Modélisation de la qualité des séparations énantiomériques utilisant le critère d'information mutuelle

Sónia **Caetano**[1*], Catherine **Krier**[2], Michel **Verleysen**[2], Yvan Vander **Heyden**[1]

[1]FABI, Département de Chimie Analytique et Technologie Pharmaceutique, Vrije Universiteit Brussel- VUB, Laarbeeklaan 103, 1090 Brussels, Belgium
[2]Université Catholique de Louvain, Machine Learning Group, Place du Levant 3, 1348 Louvain-la-Neuve, Belgium
*sonia.caetano@vub.ac.be

MOTS CLÉS: Séparations énantiomériques, critère d'information mutuelle, sélection de variables.

1. Introduction

Les molécules chirales sont très importantes dans le domaine pharmaceutique puisque les énantiomères d'un composé peuvent présenter des différences considérables d'interaction avec d'autres composés. Par conséquent le développement de techniques permettant l'évaluation, l'identification et la séparation des énantiomères d'un composé est très important.

La Chromatographie Liquide de Haute Performance (HPLC) utilisant des phases stationnaires chirales (CSP's) est une des techniques les plus utilisées pour séparer des énantiomères. Pourtant, la CSP adaptée á la séparation chirale (CS) d'une substance doit encore être trouvée par essais successifs, ce qui rend la séparation lente et coûteuse. Ainsi, la prédiction de la capacité d'une certaine CSP à réaliser la CS d'une substance donnée serait très bénéfique.

Le critère d'information mutuelle (MIC) est une mesure statistique de la relation qui existe entre deux variables. Contrairement à d'autres estimateurs, tel que la corrélation par exemple, le MIC ne fait pas de supposition quant à la relation existante entre les variables et, en conséquence, ce critère peut être utilisé dans de nombreux contextes, y compris la sélection de variables.

Cette étude présente le MIC en tant que technique de sélection de variables pour les données de type quantitative structure-property relationship (QSPR). Pour évaluer la performance de ce critère, l'énantiosélectivité de 67 molécules pour trois CSPs différentes a été modélisé. Pour chaque CSP, les descripteurs moléculaires les plus informatifs, 12 au maximum, ont été choisis avec le MIC. Ensuite, des modèles de régression basés sur les descripteurs sélectionnés par le MIC ont été développés à l'aide des Support Vector Machines (SVM) [1]. Le pouvoir de prédiction de chaque modèle a alors été évalué grâce à un jeu de test indépendant et le modèle ayant le meilleur pouvoir de prédiction a été trouvé pour chaque jeu de données. Les variables qui ont été choisies par le MIC pour ces trois modèles ont aussi été utilisées pour construire des modèles *k*-NN et Radial Basis Function Networks (RBFN) [2] afin de démontrer l'applicabilité du MIC comme technique de sélection de variables.

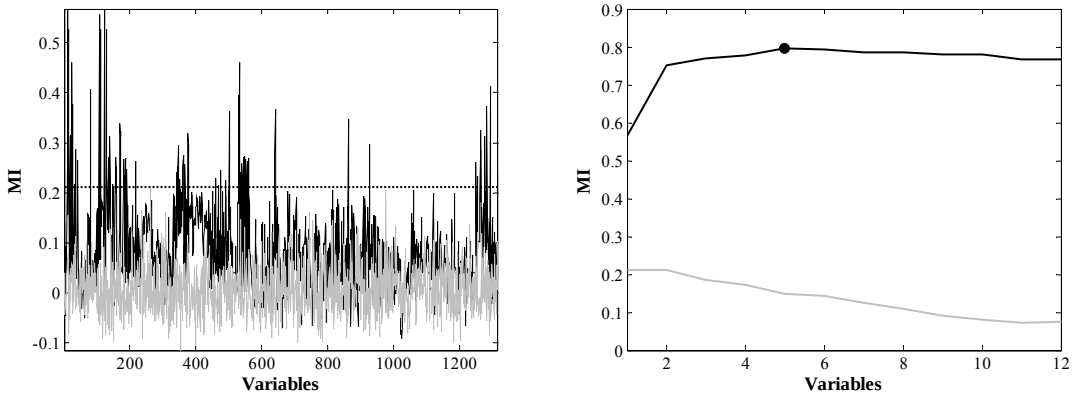
2. Résultats expérimentaux

Les données consistent en 63 hydantoines séparées dans des colonnes HPLC avec trois CSP [3] (jeux des données 1 a 3).
Le calcul des descripteurs est basé sur la structure géométrique de chaque molécule. Ces structures ont été optimisées avec Hyperchem® 6.03 professional software (Hypercube, Gainesville, FL, USA). Les descripteurs (1317 en tout), ont été calculés avec Dragon 5.0 version professionnelle [4].

3. Résultats et discussion

3.1 Sélection des variables : Critère d'information mutuelle (MIC)

Pour modéliser la sélectivité (α) des séparations énantiomèriques des 67 molécules avec les trois CSPs, les procédures ranking et forward pour le MIC ont été appliquées afin de choisir les variables les mieux adaptées. Pour les deux procédures l'influence du nombre de voisins (*k*) dans la sélection des variables a été évaluée. La sélection des variables a été faite en utilisant toutes les données, c'est-à-dire les 1317 descripteurs calculés pour chacune des 67 molécules. Dans les deux procédures *k* pouvait varier entre 1 et 10 tandis que le nombre maximum de variables retenues était de 12. Ce nombre a été choisi de manière à limiter le temps de calcul du MIC ainsi que la complexité des modèles.
La fig. 1 montre, pour le 2éme jeu de données, comment les descripteurs ont étés choisis par chaque procédure. Dans la procédure ranking (fig. 1 (a)) les 12 variables ayant le plus haut MIC sont gardées alors que pour la procédure forward (fig. 1 (b)) le nombre optimal de variables est déterminé par la valeur maximum du MIC. Dans cet exemple, le nombre de variables retenues est égal à cinq. Les tables 1 à 3 représentent le nombre de variables choisies pour chaque jeu des données.



(a) (b)

Fig. 1- Procédures (a) Ranking et (b) Forward pour la détermination du MIC pour le jeu de données 2 en utilisant k=6. Valeurs du MIC calculées respectivement pour des descripteurs moléculaires (ligne noire) et des variables aléatoires (ligne grise); −−− cut-off (valeur maximale du MIC pour les variables aléatoires); • nombre de variables optimal, c'est à dire les descripteurs qui donnent le maximum d'information.

Pour un même jeu de données différentes valeurs de *k* ont été gardées. En effet, vu qu'il n'existe pas une vraie valeur optimale pour *k*, dès que le MIC donne des résultats tels que prévu, c'est-à-dire présentant un profil similaire à ceux montrés dans la fig. 1, il n'y a pas de raison pour choisir une seule valeur de *k*. Par conséquent, et bien que les variables choisies ne soient pas les mêmes, si les résultats du MIC sont similaires pour différentes valeurs de *k*, les différentes sous-groupes de variables sélectionnées sont gardées.

3.2 Modélisation de la sélectivité

Les variables sélectionnées avec le MIC ont été utilisées pour construire des modèles SVM. Pour pouvoir évaluer le pouvoir de prédiction de ces modèles les données ont été divisées en nu jeu de calibration (40 molécules) et un jeu de test (27 molécules). Les résultats du meilleur modèle obtenu pour chaque jeu de données sont résumés en table 1.

Table 1- Résultats du meilleur modèle SVM pour chaque jeu de données

Jeu de données	*k*	Nr. de variables sélectionnées	RMSECV	RMSEP
1	6	12	0.67	0.72
2	6	12	0.85	0.84
3	8	4	0.21	0.35

Les variables utilisées pour ces modèles ont aussi été utilisées pour construire des modèles *k*-NN et RBFN afin de démontrer l'utilité du MIC comme technique de sélection de variables. Les résultats observés sont décrits en tables 2 et 3.

Table 2- Résultats obtenus avec k-NN pour chaque jeu de données.

Jeu de données	*k*	RMSECV	RMSEP
1	6	0.65	1.12
2	4	0.50	1.38
3	3	0.25	0.89

Table 3- Résultats obtenus avec RBFN pour chaque jeu de données.

Jeu de données	RMSECV	RMSEP	Nr. d'unités cachées
1	1.25	0.79	18
2	1.26	1.11	1
3	0.26	0.45	13

Les résultats des trois techniques confirment que c'est pour le 3éme jeu de données que les meilleures prédictions sont obtenues.

4. Conclusion

Les résultats montrent que non seulement le MIC peut être utilisé efficacement pour la sélection de variables, mais aussi que différents sous-groupes de descripteurs avec un pouvoir de prédiction similaire peuvent être trouvés. Ce dernier résultat permet une

sélection intelligente des descripteurs, puisque certains sont plus faciles à obtenir que d'autres.

Bibliographie

[1] N V. Vapnik, *"The Nature of Statistical Learning Theory"*, Springer, New York 1995.
[2] N. Benoudjit, M. Verleysen, Neural Processing Letters 18 (2003) 139– 154.
[3] W. H. Pirkle, M. Ho Hyun, Journal of Chromatography 322 (1985) 295-307.
[4] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, Dragon Professional Version 5.0, Milano Chemometrics and QSAR Research Group, 2004 (software available in: http://www.talete.mi.it/dragon.htm).