MODÈLES QSRR HYBRIDES ALGORITHME GÉNÉTIQUE-RÉGRESSION LINÉAIRE MULTIPLE DES INDICES DE RÉTENTION DE PYRAZINES EN CHROMATOGRAPHIE GAZEUSE

Imen Touhami, Karima Mokrani et Djelloul Messadi
Laboratoire de Sécurité Environnementale et Alimentaire, Université Badji Mokhtar Annaba, B.P. 12, 23000 Annaba, Algérie
d_messadi@yahoo.fr

(Received 8 October 2010 - Accepted 19 September 2011)

RÉSUMÉ

L'approche hybride algorithme génétique/régression linéaire multiple a été appliquée pour modéliser, séparément, les indices de rétention d'un même ensemble de 27 pyrazines éluées tour à tour sur les colonnes OV-101 et Carbowax-20M, en utilisant des descripteurs moléculaires théoriques calculés à l'aide du logiciel DRAGON. Un ensemble de 8 autres pyrazines, séparées dans les mêmes conditions, a servi d'ensemble de test. Pour éviter les modèles présentant des problèmes de colinéarité, et sans réelle capacité de prédiction, nous avons appliqué la règle QUIK basée sur l'indice de corrélation multivariable K. Les modèles optimaux ont été sélectionnés en maximisant le coefficient de prédiction (Q^2_{LOO}). Les descripteurs des modèles obtenus pour les colonnes OV-101 et Carbowax-20M (respectivement 2 et 3) montrent que ce sont les interactions de dispersion et la complexité des molécules qui gouvernent le mécanisme de rétention sur la colonne non polaire, alors que ce sont les interactions spécifiques et la symétrie des molécules qui sont prépondérantes sur la colonne polaire. La pyrazine est un élément de l'ensemble de calibration très influent pour le modèle obtenu sur cette dernière colonne.

Mots-clés: chromatographie gaz-liquide, méthode des phases multiples, modèles QSRR hybrides, descripteurs moléculaires théoriques, validation externe

ABSTRACT

Hybrid genetic algorithm/multiple linear regression approach was applied to model, separately, the retention indices of the same set of 27 pyrazines eluted in turn on OV-101 and Carbowax-20M, using theoretical molecular descriptors derived from DRAGON software. To avoid models with collinearity without prediction power the QUIK procedure based on the K multivariate correlation index was observed. The models proposed here were chosen maximizing the leave-one-out cross-validation squared correlation coefficient (Q^2_{LOO}). A two-dimensional model shows that the dispersion interactions and the complexity of the molecules control the retention mechanism on OV-101, whereas a three – dimensional model

points out the importance of specific interactions and the symmetry of the molecules in predicting retention indices on Carbowax-20M. Pyrazine, a calibration set object, is very influential for the model obtained on Carbowax-20M.

Keywords: gas-liquid chromatography, multiple stationary phases method, hybrid QSRR models, theoretical molecular descriptors, external validation

INTRODUCTION

Les pyrazines, ou 1,4-diazines, sont des hétérocycles azotés très largement distribués dans le règne animal et végétal, et très présents dans l'arôme des aliments (Parliment & Epstein, 1973 ; Masuda *et al*., 1981 ; Barlin, 1982 ; Buchbauer, 2000). Leur identification se fait généralement par chromatographie gazeuse (CG) en comparant leurs pics à ceux obtenus pour les standards des composés suspectés. La disponibilité de tels standards pouvant faire défaut, la recherche d'autres voies d'identification est souhaitable.

La Relation Quantitative Structure/Activité (QSAR) initiée par Hansch et Fujita (1964), a trouvé de nombreuses applications en chimie, en particulier dans la prédiction de la rétention chromatographique (Kaliszan, 1986 ; Kaliszan, 1987 ; Wang *et al*., 1999 ; Lee *et al*., 2004 ; Nacer & Messadi, 2006).

Mihara et Enomoto (1985), ont décrit une relation structure/rétention pour un ensemble de pyrazines substituées pour lesquelles les incréments d'indices relatifs à différents substituants sur le cycle ont été déterminés pour une petite série de substituants présents. La méthode fut ensuite étendue pour intégrer d'autres substituants, et ajouter un terme qui tient compte de la position sur le cycle d'un substituant par rapport aux autres (Mihara & Masuda, 1987). Dans une approche analogue, Masuda et Mihara (1986) décrivent l'utilisation d'indices de connectivité modifiés pour calculer à l'avance les indices de rétention d'une série de pyrazines substituées. Les méthodes conduisent à de bons résultats, pour autant que les incréments d'indices déterminés expérimentalement soient disponibles pour les composés inconnus impliqués, ce qui constitue leur défaut principal.

Stanton et Jurs (1989), ont utilisé la méthodologie QSRR pour développer des modèles reliant les caractéristiques structurales de 107 pyrazines diversement substituées, à leurs indices de rétention obtenus sur deux colonnes de polarités très différentes (OV-101 et Carbowax-20M). Les équations ont été calculées à l'aide de la régression multilinéaire, le choix des variables explicatives (topologiques, électroniques et propriétés physiques) étant réalisé par élimination progressive (Swall & Jurs, 1983), parmi les 85 descripteurs moléculaires individuels obtenus pour chaque molécule entière. Les indices de rétention (Ir) obtenus sur chaque colonne ont été traités séparément, en puisant dans les mêmes ensembles de descripteurs. Les modèles calculés avec 6 variables explicatives fournissent des erreurs standards élevées (S = 23 unités d'indice - u.i.- sur OV-101 et S= 36.33 u.i. sur Carbowax -20 M) qui ne présagent pas de bonnes capacités prédictives pour ces modèles, et qui laissent supposer des relations non linéaires entre descripteurs et propriété (Ir) étudiée.

L'objectif de ce travail vise à utiliser la méthodologie QSRR, dans l'approche algorithme génétique/ régression multilinéaire (AG/RLM), pour modéliser les indices de rétention des (27) pyrazines rapportés par Mihara et Enomoto (1985), les descripteurs moléculaires étant uniquement calculés à partir de la structure chimique des composés. Les

indices de rétention de 8 autres pyrazines, prélevés dans le travail de Mihara et Masuda (1987), serviront d'ensemble de test pour les modèles calculés.

Les hypothèses d'un modèle statistique linéaire à effets fixes, de même que la qualité de l'ajustement, ainsi que la robustesse du modèle, et ses capacités prédictives (interne et externe) seront examinées. Enfin, le domaine d'application (DA) sera discuté à l'aide du diagramme de Williams qui représente les résidus de prédiction standardisés en fonction des valeurs des leviers (hi) (Eriksson *et al*., 2003 ; Tropsha *et al*., 2003).

MÉTHODOLOGIE

Calcul et choix des descripteurs

On a utilisé le logiciel de modélisation moléculaire Hyperchem 6.03, pour représenter les molécules, puis à l'aide de la méthode semi-empirique AM1 (Dewar *et al*., 1985; Holder 1998) obtenir les géométries finales. Il est établi (Levine, 2000) que cette méthode donne de bons résultats quand on traite de petites molécules (de moins de cent atomes), comme celles considérées dans ce travail. Tous les calculs ont été exécutés dans le cadre du formalisme de Hartree-Fock avec contrainte de spin (ou RHF pour restricted Hartree-Fock) sans interaction de configuration (Levine, 2000). Les structures moléculaires ont été optimisées à l'aide de l'algorithme Polak-Ribière avec pour critère une racine du carré moyen du gradient égale à 0,001 kcal.mol^{-1}. Les géométries ainsi optimisées ont été transférées dans le logiciel informatique Dragon version 5.4, pour le calcul de 1664 descripteurs (en plus de ceux, différents, fournis par le logiciel Hyperchem) appartenant à diverses classes. Les descripteurs d'une même classe à valeurs constantes (écarts types inférieurs à 0,0001), et ceux hautement corrélés ($R \geq 0{,}95$), ont été exclus.

En opérant sur les 27 pyrazines d'essai, des sous-ensembles de descripteurs ont été sélectionnés par algorithme génétique, en utilisant le logiciel MobyDigs (Todeschini *et al.,* 2009) et en maximisant le coefficient de prédiction Q^2 .

Pour éviter les modèles présentant des problèmes de colinéarité, et sans réelle capacité de prédiction, on a appliqué la règle QUIK (Todeschini *et al*., 1999), basée sur l'indice de corrélation multivariable K (Todeschini, 1997), défini par :

$$K = \frac{\sum_j \left(\lambda_j \Big/ \sum_j \lambda_j \right) - 1/p}{2(p-1)/p}; \qquad \text{j=1, …, p et } 0 \leq \text{K} \leq 1 \qquad (1)$$

Les λ_j sont les valeurs propres de la matrice de corrélation de l'ensemble de données X (n, p), n étant le nombre d'objets et p le nombre de variables explicatives.

Cette règle est déduite de l'hypothèse, que la corrélation totale dans l'ensemble formé par les prédicteurs X du modèle plus la réponse Y (K_{xy}) doit toujours être plus grande que celle uniquement mesurée dans l'ensemble des prédicteurs (K_x). Le calcul de

K_{xy} est réalisé en considérant la réponse Y comme une variable et en calculant la matrice de corrélation correspondante.

Développement et validation du modèle

Régression linéaire multiple

L'analyse de régression linéaire multiple a été réalisée avec le logiciel MobyDigs, (2009), en utilisant la méthode des moindres carrés ordinaires.

La qualité de l'ajustement a été évaluée par le coefficient de détermination, R^2 , et l'écart quadratique moyen calculé sur l'ensemble de calibration:

$$EQMC = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{2}$$

y_i et $\hat{y}_i$ étant les valeurs observées et calculées de la variable dépendante.

Les techniques de validation croisée ont été appliquées pour l'évaluation de la prédiction interne (Q^2_{LMO} ; bootstrap), et de la robustesse (Q^2_{LOO} ; Y-scrambling) du modèle.

La validation croisée par « leave-one-out » (LOO) (Allen, 1974), consiste à recalculer le modèle sur (n-1) objets, et à utiliser le modèle ainsi obtenu pour prédire la valeur de la variable dépendante du composé écarté. Le procédé est répété pour chacun des n objets de l'ensemble d'essai. La somme des carrés des erreurs de prédiction (désignée par l'acronyme PRESS, pour Predictive Residual Sum of Squares) est une mesure de la dispersion des estimations. On l'utilise pour définir le coefficient de prédiction (Q^2_{LOO}), et l'écart quadratique moyen de prédiction (ou $EQMP$) :

$$Q^2 = 1 - \frac{PRESS}{SCT} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{i/i}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2} \tag{3}$$

$$EQMP = \sqrt{\frac{PRESS}{n}} \tag{4}$$

SCT est la somme totale des carrés; $\hat{y}_{i/i}$ désigne la réponse du i^{ème}objet estimée en utilisant un modèle obtenu sans faire intervenir cet i^{ème}objet, et $\overline{y}$ la valeur moyenne des n observations ; la sommation court sur l'ensemble des composés de calibration.

Une valeur Q^2_{LOO} > 0,5 est considérée comme satisfaisante, une valeur Q^2_{LOO} > 0,9 est excellente (Eriksson *et al.*, 2003).

En fait, si une forte valeur de Q^2_{LOO} est une condition nécessaire d'une possible capacité prédictive élevée d'un modèle, cette condition seule n'est pas suffisante.

Pour éviter une surestimation de la capacité prédictive du modèle on a également appliqué la procédure « leave-more-out » (LMO), répétée 8000 fois, en excluant 50% des objets à chaque étape ($Q^2_{LMO/50}$).

Les modèles QSAR/QSRR, à cause (souvent) de leur complexité et de la sophistication des outils de chimiométrie employés, peuvent constituer une source de corrélations fortuites. Dans le but d'établir que le modèle n'est pas dû au hasard, on a appliqué le test de randomisation de Y (Y-scrambling) (Wold & Eriksson, 1995). Ce test consiste à générer un vecteur de la propriété étudiée par permutation aléatoire des composantes du vecteur réel. On calcule alors sur le vecteur obtenu un modèle QSRR, selon la méthode habituelle. Ce procédé est répété 500 fois dans cette étude.

Dans la technique de validation par bootstrap on simule de nouveaux échantillons de taille (n), par tirages aléatoires avec remise. De cette façon l'ensemble de calibration, qui conserve sa taille initiale (n), se compose, en général, d'objets répétés, l'ensemble de tests rassemblant les objets exclus (Efron & Tibshirani, 1993 ; Wehrens *et al*., 2000). Le modèle est calculé sur l'ensemble de calibration et les réponses prédites pour l'ensemble du test. Tous les carrés des différences entre les valeurs prédites et observées des objets de l'ensemble de tests sont collectés dans le PRESS. Cette procédure de construction des ensembles de calibration et d'évaluation est répétée plusieurs milliers de fois (8000 dans cette étude), les PRESS sont additionnés, et une capacité de prédiction moyenne est calculée (Wehrens *et al.,* 2000).

L'application du modèle, calculé sur l'ensemble de calibration, aux composés de l'ensemble de validation externe permet de vérifier de manière fiable la capacité prédictive du modèle obtenu.

L'équation(5) permet le calcul de Q^2_{EXT} :

$$Q^2_{EXT} = 1 - \frac{\sum_{i=1}^{n_{EXT}} \left(\hat{y}_{i/i} - y_i\right)^2 \Big/ n_{EXT}}{\sum_{i=1}^{n_{EXT}} \left(y_i - \overline{y}_{tr}\right)^2 \Big/ n_{tr}} = 1 - \frac{PRESS/n_{EXT}}{SCT/n_{tr}} \qquad (5)$$

L'indice (EXT) se rapportant aux objets de l'ensemble de validation externe (ou à ceux de l'ensemble d'évaluation obtenu par bootstrap), et l'indice (tr) à ceux de l'ensemble de calibration (training set).

Avec R^2 , le paramètre $EQMP_{EXT}$ est également utile. On le calcule selon :

$$EQMP_{EXT} = \sqrt{\frac{1}{n_{EXT}} \sum_{i=1}^{n_{EXT}} \left(\hat{y}_{i/i} - y_i \right)^2} \qquad (6)$$

La somme portant sur les objets de l'ensemble de validation (n_{EXT}).

Domaine d'application (DA)

Le domaine d'application a été discuté à l'aide du diagramme de Williams (traité en détail dans (Eriksson *et al*., 2003 ; Tropsha *et al*., 2003)), représentant les résidus de prédiction standardisés en fonction des valeurs des leviers h_i . L'équation (7) définit le levier d'un composé dans l'espace original des variables indépendantes (x_i):

$$h_i = \boldsymbol{x_i} \left(\boldsymbol{X^T X} \right)^{-1} \boldsymbol{x_i^T} \quad \text{(i=1,...,n)} \qquad (7)$$

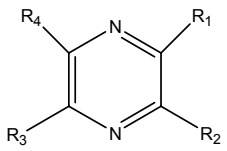
Où $\boldsymbol{x_i}$ est le vecteur ligne des descripteurs du composé i et $\boldsymbol{X}$ (n*p) la matrice du modèle déduite des valeurs des descripteurs de l'ensemble de calibration ; l'indice T désigne le vecteur (ou la matrice) transposé (e).

La valeur critique du levier (h^*) est fixée à (3p+1)/n. Si $h_i < h^*$, la probabilité d'accord entre les valeurs mesurée et prédite du composé i est aussi élevée que celle des composés de calibration. Les composés avec $h_i > h^*$ renforcent le modèle quand ils appartiennent à l'ensemble de calibration, mais auront, sinon, des valeurs prédites douteuses sans pour autant être forcément aberrantes, les résidus pouvant être bas.

DONNÉES EXPÉRIMENTALES

Les indices de rétention de 27 pyrazines chromatographiées en programmation de température, séparément sur les colonnes OV-101 et Carbowax-20M, utilisés comme ensembles d'essai, ont été prélevés dans le travail de Mihara et Enomoto, (1985). De plus, les indices de rétention de 8 autres pyrazines obtenus dans les mêmes conditions par Mihara et Masuda, (1987), ont été utilisés comme ensembles de validation externe.

Les composés impliqués dans cette étude, rassemblés dans *le tableau 1*, présentent la structure générale suivante :



R_1: H, alkyl, alkoxy, méthylthio, acetyl ;
R_2: H, alkyl, alkoxy, acetyl, vinyl;
R_3: H, alkyl;
R_4: H, alkyl.

RÉSULTATS ET DISCUSSION

Développement et validation des modèles

L'optimisation par algorithme génétique conduit à de nombreux modèles de différentes dimensions. Les symboles et la signification des descripteurs optimaux sélectionnés (Tableau1) sont les suivants (Todeschini *et al.*, 2000):

- IC0 : Indice du taux d'information (symétrie de proximité d'ordre zéro).
- ATS1p : Autocorrélation de Broto-Moreau d'une structure topologique de distance 1/ pondérée par les polarisabilités atomiques.
- IVDE : Taux d'information moyen sur l'égalité des degrés des sommets.
- ESpm04d: Moment spectral 04 de la matrice d'adjacence des arêtes pondéré par le moment dipolaire.
- nCconj : Nombre de carbones conjugués (sp2).
-

Les deux premiers permettent de modéliser les indices de rétention obtenus sur la colonne OV-101 (éq. 8), et les trois derniers les indices de rétention calculés sur la colonne Carbowax-20M (éq. 9).

Si les descripteurs IC0 et ATS1p sont très peu corrélés (kx = 4,04), les descripteurs IVDE, ESpm04d et nCconj présentent une certaine colinéarité (kx = 38,71). Cependant, ce qui est le plus important pour ces derniers, c'est que la différence dans la corrélation des variables du bloc $\boldsymbol{X}$ plus la réponse $\boldsymbol{Y}$ (k_{xy}) et celle du bloc $\boldsymbol{X}$ (k_x) est assez grande ($\Delta = k_{xy} - k_x = 15{,}18$) (Tableau 2).

Les modèles basés sur ces descripteurs, calculés en utilisant leurs valeurs centrées réduites, ont pour équations :

Colonne OV-101 :

$$Ir = 1030{,}11\ (\pm 2{,}47)\ IC0 + 126{,}35\ (\pm 2{,}41)\ ATS1p \qquad (8)$$

Colonne Carbowax-20M:

$$Ir = 1438{,}64\ (\pm 3{,}75) - 27{,}72\ (\pm 4{,}98)\ IVDE + 134{,}48\ (\pm 4{,}82)\ ESpm04d + 40{,}31(\pm 3{,}6)nCconj \qquad (9)$$

Ces deux modèles vérifient les hypothèses d'un modèle statistique linéaire à effets fixes. En effet la Figure 1 reproduit les distributions des résidus standards di (rapport résidu ordinaire/racine du carré moyen des écarts) en fonction des valeurs ajustées, qui semblent aléatoires (sans tendances particulières). Cela montre la constance des variances σ^2 , c'est-à-dire leur indépendance des régresseurs et de la variable dépendante ajustée.

La quasi-linéarité (r = 0,9951 ; OV-101 –r = 0,9835; Carbowax-20M – $r_{critique}$= 0,96048) du diagramme des scores normaux (Figure 2) est un indice de normalité. Les valeurs de la statistique de Durbin-Watson (Durbin, & Watson, 1951), [d=2,11 ; OV-101/ d = 1,75 ; Carbowax-20M] sont plus grandes que les valeurs supérieures données par les tables,

respectivement pour 2 et 3 régresseurs, et pour tout risque raisonnable α, ce qui établit à chaque fois l'indépendance des résidus.

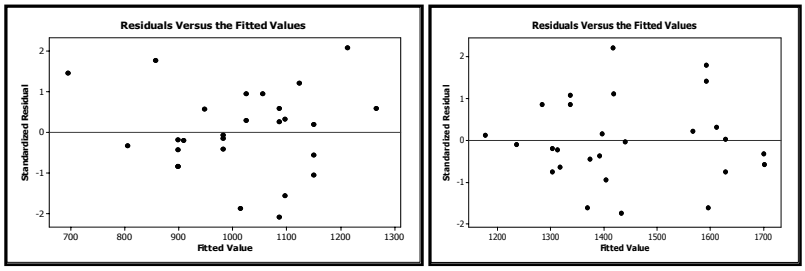


Figure 1. Graphe des résidus standards en fonction des indices de rétention estimés. A gauche: colonne OV-101; à droite: colonne Carbowax-20 M.

TABLEAU 1

Indices de Rétention Observés sur les Deux Colonnes, et Descripteurs Optimaux Sélectionnés

Object	Ir Carb	Ir OV-101	IVDE	ESpm04d	nCconj	IC0	ATS1p
2-Acetyl 3,6-diMePyrazine	1615	1144	1.573	5.635	1	1.572	2.298
2-Acetyl 3-EtPyrazine	1617	1138	1.539	5.555	1	1.572	2.298
2-Methoxy 3-MePyrazine	1339	954	1.436	4.876	0	1.646	2.003
2-EtPyrazine	1300	894	1.061	4.409	0	1.406	2.015
2-Et 5-MePyrazine	1357	980	1.436	4.75	0	1.36	2.14
2-Et 6-MePyrazine	1353	977	1.436	4.75	0	1.36	2.14
2-Ethoxy 3-MePyrazine	1385	1029	1.371	4.942	0	1.578	2.129
2-Et 3-MethioxyPyrazine	1695	1237	1.371	6.151	0	1.578	2.379
2-Et 3-MethoxyPyrazine	1400	1037	1.371	4.963	0	1.578	2.129
2,3diEt 5-MePyrazine	1459	1137	1.539	5.104	0	1.291	2.351
2-VinylPyrazine	1392	907	1.061	4.263	2	1.449	2.015
2-MethoxyPyrazine	1306	877	1.061	4.541	0	1.727	1.858
2-isoPropyl3-MethoxyPyrazine	1400	1078	1.539	5.161	0	1.523	2.242
2-isoPropyl-3-MethioxyPyrazine	1692	1273	1.539	6.216	0	1.523	2.468
Pyrazine	1179	710	0	3.611	0	1.522	1.705
2-MePyrazine	1235	801	1.149	4.254	0	1.46	1.872
2,3-diMePyrazine	1309	897	1.5	4.679	0	1.406	2.015
2,5-diMePyrazine	1290	889	1.5	4.642	0	1.406	2.015
2,6-diMePyrazine	1300	889	1.5	4.642	0	1.406	2.015
2,3,5-triMePyrazine	1366	981	1.585	4.949	0	1.36	2.14
TetraMePyrazine	1439	1067	1.522	5.184	0	1.322	2.251
2-AcetylPyrazine	1571	993	1.436	5.344	1	1.706	2.074
2-Acetyl 5-MePyrazine	1625	1093	1.571	5.492	1	1.634	2.192
2-Acetyl 6-MePyrazine	1618	1089	1.571	5.492	1	1.634	2.192

2-Acetyl 3,5-diMePyrazine	1629	1153	1.573	5.635	1	1.572	2.298
2-Acetyl 3-MePyrazine	1567	1061	1.571	5.508	1	1.634	2.192
2-Ethoxy-3-EtPyrazine	1439	1101	1.309	5.023	0	1.523	2.242
Acetylpyrazine*	1571	993	1.436	5.344	1	1.706	2.074
2,5-diMe-6-ethylpyrazine*	1400	1059	1.571	5.03	0	1.322	2.251
Butylpyrazine*	1474	1088	1.685	5.299	0	1.322	2.251
5-isopropyl-3Me-2-methoxypyrazine*	1467	1170	1.585	5.336	0	1.476	2.343
Ethoxypyrazine*	1348	959	0.986	4.632	0	1.646	2.003
2-ethoxy-3-ethylpyrazine*	1439	1101	1.309	5.023	0	1.523	2.242
2-Me-3-propylpyrazine*	1438	1072	1.571	5.016	0	1.322	2.251
3-Butyl-2,6-diMepyrazine*	1514	1196	1.825	5.602	0	1.264	2.442

* composés de validation externe.

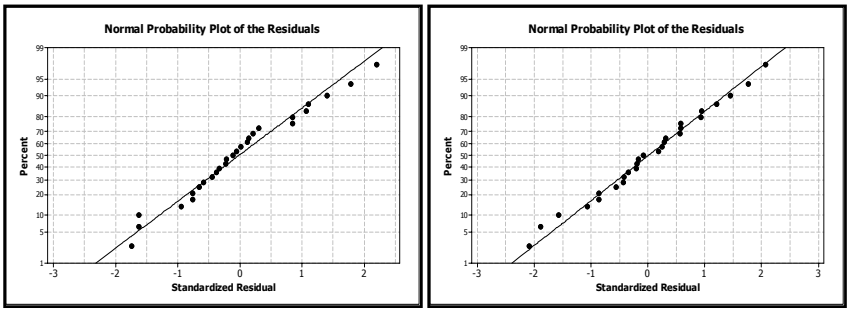


Figure 2. Diagramme des scores normaux. A gauche: colonne OV-101; à droite: colonne Carbowax-20M.

Les diagnostiques statistiques réunis dans le Tableau 2 permettent de faire des comparaisons et de tirer plusieurs conclusions.

TABLEAU 2

Diagnostiques Statistiques pour les Modèles Sélectionnés

Colonne	Descripteurs	n_{tr}	n_{valid}	R^2	Q^2	$Q^2_{LMO/50}$	Q^2_{BOOT}	Q^2_{EXT}
OV-101	IC0 ATS1p	27	8	99,16	98,86	96,35	98,68	98,42
Carbowax-20M	IVDE ESpm04d nCconj	27	8	98,50	98,10	97,88	97,50	98,25
Colonne	R^2_{adj}	*EQMC*	*EQMP*	$EQMP_{EXT}$	k_x	k_{xy}	F	SE
OV-101	99,09	11,80	19,76	16,17	4,04	49,79	1417,57	12,52
Carbowax-20M	98,31	17,80	20,03	19,24	38,71	53,89	504,53	19,28

Les valeurs de R^2 et de R^2_{adj} montrent, à chaque fois, la qualité de l'ajustement, alors que les très faibles différences entre R^2 et Q^2 renseignent sur la robustesse des modèles qui sont, en outre, très hautement significatifs (valeurs élevées de la statistique F de Fisher). De plus, la similitude de $EQMC$ et $EQMP$ signifie que les capacités de prédiction internes des modèles ne sont pas trop dissemblables de leurs pouvoirs d'ajustement.

Les très faibles écarts entre Q^2 et $Q^2_{LMO/50}$ démontrent la bonne stabilité dans la validation interne, et la validation par bootstrap (Q^2_{BOOT}) confirme tout à la fois la capacité de prédiction interne et la stabilité des modèles.

La validation statistique externe (Q^2_{EXT} ; $EQMP_{EXT}$) atteste de la bonne capacité prédictive des composés n'ayant pas participé au calcul des modèles, mais qui appartiennent cependant au domaine chimique de l'ensemble d'essai (voir à la suite : domaine d'application). Il est à noter que le modèle d'équation (9) est à accréditer des meilleures performances ($Q^2_{EXT} > Q^2$; $EQMP_{EXT} < EQMP$).

La Figure 3 qui représente le graphe des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés (carrés) au modèle de départ (étoile). Il est clair que les statistiques obtenues pour les vecteurs modifiés des indices de rétention sont plus petites que celles des modèles réels. Ceci permet d'assurer que des relations structure/rétention réelles ont été établies pour chacune des colonnes.

Domaine d'application

Comme on peut le voir sur les diagrammes de Williams (Figure 4), les valeurs hi de tous les composés de calibration et de test séparés sur la colonne OV-101 sont inférieures à la valeur critique ($h^* = 0{,}333$) et aucun des ces composés n'est influent.

Par contre, pour les composés séparés sur la colonne Carbowax-20M, la 2-Vinylpyrazine (seul composé comportant un substituant vinyl) et la pyrazine (seul composé ne comportant pas de substituant) possèdent des $h_i > h^* = 0{,}444$.

Dans les deux cas, tous les composés de calibration et de test présentent des résidus de prédiction standardisés inférieurs, en valeur absolue, à 3 unités d'écart type (3 σ), ce qui montre qu'il n'y a pas de données aberrantes pour les deux modèles QSRR développés.

Interprétation des modèles

Sur la colonne OV-101, le descripteur ATS1p qui est très corrélé avec Ir, régit notablement celui-ci, comme le montrent les valeurs de son coefficient dans le modèle (éq. 8) et du t de Student associé (Tableau 3). Ce descripteur AUTO 2D trouve son origine dans l'auto corrélation de la structure topologique de Moreau *et al.* (1984) ; sa valeur est égale à la somme de toutes les contributions des fragments structurels de distance l, réunies dans la

matrice de distance topologique. En même temps, ce descripteur indique le rôle décisif de la polarisabilité dans le phénomène de rétention. En effet, la polarisabilité peut être reliée à la capacité du soluté d'entreprendre des interactions de dispersion avec les constituants de la phase stationnaire OV-101. L'indice d'information IC0 (Magnuson *et al*., 1983), insensible à la géométrie moléculaire, représente une mesure de la complexité structurelle par atome ; sa contribution négative à la rétention est faible.

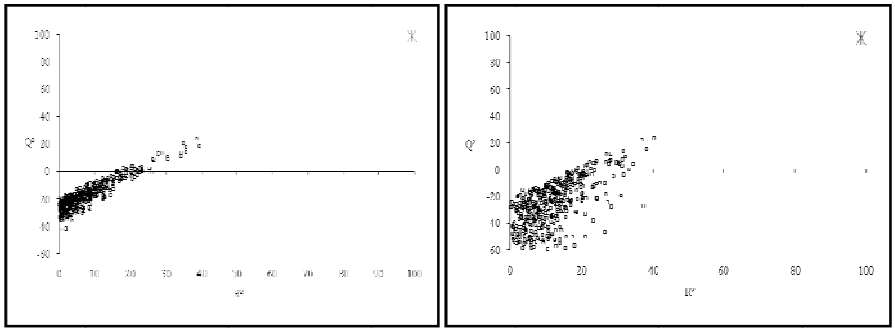


Figure 3. Tests de randomisation associés aux modèles QSRR. A gauche: colonne OV-101; à droite: colonne Carbowax-20 M. Les carrés correspondent aux propriétés ordonnées aléatoirement, et les étoiles aux propriétés réelles.

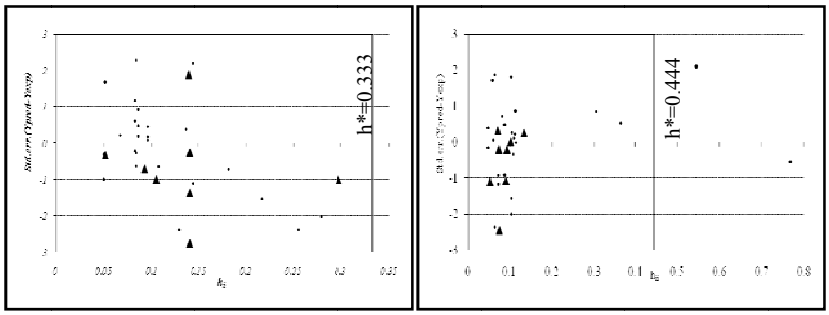


Figure 4. Diagrammes de Williams des résidus de prédiction standardisés en fonction des leviers. A gauche: colonne OV-101; à droite: colonne Carbowax-20M. • : Calibration. ▲: Validation externe.

Les moments spectraux de la matrice d'adjacence des arêtes (Estrada & Gutsman, 1996), permettent de relier les propriétés physiques (et biologiques) des molécules directement à leurs composantes structurelles (Estrada *et al*., 1998). Le moment spectral d'ordre 4, ESpm04d, défini par la trace de la puissance quatrième de la matrice d'adjacence des arêtes, est le plus corrélé avec les indices de rétention obtenus sur la colonne polaire Carbowax-20M ; c'est aussi le descripteur dont le coefficient est le plus élevé dans l'équation (9), et le plus significatif (t = 27,92). En outre, sa pondération par le moment dipolaire met en relief le rôle important des interactions spécifiques soluté-phase stationnaire Carbowax-20M, dans la rétention chromatographique. Le nombre de carbones aromatiques conjugués

(nCconj), renseigne sur la densité en électrons π des composés aromatiques qui constituent un bon groupe donneur de protons.

TABLEAU 3

Importance des Prédicteurs (*X*) dans les Modèles; r(*X*, Ir) est le Coefficient de Corrélation Prédicteur-Indice de Rétention

Colonne	***X***	r(***X***, Ir)	Coefficient	t	p
OV-101	Cste	-	1090	-21,32	0,000
	IC0	-0,065	32	11,65	0,000
	ATS1p	0,970	126	52,38	0,000
Carbowax-20M	Cste	-	1439	7,51	0,000
	IVDE	0,524	-28	-5,57	0,000
	ESpm04d	0,910	134	27,92	0,000
	nCconj	0,478	40	11,20	0,000

Ainsi, le descripteur nCconj indique (encore) le rôle des interactions spécifiques, par établissement de liaisons hydrogène, dans la rétention sur la colonne polaire Carbowax-20M. Le taux d'information moyen sur l'égalité des degrés des atomes, IVDE (Bonchev, 1983), c'est-à-dire du nombre d'atomes, autres que H, liés à l'atome considéré, renseigne sur le niveau de symétrie de la structure moléculaire; la symétrie, avec la taille et la forme des molécules, peut jouer un rôle non négligeable dans le processus de distribution du soluté entre les 2 phases (mobile/stationnaire) chromatographiques. Sa contribution négative à la rétention est faible.

CONCLUSION

Les indices de rétention de 27 pyrazines éluées sur 2 colonnes de polarités très différentes ont été corrélés avec 2 (colonne OV-101) et 3 (colonne Carbowax-20M) descripteurs théoriques calculés uniquement à partir de la structure des molécules, et sélectionnés par algorithme génétique parmi plus de 1600 descripteurs moléculaires obtenus à l'aide du logiciel DRAGON. Les indices de rétention de 8 autres pyrazines, séparées dans les mêmes conditions, ont été choisis pour former les ensembles de validation externe pour les modèles calculés.

Les modèles QSRR présentés sont robustes, avec de bonnes capacités prédictives internes et externes, et une bonne qualité de l'ajustement. Pour le modèle de rétention sur la colonne très polaire Carbowax-20M, la pyrazine, seul élément non substitué de l'ensemble de calibration, est un composé très influent de l'ensemble de calibration.

Sur la colonne non polaire OV-101 ce sont les interactions de dispersion et la complexité des molécules qui gouvernent la rétention, alors que sur la colonne Carbowax-20M ce sont les interactions spécifiques et la symétrie des molécules qui s'imposent.

RÉFÉRENCES

Allen, D.M. 1974. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16: 125-127.

Barlin, G.B. 1982. *The pyrazines*. Wiley-Interscience, New York.

Bonchev, D. 1983. *Information theoretic indices for characterization of chemical structures*. Research Studies Press, Chichester (UK).

Buchbauer, G. 2000. *Treshold-based structure-activity relationships of pyrazines with bell-pepper flavor*.

Dewar, M.J.S., Zoebisch, E.G., Ealy, E.F., Stewart, J.J.P. 1985. AM1: a new general purpose quantum mechanical model. *J. Am.Chem. Soc.*, 107: 3902-3909.

Durbin, J., Watson, G.S. 1951. Testing for serial correlation in least squares regression. II. *Biometrika,* 38(1-2): 159-178.

Dragon 5.4, http://www.disat.unimib.it

Efron, B., Tibshirani, R.J. 1993. *An introduction to the bootstrap*. Chapman & Hall.

Eriksson, L., Jaworska, J., Worth, A., Cronin, M., Mc Dowell, R.M., Gramatica, P. 2003. Methods for reliability, uncertainty assessment, and applicability evaluations of regression based and classification QSARs. *Environ. Health Perspect*., 111(10): 1361-1375.

Estrada, E., Gutsman, I. 1996. A topological index based on distances of edges of molecular graphs. *J. Chem. Inf. Comput. Sci*., 36: 850-853.

Estrada, E., Peňa, A., Garcia-Domenech, R. 1998. Designing sedative/hypnotic compounds from a novel substructural graph–theoretical approach. *J. Chem. Inf. Comput. Aid. Molec. Des*., 12: 583-595.

Hansch, C., Fujita, T. 1964. ρ-σ-π analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc*., 86: 1616-1662.

Holder, A.J. 1998. AM1, *Encyclopedia of Computational Chemistry*, P.V.R. Scheleyer, N.L. Allinger, T. Clarck, J. Gasteiger, P.A. Kollman, H.F Schaefer, III and P.R. Schreiner (Eds), Wiley, Chichester, 1, 8.

Hyperchem 6.03, (Hypercube), http://www.hyper.com.

Kaliszan, R. 1986. Quantitative relationships between molecular structure and chromatographic retention. *CRC Crit. Rev. Anal. Chem*., 16: 323-383.

Kaliszan, R. 1987. *Quantitative structure-chromatographic retention relationships.* J. Wiley, New York.

Lee, Seung Ki., Polyakova, Yulia., Row, Kyung Ho. 2004. Evaluation of predictive retention factors for phenolic compounds with QSPR equations. *J. Liq. Chromatogr. and Rel. Tech*., 27(4): 629-639.

Levine, I.N. 2000. *Quantum chemistry*. 5th ed., New Jersey, Prentice-Hall.

Magnuson, V.R., Harriss, D.K., Basak, S.C. 1983. Topological indices based on neighbor symmetry: chemical and biological applications. In: *Chemical Applications of Topology and Graph Theory.,* R.B. King, ed., Elsevier, Amsterdam., 178-191.

Masuda, H., Misaku, Y., Shibamoto, T. 1981. Synthesis of new pyrazines for flavor use. *J. Agric. Food Chem.*, 29: 944-947.

Masuda, H., Mihara, S. 1986. Use of modified molecular connectivity indices to predict retention indices of monosubstituted alkyl, alkoxy, alkylthio, phenoxy and (phenylthio) pyrazines. *J. Chromatogr*., 366: 373-377.

Mihara, S., Enomoto, N. 1985. Calculation of retention indices of pyrazines on the basis of molecular structure. *J. Chromatogr.,* 324: 428-430.

Mihara, S., Masuda, H. 1987. Correlation between molecular structures and retention indices of pyrazines. *J. Chromatogr*., 402:309-317.

MobyDigs 1.1, http://www.disat.unimib.it

Moreau, G., Broto, P., Vandycke, C. 1984. Molecular Structures: Perception. Autocorrelation descriptor and SAR studies. *Eur. J. Med. Chem*., 19: 66-70.

Nacer, H., Messadi, D. 2006. Relation structure-rétention chromatographique pour les phenols. *Rev. COST*., 4: 81-87.

Parliment, T.H., Epstein, M.F. 1973. Organoleptic properties of some alkyl-substituted alkoxy- and alkylthiopyrazines. *J. Agric. Food Chem*., 21: 714-716.

Stanton, D.T., Jurs, P.C. 1989. Computer-assisted prediction of gas chromatographic retention indexes of pyrazines. *Anal. Chem*., 61: 1328-1332.

Swall, G.W., Jurs, P.C. 1983. Interactive computer system for the simulation of carbon-13 nuclear magnetic resonance spectra. *Anal. Chem*., 55: 1121-1127.

Todeschini, R. 1997. Data correlation, number of significant principal components and shape of molecules. The K correlation index. *Anal. Chim. Acta*, 348: 419-430.

Todeschini, R., Consonni, V., Maiocchi, A. 1999. The K correlation index: theory development and its applications in chemometrics. *Chemom. Intell. Lab. Syst*., 46 : 13-29.

Todeschini, R., Consonni,V., Mannhold, R. 2000. *Handbook of molecular descriptors*. Eds. Kubinyi, H., Timmerman, H Wiley-VCH Verlag GmbH, Weinheim.

Todeschini , R., Ballabio, D., Consonni, V., Mauri, A., Pavan, V. 2009. MobyDigs 1.1, Copyright TALETE srl.

Tropsha, A., Gramatica, P., Grombar, V.K. 2003. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combi. Sci*., 22: 69-76.

Wang, Q.S., Zhang, L.M., Zhang, X.D., Xing, G., Tang, Z. 1999. A system for predicting the retentions of O-alkyl, n-(1-methylthioethylideneamino) phosphoramidates on RP-HP. *Chromatographia*, 49(7/8): 444-448.

Wehrens, R., Putter, H., Buydens, L.M.C. 2000. The bootstrap: a tutorial. *Chemom. Intell. Lab. Syst.*, 54(1): 35-52.

Wold, S., Eriksson, L. 1995. Statistical validation of QSAR results. *In:* H. Van de Waterbeemd ed. Chemometrics methods in molecular design. VCH, New York., 2: 309-318.