue (par exemple, des molécules considérées actives ou inactives envers des kinases), le but de la classification supervisée est de construire un classifieur pour prédire l'activité de nouvelles molécules. Le travail présenté ici est un pré-traitement des données moléculaires dont nous disposons. Notre objectif est de raffiner l'activité des molécules. Elles sont actuellement réparties entre deux classes : les molécules actives et les molécules inactives. Or une répartition plus fine et graduelle entre plusieurs classes permettrait aux chimistes et pharmaciens une meilleure visualisation et compréhension de l'activité des molécules. En concertation avec ces derniers, notre objectif est de trouver la meilleure répartition des molécules entre 4 classes d'activité

: classe des molécules très actives, moyennement actives, faiblement actives et inactives.
Dans la suite de cet article, nous commencerons par présenter le contexte de notre travail et les données moléculaires dont nous disposons. Nous détaillerons ensuite l'objectif de ce travail et la démarche suivie. Après la présentation et l'interprétation des résultats expérimentaux, nous conclurons sur les perspectives de notre travail.

II. Contexte et données chimiques

Nous disposons de 1517 molécules dont on connaît l'activité sur les protéines kinases BCR-ABL préparées par le laboratoire CERMN avec notamment 6 familles de descripteurs incluant plusieurs milliers de fingerprints, des voisinages d'atomes selon différentes distances, le poids moléculaire, etc. Les descripteurs appelés aussi pharmacophores sont des graphes complets reliant des caractéristiques chimiques (Accepteur (A) , Donneur (D), Positif (P), Négatif (N), Cycle aromatique (R), structure hydrophobique (H)). Un fingerprint est une liste de pharmacophores avec une indication de présence dans les molécules sous forme d'une description binaire.
La Figure 1 est un exemple d'un pharmacophore représentant 4 caractéristiques chimiques d'une molécule généré par un outil appelée Normastic et utilisé pour la génération des graphes des pharmacophores [2]. Nous travaillons plus précisément sur

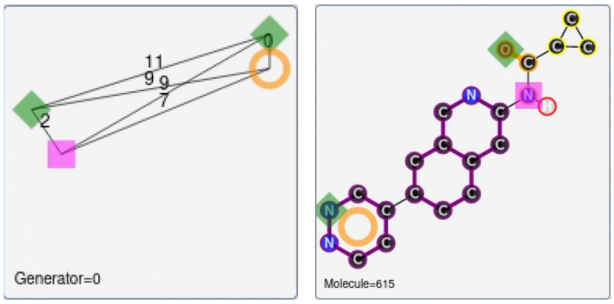

Fig. 1. Transformation d'une molécule en un graphe de pharmacophore

deux jeux de données .
Le premier jeu de données contient une collection de molécules avec leur affinité pour la cible BCR-ABL. L'activité d'une molécule se mesure avec la quantité exprimée en nanomolaire (nM) qu'il faut utiliser pour inhiber ou activer de 50 % l'activité d'une kinase. Cette activité désigne la capacité des molécules à produire un effet biologique. Plus la valeur est petite, plus la molécule est active donc efficace. La Figure 2 présente un extrait de ce fichier. Par exemple la molécule 1 avec un pourcentage d'inhibition de 0.0200 (nM) a une activité élevée : il suffit d'une très faible quantité de cette molécule pour produire une réaction de la cible.

Le deuxième jeu de données s'appuie sur un ensemble de descripteurs de molécules (les pharmacophores).
Comme illustré dans la Figure 3 ci-dessous, les pharmacophores sont représentés en colonne et les molécules en ligne. La cellule contient 1 si la molécule possède ce pharmacophore, et 0 s'il n'est pas présent dans la molécule. Les molécules

0	0.0186
1	0.0200
2	0.0200
3	0.0200
4	0.0200
...	...
1512	157000.0000
1513	236000.0000
1514	243000.0000
1515	267000.0000
1516	392000.0000

Fig. 2. Extrait du fichier des molécules avec leurs activités

peuvent contenir plusieurs pharmacophores.
L'objectif des chimistes est de trouver des pharmacophores qu'ils considèrent intéressants ; ce sont ceux présents dans plusieurs molécules actives.

	|P|P|D|	|3|7|10|	|P|P|A|	|3|15|12|	|P|P|A|	|3|13|16|
1		0		0		0
2		0		0		0
3		0		0		0
4		0		0		0
..		...		...		...
567		0		0		1
568		1		0		0
569		0		1		0
570		0		0		0

Fig. 3. Extrait du fichier des molécules avec leurs activités

Dans [2], les auteurs ont considéré qu'une molécule est active si son activité est entre 0 et 100 (nM) et inactive si elle dépasse 1000 (nM). Les molécules entre 100 et 1000 (nM) ont été retirées du jeu de données.
L'activité des molécules est utilisée pour évaluer la qualité des pharmacophores en calculant la mesure Growth Rate (GR). C'est un rapport de la fréquence des molécules actives contenant ce pharmacophore par rapport à la fréquence de celles qui sont inactives. Pour chaque pharmacophore Pi, la formule de GR est calculée comme suit :

$$GR(Pi) = \frac{Fréquence(Actives)}{Fréquence(Inactives)} \quad (1)$$

Avec :
Fréquence(Actives) : le rapport du nombre de molécules actives dans Pi par le nombre de molécules actives du jeu de données.
Fréquence(Inactives) : le rapport du nombre de molécules inactives dans Pi par le nombre de molécules inactives du jeu de données.

Plus la valeur du GR est élevée, plus le pharmacophore couvre des molécules actives. Un GR égal à 1 indique que le pharmacophore couvre le même rapport de molécules actives et inactives de la base de données. Un GR inférieur à 1 indique que le pharmacophore couvre plus de molécules inactives. Un GR infini indique qu'un pharmacophore ne couvre que des molécules actives. Inversement, avec un GR égal à 0, un pharmacophore ne couvre que des molécules inactives.

Dans les travaux précédents, les pharmacophores dont le GR était supérieur à 3, étaient considérés comme particulièrement intéressants pour les chimistes.

III. Objectifs et démarche

L'objectif du travail présenté dans cet article est le raffinement de l'activité des molécules en intervalles ou classes. Au lieu d'une activité binaire nous souhaitons obtenir des catégories ou classes d'activité. Dans les travaux antérieurs, les molécules ayant une activité entre 100 (nM) et 1000 (nM) étaient exclues. Nous proposons de prendre en compte un dataset contenant des molécules dans cet intervalle, qui peuvent être utiles plus tard dans la classification.

Nous avons classé les molécules en 4 catégories : très active, moyennement active, faiblement active et inactive. Nous avons proposé de faire un clustering sur la plage d'activité des molécules pour déterminer les points de coupure de chaque classe. Nous avons testé plusieurs algorithmes, et avons présenté les résultats obtenus aux chimistes afin de retenir les classes d'activité les plus pertinentes. Cette partie est détaillée dans la section suivante.

Le choix du nombre de classes a été fixé à 4 en prévision des possibilités de représentation graphique. Les chimistes disposent actuellement d'un outil de visualisation et d'exploration de l'espace des pharmacophores n'indiquant que deux classes : actives ou inactives. Le raffinement de l'activité des molécules que nous proposons permettra de former de nouvelles visualisations sur l'espace des pharmacophores en fonction des nouvelles catégories d'activité, et de donner aux chimistes une idée plus fine sur l'intérêt des pharmacophores. Les catégories d'activité obtenues sont utilisées pour évaluer les pharmacophores.

Nous avons adapté la formule du GR d'origine pour exprimer le GR par catégorie d'activité en rapport avec la fréquences des molécules inactives. Chaque pharmacophore est défini par 3 mesures de GR (GR1, GR2 et GR3) exprimées de la manière suivante :

$$GR1(Pi) = \frac{Fréquence(Très \quad Actives)}{Fréquence(Inactives)} \tag{2}$$

$$GR2(Pi) = \frac{Fréquence(Moyennement \quad Actives)}{Fréquence(Inactives)} \tag{3}$$

$$GR3(Pi) = \frac{Fréquence(Faiblement \quad Actives)}{Fréquence(Inactives)} \tag{4}$$

- Le GR1 reflète l'association d'un pharmacophore à la classe des molécules très actives.
- Le GR2 reflète l'association d'un pharmacophore à la classe des molécules moyennement actives.
- Le GR3 reflète l'association d'un pharmacophore à la classe des molécules faiblement actives.
- Un pharmacophore est inactif si les valeurs de ses trois mesures GR1, GR2 et GR3 sont nulles.

Dans le cas où un pharmacophore ne porte que des molécules actives (c'est-à-dire la fréquence des molécules inactives est égale à 0) les 3 valeurs des 3 GRs seront à l'infini (∞) par conséquent nous ne pouvons pas connaître la classe d'activité la plus liée au pharmacophore. C'est pour cette raison que nous proposons un nouvel indicateur pour éviter que le dénominateur (la fréquence de molécules inactives) soit nul. Ceci rend les valeurs des 3 GRs comparables pour connaître la classe à laquelle il appartient le plus. Ces nouveaux indicateurs, notés GRA (Growth Rate Adapté), sont définis comme suit :

$$GRA1(Pi) = \frac{Fréquence(Très \quad Actives)}{Fréquence(Inactives) + 1} \tag{5}$$

$$GRA2(Pi) = \frac{Fréquence(Moyennement \quad Actives)}{Fréquence(Inactives) + 1} \tag{6}$$

$$GRA3(Pi) = \frac{Fréquence(Faiblement \quad Actives)}{Fréquence(Inactives) + 1} \tag{7}$$

Il existe des pharmacophores qui sont plus liés à la classe très active, d'autres à la classe moyennement active et d'autres à la classe faiblement active. Les pharmacophores dont les GRAs sont nuls (GRA1 = 0, GRA2 = 0 et GRA3 = 0) sont liés à la classe inactive. Le tableau I présente quelques exemples de pharmacophores.

TABLE I
EXEMPLES DE 4 PHARMACOPHORES ET LEURS VALEURS DE GRA

	GRA1	GRA2	GRA3
Ph1	10	3	0
Ph2	0	20	10
Ph3	30	10	5
Ph4	20	0	20

Ph1 porte à la fois des molécules très actives et des molécules moyennement actives avec le GRA1 plus élevé donc il est plus lié à la classe très active. Le Ph2 porte plus de molécules moyennement actives que faiblement actives. Le Ph3 possède des molécules de chaque classe mais son GRA1 est plus élevé donc il est considéré comme étant très actif. Le Ph4 possède une fréquence équivalente de molécules très actives et de molécules faiblement actives.

Les interprétations sur la qualité des pharmacophores peuvent être faites en analysant les GRAs de chaque classe d'activité et voir celle la plus liée.

IV. Test et résultats expérimentaux

Afin de raffiner l'activité des molécules, celles-ci sont classées en 4 catégories en fonction de leur activité : très active, moyennement active, faiblement active et inactives. Des algorithmes de clustering sont utilisés. L'objectif est de trouver les points de coupure du domaine de variation de l'activité les plus pertinents, permettant de définir des classes d'activité homogènes.

Plusieurs méthodes de clustering ont été étudiées avec WEKA : K-means [3], Make Density Based K-means, ou MDBSCAN [4], Filtered Clustering [5], Cascade Simple K-means [6], Xmeans [7], EM (Expectation-Maximization) [8], Farthest

First [9],et Hierarchical Clustering [10].
Le tableau II présente les résultats des différents algorithmes de clustering. Chaque case contient en gras, le nombre de molécules de la classe, et entre crochet les valeurs minimum et maximum de l'activité des molécules de la classe.
Les valeurs d'activité des molécules à regrouper en 3 classes (Très active, Moyennement active, Faiblement active) étaient sur [0.01, 1000]. Les molécules avec une activité supérieure à 1000 (nM) ne sont pas prises en compte pour le clustering car elles sont considérées inactives pour les chimistes. Les valeurs de l'activité étaient reparties sur 5 niveaux de grandeur [0.01, 1000], il a été décidé d'utiliser le logarithme de l'activité pour effectuer le clustering.

TABLE II
RÉSULTATS DES ALGORITHMES DE CLUSTERING DES MOLÉCULES

Algorithmes de clustering de WEKA	Cluster 1	Cluster 2	Cluster 3	Cluster 4
SimpleKmeans	**346**[0.0186, 3.02]	**367**[3.16, 70]	**435**[72, 1000]	> 1000
Make density based clustering (kmeans)	**357**[0.0186,3.6]	**362** [3.98,73.3]	**429**[75, 1000]	> 1000
Filtered Clusterer	**346**[0.0186, 3.02]	**367**[3.16, 70]	**435**[72, 1000]	> 1000
Cascade SimpleKmeans	**322**[0.0186, 2.6]	**378**[2.72 ,67]	**448**[68, 1000]	> 1000
Xmeans	**322**[0.0186, 2.6]	**378**[9.1, 67]	**448**[68, 1000]	> 1000
EM	**468**[0.0186,11]	**372**[12,189]	**308**[190,1000]	> 1000
FarthestFirst	**54**[0.0186, 0.116]	**343**[0.150 , 5.02]	**751**[5.1, 1000]	> 1000
Hierarchical clustering	**10**[0.0186, 0.02]	**6**[0.03, 0.04]	**1132**[0.04, 1000]	> 1000

Ces résultats ont été analysés par les chimistes qui ont estimé que les intervalles trouvés avec K-means sont les plus significatifs. Ils sont illustrés sur la Figure 4
Les 4 catégories d'activité obtenues avec les plages d'activité sont donc les suivantes :

- Très active [0, 3.02]
- Moyennement active [3.16, 70]
- Faiblement active [72, 1000]
- Inactive $>$ 1000 nM

Fig. 4. Intervalles d'activités obtenus par K-means

Les intervalles obtenus ne sont pas continus car il n'existe pas des molécules ayant une activité entre 3.02 et 3.16 (nM) ainsi qu'entre 70 et 72 (nM).
Les 4 catégories d'activité ainsi, obtenues seront utilisées dans nos travaux futurs pour une meilleure visualisation des pharmacophores et dans le développement d'un modèle de prédiction des molécules d'intérêt thérapeutique.

La Figure 5 montre le nuage des points dans chaque cluster de K-means. Le cluster 1 est représenté en bleu, le cluster 2 en rouge et le cluster 3 en vert. Les points de chaque cluster sont regroupés étroitement et forment un groupe compact, ce qui signifie que les instances du cluster sont similaires et différentes des instances des autres clusters. K-means a bien séparé les différentes catégories de données même s'ils existent des petits chevauchements entre les clusters. Avec les autres algorithmes, nous avons obtenu plus de chevauchements.

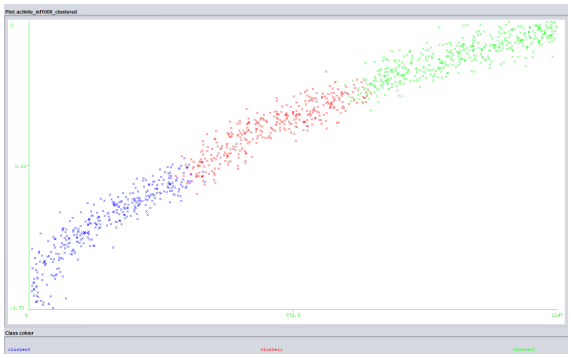

Fig. 5. Nuage de points obtenu par K-means

V. CONCLUSION

Dans ce travail, nous avons proposé de raffiner l'activité des molécules, initialement réparties en 2 classes (actives et inactives) en déterminant 4 classes d'activité : très active, moyennement active, faiblement active et inactive et différents algorithmes de clustering ont été testés. L'analyse des résultats par les chimistes a permis de valider les plages d'activité de chaque classe. Nous avons également adapté la mesure Growth Rate (GR) d'origine pour l'exprimer par catégorie d'activité. Ces nouveaux indicateurs sont utiles pour une évaluation plus fine de la qualité des pharmacophores. Une perspective immédiate de notre travail consiste à convertir cette information fine sur la qualité des pharmacophores en une représentation visuelle interprétable appelée "Réseau de pharmacophores" pour aider les pharmaciens et les chimistes à mieux comprendre la qualité des pharmacophores et à les distinguer visuellement. La répartition plus fine des molécules selon leur activité sera également prise en compte pour le développement d'un modèle de prédiction des molécules d'intérêt thérapeutique.

REMERCIEMENTS

Ce travail a été réalisé au sein des laboratoires, LIPAH (Laboratoire en Informatique en Programmation Algorithmique et Heuristique) de l'Université de Tunis El Manar, GREYC (Groupe de Recherche en Informatique, Image et Instrumentation de Caen) et le CERMN (Centre d'Etudes et de Recherche sur le Médicament de Normandie) de l'Université de Caen Normandie.
Ce travail a été réalisé avec le soutien financier du partenariat Hubert Curien Utique du Ministère de l'Europe et des Affaires Étrangères français et du Ministère de l'Enseignement Supérieur et de la Recherche Scientifique tunisien, dans le cadre du projet PAPRICA (Code Campus France : 47638 VM, Code CMCU : 22G1405)

REFERENCES

[1] Willett, P. (2017). Chemoinformatics: past, present and future. Journal of Computer-Aided Molecular Design, 31(1), 1-6. doi: 10.1007/s10822-016-9978-0

[2] MÉTIVIER, Jean-Philippe, CUISSART, Bertrand, BUREAU, Ronan, et al. The pharmacophore network: a computational method for exploring structure–activity relationships from a large chemical data set. Journal of medicinal chemistry, 2018, vol. 61, no 8, p. 3551-3564.
[3] MACQUEEN, J. Classification and analysis of multivariate observations. In : 5th Berkeley Symp. Math. Statist. Probability. Los Angeles LA USA : University of California, 1967. p. 281-297.
[4] HINNEBURG, A. A density based algorithm for discovering clusters in large spatial databases with noise. In : KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996.
[5] BEYER, K., Goldstein, J., Ramakrishnan, R., Shaft, U. (1999). When is "nearest neighbor" meaningful Database issues in the analysis of large graphs. International Conference on Database Theory, 217-235.
[6] SHEIKHOLESLAMI, Gholamhosein, CHATTERJEE, Surojit, et ZHANG, Aidong. Wavecluster: A multi-resolution clustering approach for very large spatial databases. In : VLDB. 1998. p. 428-439.
[7] PELLEG, Dan, MOORE, Andrew W., et al. X-means: Extending k-means with efficient estimation of the number of clusters. In : Icml. 2000. p. 727-734.
[8] BOYLES, Russell A. On the convergence of the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 1983, vol. 45, no 1, p. 47-50.
[9] MACQUEEN, J. Classification and analysis of multivariate observations. In : 5th Berkeley Symp. Math. Statist. Probability. Los Angeles LA USA : University of California, 1967. p. 281-297.
[10] JOHNSON, Stephen C. Hierarchical clustering schemes. Psychometrika, 1967, vol. 32, no 3, p. 241-254.