Indices de similarité et aide à la décision multicritère appliqués à la découverte et au ciblage de nouvelles substances naturelles

Yassine M. Mejri[1,2], Olivier Cailloux[1],Meltem Öztürk[1], Mehdi A. Beniddir[2]

[1] Université Paris Dauphine, PSL Research University, CNRS, Lamsade, 75016 Paris, France.
mohamed-yassine.mejri@dauphine.eu, olivier.cailloux@dauphine.fr,
meltem.ozturk@dauphine.fr

[2] Université Paris-Saclay Faculty of Pharmacy, CNRS, BioCIS, 91400 Orsay, France France
mehdi.beniddir@universite-paris-saclay.fr

Mots-clés : *aide multicritère à la décision, expérimentation, indices de similarité, chimie des substances naturelles, traitement chimioinformatique.*

1 Introduction

Dans ce travail nous nous focalisons sur l'analyse des indices de similarité utilisés pour comparer la similarité structurale de molécules à travers leurs données spectrales. Les indices de similarité sont au coeur des procédures de décision impliquant des experts chimistes. Ils sont très souvent utilisés lors de la recherche de nouvelles substances naturelles bioactives (point de départ pour un développement en chimie médicinale).

L'étude que nous détaillons dans cet article s'inscrit dans un projet pluridisciplinaire (chimie et informatique), qui a pour but d'automatiser la procédure de recherche et de ciblage de nouvelles molécules basée sur l'analyse de données issues de la spectrométrie de masse en tandem couplée à la chromatographie en phase liquide (LC-MS/MS). Lors d'une telle analyse, le spectromètre de masse tandem sélectionne un ion précurseur avec le premier analyseur, puis cet ion est fragmenté par dissociation induite par collision et les fragments ainsi formés sont analysés afin d'obtenir un spectre MS/MS. Ce dernier est composé uniquement d'ions produits (rapports mass/charge : m/z) par la fragmentation de l'ion précurseur et est donc riche en informations structurales. Les chimistes exploitent ces données pour savoir si une plante contient ou non des molécules non-connues dans la littérature. Pour cela, ils se basent sur la technique des réseaux moléculaires, un graphe où chaque nœud représente un ion issu de la molécule de la plante étudiée (représentée par son spectre MS/MS) et où les arcs représentent la similarité spectrale entre les nœuds. Plus un nœud a un spectre non similaire aux spectres déjà étudiés dans la littérature, et a également des voisins non connus, plus il a des chances de faire référence à une nouvelle molécule. Les chimistes interprètent les métadonnées de ce graphe ainsi que sa topologie afin d'en déduire une décision leur permettant de déployer les efforts de ciblage et d'isolement de molécules originales. Ce graphe n'est pas le seul outil utilisé par les experts mais il est très central dans la procédure de recherche. La première étape de notre projet consistait à définir les critères utilisés par les experts lors de la découverte de nouvelles molécules. Nos entretiens avec nos experts chimistes nous ont permis de définir les critères suivants : la *faisabilité de l'isolement* (certaines molécules peuvent se trouver en quantité très faible, rendant impossible leur analyse), l'*annotation des spectres MS/MS* (il existe différentes bases de données, plus ou moins fiables, contenant des spectres déjà connus), la *topologie des réseaux moléculaires* (le graphe de similarité des spectres plus d'autres informations), la *classe chimique de la molécule* et l'*activité biologique de la molécule*. La *topologie des réseaux moléculaires* est très centrale dans l'analyse des chimistes ; elle donne des informations sur les similarités des spectres des molécules de l'organisme étudié mais aussi indique si le spectre d'une molécule ressemble ou non à d'autres spectres connus dans la littérature. Cette topologie est donc fortement liée à l'indice de similarité choisi par les experts chimistes. Ce choix étant primordial, nous avons voulu tester la fiabilité des résultats de similarité que l'on peut obtenir avec différents indices. La section suivante présente cette étude.

2 Topologie des réseaux moléculaires : indice de similarité

La technique privilégiée de nos experts étant la spectrométrie de masse en tandem, nous nous sommes focalisés sur les indices de similarité communément utilisés pour comparer ce type de données. Ces indices de similarité ont pour but d'**approximer une similarité structurale** entre les molécules car les molécules n'étant pas connues d'avance, il est difficile de connaître leur structure chimique.

Pour définir la qualité des résultats des indices de similarités spectrales, nous avons effectué des tests sur les molécules d'une famille de substances naturelles, les MIA (*Monoterpene Indole Alkaloids*). Les MIA sont des produits naturels végétaux qui comprennent un certain nombre de composés importants sur le plan médicinal. Dans nos tests, nous avons étudié trois indices de similarité spectrales :

— *Cosinus modifié* [1] : chaque spectre est vu comme un vecteur et on calcule le cosinus entre ces vecteurs.
— *Spec2Vec*[2] : indice basé sur l'apprentissage automatique des textes.
— *MS2Deep score*[3] : indice basé sur l'apprentissage profond.

Nous avons comparé les résultats de ces indices à ceux obtenus par **un indice structural**, l'indice de *Morgan/Tanimoto* [4] qui est basé sur une technique de *hashing* de la structure de la molécule.

Nos expériences ont été les premières faites sur cette famille de produits naturels. Elles ont nécessité un travail conséquent de pré-traitement des données sur les spectres MS/MS. Nous avons également ré-entrainé CFM-ID(ref : DOI : 10.1093/nar/gku436), un algorithme d'apprentissage profond de prédiction de données spectrales car ce dernier a été conçu avec des ensembles d'apprentissage contenant des métabolites humains, des molécules très différentes des MIA. Nos résultats montrent entre autre que l'indice de cosinus modifié est celui qui donne les résultats les plus proches de l'indice structurale et un seuil de 0.6 pourra être utilisé pour dire que deux molécules sont similaires. Ces résultats sont très importants pour les chimistes utilisant les spectres de masse pour définir un seuil de similarité structurale.

3 Conclusions et perspectives

La procédure de recherche de nouvelles molécules peut être vue comme un problème multi-critère où chaque critère doit être analysé en profondeur. Dans cet article nous avons présenté rapidement le problème et souligné l'importance des indices de similarités spectrales pour les experts chimistes. Nos expériences nous permettent de faire des recommandation sur le choix de l'indice à utiliser. Nous espérons compléter notre travail en analysant de plus près les quatre autres critères et en proposant des opérateurs d'agrégation qui seront fidèles aux raisonnements des experts chimistes.

Références

[1] G. J. Dolecek, J. R. G. Baez, and M. Laddomada. Design of efficient multiplierless modified cosine-based comb decimation filters : Analysis and implementation. *IEEE Transactions on Circuits and Systems I : Regular Papers*, 64(5) :1051–1063, 2017.

[2] F. Huber, L. Ridder, S. Verhoeven, J. H. Spaaks, F. Diblen, S. Rogers, and J. Van Der Hooft. Spec2vec : Improved mass spectral similarity scoring through learning of structural relationships. *PLoS computational biology*, 17(2) :e1008724, 2021.

[3] F. Huber, S. van der Burg, J. van der Hooft, and L. Ridder. Ms2deepscore : a novel deep learning similarity measure to compare tandem mass spectra. *Journal of cheminformatics*, 13(1) :1–14, 2021.

[4] H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2) :107–113, 1965.