Análisis QSPR para índices de retención medidos en la columna OV-101

Cristian Rojas[a,*], Pablo R. Duchowicz[a], Piercosimo Tripaldi[b] y Reinaldo Pis Diez[c]

[a] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (CCT La Plata-CONICET, UNLP), Diag. 113 y 64, C.C. 16, Sucursal 4, 1900 La Plata, Argentina
[b] Laboratorio UDALAB, Facultad de Ciencia y Tecnología, Universidad del Azuay, Av. 24 de Mayo 7-77 y Hernán Malo. Cuenca, Ecuador
[c] CEQUINOR, Centro de Química Inorgánica (CONICET, UNLP), Departamento de Química, Facultad de Ciencias Exactas, UNLP, C.C. 962, 1900 La Plata, Argentina

*Correspondencia: crojasvilla@gmail.com

Resumen - En el presente trabajo se ha desarrollado una relación cuantitativa estructura-propiedad (QSPR) predictiva para modelar el índice de retención medido en la columna capilar OV-101, usando un conjunto de 1208 compuestos aromáticos. Se ha procedido en dos etapas: en la primera se han considerado todos los bloques de descriptores, mientras que en la segunda se usaron únicamente descriptores topológicos. Los modelos obtenidos han sido apropiadamente validados y los resultados obtenidos muestran claramente que los descriptores 3D no aportan información relevante para los modelos; mientras que, el índice de conectividad de solvatación de primer orden posee gran relevancia para este propósito.

1. Introducción

En 1977 aparecen por primera vez tres publicaciones sobre la teoría QSPR, o lo que actualmente se conoce como relación cuantitativa estructura-retención (QSRR) [1], que es muy útil para la acertada predicción de los índices de retención (*RI*) [2, 3]. Desde entonces, varios estudios QSPR han sido publicados para predecir el *RI* de compuestos volátiles [4-11]. Dados estos antecedentes, el objetivo principal del presente estudio es desarrollar un modelo predictivo QSPR para datos con *RI* medidos en la columna OV-101.

2. Materiales y Métodos

El conjunto de datos analizado involucró 1208 sustancias aromáticas cuyo *RI* se midió en la columna estacionaria no-polar OV-101 [12]. Las geometrías moleculares de dichas sustancias fueron optimizadas en el programa HyperChem [13] por el método de los campos de fuerza de mecánica molecular (MM+) seguidas por el método semiempírico PM3. Posteriormente, el programa Dragon [14] se usó para calcular los descriptores [15]. El Método de Reemplazo (RM) [16, 17] se usó como herramienta de selección de variables; mientras que la función de utilidad [18] implementada en DART [19] se usó para seleccionar el mejor modelo. Para la validación se dividió el conjunto en tres grupos [20] usando una metodología basada en k-medias [21]. Se aplicó la validación cruzada de dejar uno fuera (LOO), dejar varios fuera (LMO) y la randomización-Y [22]. El dominio de aplicabilidad (AD) [23] permitió verificar los

compuestos considerados extrapolación. Finalmente, se estandarizaron los coeficientes (b_j^s) para medir el grado de contribución de cada descriptor [24].

3. Resultados y Discusión

De la partición k-medias se obtuvieron tres grupos: $N_{train} = 400$, $N_{val} = 405$, y $N_{test} = 403$. A continuación, el RM exploró conjuntos conteniendo a) 2895 descriptores moleculares de todas las familias y b) 1815 descriptores topológicos. Los parámetros $RMSD_{train}$ y $RMSD_{test}$ no muestran una variación significativa entre modelos de una misma dimensión (d).

Tabla 1. Mejores modelos QSPR obtenidos considerando todas las familias de descriptores.

d	R^2_{train}	$RMSD_{train}$	R^2_{val}	$RMSD_{val}$	$R^2_{ij\max}$	U	Descriptores moleculares
1	0.87	124.38	0.89	107.12	0.00	0.167	*X1sol*
2	0.89	112.46	0.92	91.93	0.00	0.470	*Chi0_EA, GATS1p*
3	0.91	101.71	0.93	83.58	0.88	0.568	*nHDon, RDF010e, Sp*
4	0.92	97.75	0.94	79.65	0.88	0.721	*nHDon, DP02, RDF010e, Sp*
5	**0.92**	**95.44**	**0.94**	**78.31**	**0.32**	**0.784**	***PDI, Hy, ATSC2s, EE_B(s), X1sol***
6	0.93	91.47	0.95	76.19	0.76	0.682	*H-050, R1s+, Mor05p, RDF010s, ATSC2s, X1sol*
7	0.93	89.07	0.94	76.94	0.76	0.683	*H-050, nCconj, R2s+, Mor05p, RDF010s, ATSC2s, X1sol*

Tabla 2. Mejores modelos QSPR obtenidos considerando descriptores topológicos

d	R^2_{train}	$RMSD_{train}$	R^2_{val}	$RMSD_{val}$	$R^2_{ij\max}$	U	Descriptores moleculares
1	0.87	124.38	0.89	107.12	0.00	0.167	*X1sol*
2	0.89	112.46	0.92	91.93	0.00	0.489	*Chi0_EA, GATS1p*
3	0.91	105.09	0.93	83.80	0.19	0.699	*PDI, Hy, X1sol*
4	**0.91**	**100.94**	**0.93**	**82.99**	**0.17**	**0.806**	***PDI, H-050, SpMax1_Bh(s), X1sol***
5	0.92	96.43	0.94	82.74	0.38	0.770	*PDI, O-058, H-050, ATSC4s, X1sol*
6	0.93	94.00	0.93	85.15	0.76	0.602	*O-058, H-050, C-044, ATSC4e, H_Dt, X1sol*
7	0.93	91.09	0.94	80.11	0.32	0.763	*PDI, Hy, C-044, C-033, ATSC2s, EE_B(s), X1sol*

Para calcular la utilidad (U) se usó una función lineal para representar R^2_{train} y R^2_{val}; mientras que, $RMSD_{train}$, $RMSD_{val}$ y $R^2_{ij\max}$ fueron moduladas por una función lineal inversa y finalmente se consideró una función normal para el número de descriptores d. El mejor modelo en cada tabla se indica en negrita y se observa claramente que los descriptores 3D podrían omitirse, con lo que el modelo QSPR es:

$$IR = -1104.8 + 169.3\,X1sol + 26.0\,SpMax1\_Bh(s) + 136.5\,H\text{-}050 + 1370.2\,PDI \qquad (1)$$

$N_{train} = 400$, $d = 4$, $R^2_{train} = 0.914$, $S_{train} = 100.9$, $F = 1049$, $R^2_{ij\,max} = 0.172$

$o(3S) = 5$, $R^2_{loo} = 0.912$, $S_{loo} = 102.3$, $R^2_{l20\%o} = 0.907$, $S_{l20\%o} = 105.0$

$N_{val} = 405$, $R^2_{val} = 0.935$, $S_{val} = 83.0$

$N_{test} = 403$, $R^2_{test} = 0.927$, $S_{test} = 78.6$

$S_{train} < S^{rand} = 335.1$

La Fig. 1 muestra que existe una tendencia lineal de los puntos, indicando que un modelo de regresión estable se ha alcanzado.

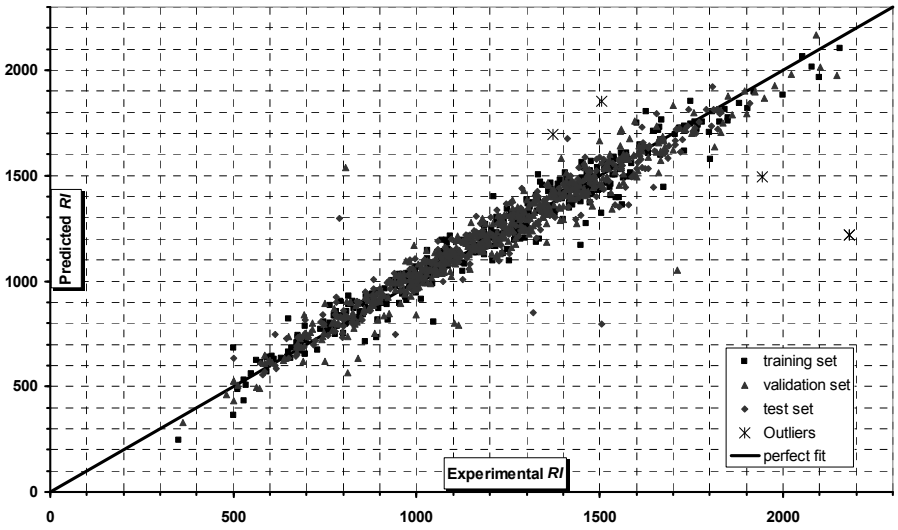

Fig. 1. *IR* experimental versus predicho de acuerdo con el modelo QSPR.

El índice de conectividad $X1sol$ [25] tiene la mayor importancia y gobierna el *RI* en cromatografía para columnas apolares. La tabla 4 muestra que esta base de datos es 4.1 veces mayor que la más grande estudiada [11]. Asimismo, los resultados son malos cuando se usan compuestos con estructuras químicas diversas [10, 11] y finalmente, tres modelos QSRR no realizan validación externa [4, 6, 26].

Tabla 4. Comparación de varios modelos QSPR

Referencia	Familia de compuestos	Num. de compuestos	Num. de descriptores	R^2_{train}	$RMSD_{train}$	R^2_{test}	$RMSD_{test}$
[10]	Pirazinas sustituidas	107	6	0.994	21.4	0.992	32.4
[4]	Alcoholes, Aldehídos, Cetonas y Esteres	115	7	0.998	11.1	-- [a]	--
[26]	Alquilbencenos	39	7	1.000	3.4	0.999	--
	Alquilnaftalenos	15	5	1.000	1.7	0.999	--
	Alquil aril carbamatos (Grupo 1)	27 para los tres grupos	3	1.000	1.4	--	--
	Alquil aril carbamatos (Grupo 2)		4	1.000	0.6	--	--
	Alquil aril carbamatos (Grupo 3)		4	0.994	7.0	--	--
[6]	Fragancias	91	5	0.994	--	--	--
[11]	Sabores	297	4	0.961	59.6	0.959	58.0
Este trabajo	Fragancias y sabores	1208	4	0.914	100.9	0.927	78.6

El AD refleja que catorce compuestos son predicciones no fiables a partir del modelo, adicionales a los 5 datos atípicos (Fig. 1). Se atribuye este comportamiento a la complejidad en términos de heterogeneidad de la base de datos.

4. Conclusiones

Se ha desarrollado un modelo QSPR para la columna apolar OV-101 con aceptable capacidad predictiva. Adicionalmente, la función de utilidad demostró ser una herramienta útil para seleccionar el mejor modelo entre siete diferentes. Por otro lado, se demostró que los descriptores 3D no mejoran el modelo y qué, el índice de conectividad de primer orden está fuertemente correlacionado con el *IR*.

Agradecimientos

Cristian Rojas agradece la beca doctoral brindada por la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT) de la República del Ecuador. Pablo R. Duchowicz agradece el apoyo financiero brindado por el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), proyecto PIP11220100100151 y al Ministerio de Ciencia, Tecnología e Innovación Productiva por permitir el uso de la biblioteca digital. Pablo R. Duchowicz y Reinaldo Pis Diez son miembros de la Carrera de Investigador Científico del CONICET.

Referencias

[1] R. Kaliszan, QSRR: Quantitative Structure-(Chromatographic) Retention Relationships, Chem. Rev., 107 (2007) 3212-3246.

[2] K. Héberger, Quantitative structure-(chromatographic) retention relationships, J. Chromatogr. A, 1158 (2007) 273-305.

[3] Q.S. Wang, L. Zhang, M. Zhang, X.D. Xing, G.Z. Tang, A system for predicting the retentions of O-alkyl, n-(1-methylthioethylideneamino) phosphoramidates on RP-HPLC, Chromatographia, 49 (1999) 444-448.

[4] L.S. Anker, P.C. Jurs, P.A. Edwards, Quantitative structure-retention relationship studies of odor-active aliphatic compounds with oxygen-containing functional groups, Anal. Chem., 62 (1990) 2676-2684.

[5] H. Dua, J. Wanga, Z. Hua, X. Yao, Quantitative Structure-Retention Relationship study of the constituents of saffron aroma in SPME-GC–MS based on the Projection Pursuit Regression method, Talanta, 77 (2008) 360-365.

[6] K.L. Goodner, Practical retention index models of OV-101, DB-1, DB-5, and DB-Wax for flavor and fragrance compounds, LWT-Food Sci. Technol., 41 (2008) 951-958.

[7] K. Héberger, M. Görgényi, M. Sjöström, Partial least squares modeling of retention data of oxo compounds in gas chromatography, Chromatographia, 51 (2000) 595-600.

[8] S. Liu, C. Yin, S. Cai, Z. Li, Molecular structural vector description and retention index of polycyclic aromatic hydrocarbons, Chemom. Intell. Lab. Syst., 61 (2002) 3-15.

[9] C. Lu, A. Jalbout, L. Adamowicz, Y. Wang, C. Yin, QSRR Study for Gas and Liquid Chromatographic Retention Indices of Polyhalogenated Biphenyls Using Two 2D Descriptors, Chromatographia, 66 (2007) 717-724.

[10] D.T. Stanton, P.C. Jurs, Computer-assisted prediction of gas chromatographic retention indexes of pyrazines, Anal. Chem., 61 (1989) 1328-1332.

[11] J. Yan, D.-S. Cao, F.-Q. Guo, L.-X. Zhang, M. He, J.-H. Huang, Q.-S. Xu, Y.-Z. Liang, Comparison of quantitative structure-retention relationship models on four stationary phases with different polarity for a diverse set of flavor compounds, J. Chromatogr. A, 1223 (2012) 118-125.

[12] W. Jennings, T. Shibamoto, Qualitative Analysis of Flavor and Fragrance Volatiles by Glass Capillary Gas Chromatography, ACADEMIC PRESS, INC, London, 1980.

[13] HyperChem, Hypercube Inc., http://www.hyper.com.

[14] Dragon, Software for Molecular Descriptor Calculation, TALETE, srl., http://www.talete.mi.it/, 2014.

[15] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, WILEY-VCH, Weinheim, 2009.

[16] P.R. Duchowicz, E.A. Castro, F.M. Fernández, Alternative Algorithm for the Search of an Optimal Set of Descriptors in QSAR-QSPR Studies, MATCH Commun. Math. Comput. Chem., 55 (2006) 179-192.

[17] P.R. Duchowicz, E.A. Castro, F.M. Fernández, M.P. González, A New Search Algorithm of QSPR/QSAR Theories: Normal Boiling Points of Some Organic Molecules, Chem. Phys. Lett., 412 (2005) 376-380.

[18] M. Pavan, R. Todeschini, Total Order Ranking Methods, in: M. Pavan, R. Todeschini (Eds.) Scientific Data Ranking Methods: Theory and Applications, Elsevier2008, pp. 51-72.
[19] DART, Decision Analysis by Ranking Techniques, TALETE, srl., http://www.talete.mi.it/, 2007.
[20] A. Miller, Subset selection in regression, CRC Press2012.
[21] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, New York, 2005.
[22] C. Rücker, G. Rücker, M. Meringer, Y-Randomization and its variants in QSPR/QSAR, J. Chem. Inf. Model., 47 (2007) 2345-2357.
[23] L. Eriksson, J. Jaworska, A.P. Worth, M.T. Cronin, R.M. McDowell, P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs, Environmental health perspectives, 111 (2003) 1361-1375.
[24] N.R. Draper, H. Smith, Applied Regression Analysis, New York, 1981.
[25] N.S. Zefirov, V.A. Palyulin, QSAR for Boiling Points of "Small" Sulfides. Are the "High-Quality Structure-Property-Activity Regressions" the Real High Quality QSAR Models?, J. Chem. Inform. Comput. Sci., 41 (2001) 1022-1027.
[26] V.A. Gerasimenko, V.M. Nabivach, Relationships between gas chromatographic retention indices and molecular structure of aromatic hydrocarbons, J. Chromatogr. A 498 (1990) 357-366.