FUNCIONES DE DESEABILIDAD COMO FUNCIÓN DE PUNTUACIÓN (FITNESS FUNCTION) EN LA UTILIZACIÓN DE LOS ALGORITMOS GENÉTICOS PARA LA SELECCIÓN DE VARIABLES

Piercosimo Tripaldi[a,*] y Cristian Rojas[b]

[a] Laboratorio UDALAB, Facultad de Ciencia y Tecnología, Universidad del Azuay, Av. 24 de Mayo 7-77 y Hernán Malo. Cuenca, Ecuador
[b] Cátedra de Química Teórica y Computacional, Departamento de Química, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Calle 115 y 47, 1900 La Plata, Argentina
*Correspondencia. tripaldi@uazuay.edu.ec

Resumen

Introducción

En la teoría QSAR/QSPR es común enfrentarse con un elevado número de variables que, en muchas ocasiones supera con creces al número de datos experimentales, por lo que es obligatorio buscar reducir a unas pocas variables útiles. Este proceso puede enfrentarse con diferentes estrategias: análisis de todos los subconjuntos posibles de variables (ASM), búsqueda secuencial ó método de reemplazo (RM) con sus variaciones (3,4), la selección mediante técnicas de bandadas de partículas (PSO) (5) y algoritmos genéticos (GAs) (6,7,8,9), entre otros.

En GAs como en PSO se retienen las variables que producen la mejor puntuación, la cuál es sustancialmente un criterio de decisión. En este trabajo se ha utilizado el método de las funciones de deseabilidad (10,11) para optimizar dicha puntuación en la selección de variables mediante algoritmos genéticos. Las funciones consideradas para los criterios son:

1) $R^2_{cv}\,LMO$ (función creciente lineal con peso 0.90)
2) 10 descriptores máximos (función lineal decreciente con peso 0.05)
3) Razón $K_{xy} / K_x > 2$ (función lineal creciente con peso 0.05)

La función de deseabilidad total D se ha calculado como media geométrica ponderada de los tres valores obtenidos para cada modelo.

Materiales y métodos

Se ha tomado el índice de retención *IR* de Kovats (12) de 1234 sustancias volátiles reportados por Jenning y Shibamoto (13) para la columna OV-101 (0.28 mm x 50 m) con 1% de carbowax. Las estructuras se han optimizado geométricamente en Hyperchem (14) con mecánica molecular MM+ y luego refinado con el método semiempírico PM3, aplicando el algoritmo de Polak-Ribiere hasta que la desviación de la raíz cuadrada sea menor a $0.01\text{kcal.(Å.mol)}^{-1}$. Para cada molécula se han calculado 4885 descriptores utilizando Dragon 6.0 (15). La selección de los descriptores óptimos se ha efectuado aplicando el método GAs bajo la modalidad implementada por Leardi et al (16). Este procedimiento ha sido modificado para utilizar en vez de puntuación original la función D.

Para validación se ha descompuesto aleatoriamente el conjunto de datos en dos subconjuntos: entrenamiento (50%) para construir el modelo con mayor D y posteriormente validarlo sobre el segundo subconjunto (validación externa) calculando el R^2_{pred}. Este proceso se ha repetido tres veces. Posteriormente, se ha dividido el conjunto en entrenamiento (60%) y validación (40%), considerando únicamente los descriptores seleccionados en las tres repeticiones para obtener el modelo final aplicando nuevamente el algoritmo GAs optimizando D.

Resultados

Los mejores modelos obtenidos en las tres repeticiones han presentado los siguientes parámetros:

Repetición No 1			Repetición No 2			Repetición No 3		
Fdes	0.697		Fdes	0.591		Fdes	0.69	
R2cvLmo	0.877		R2cvLmo	0.8592		R2cvLmo	0.8702	
R2	0.875		R2	0.8633		R2	0.874	
RMSEC CvLmo	113.1411		RMSEC CvLmo	120.904		RMSEC CvLmo	116.1092	
RMSEC	112.0516		RMSEC	119.1356		RMSEC	114.3995	
R2 externo	0.875		R2 externo	0.895		R2 externo	0.9	
Descriptor	Coeficientes	Coef. Estand.	Descriptor	Coeficientes	Coef. Estand.	Descriptor	Coeficientes	Coef. Estand.
intercepción	1218.145		intercepción	1.227		intercepción	1224	
SCBO	15.17458	0.04661	Mw	39	0.1232	Mw	52.13397	0.1666
WiDt	109.8752	0.34724	WiDt	90	0.3137	SCBO	135.0045	0.4248
HyWi_Bm	189.9725	0.58344	HyWi_Bm	181	0.5297	Eig1o_AEA(bo)	19.96293	0.0626
						PJI2	69.8287	0.2299
						HyWi_Bm	42.96017	0.1256

La aplicación del método utilizado al conjunto de todos los descriptores seleccionados ha dado los siguientes resultados:

Modelo final		
Fdes	0.817	
R2cvLmo	0.8505	
R2	0.8517	
RMSEC CvLmo	120.1339	
RMSEC	119.5603	
R2 externo	0.927	
Descriptor	Coeficientes	Coef. Estand.
intercepción	1227.137	
PJI2	104.9589	0.34963
SCBO	213.2246	0.673755

Se aprecia que con solo dos descriptores el modelo es muy estable en predicción. El descriptor $PJI2$ es el índice de forma de Petitjean que se basa en la matriz de distancias, mientras $SCBO$ es la suma de los órdenes de enlace del grafo molecular sin hidrógenos (17). Ambos coeficientes son positivos por lo que hay efecto sinérgico sobre el *IR*.

Conclusiones y discusión

Los descriptores seleccionados pueden ser relacionados con dos funciones termodinámicas importantes en el proceso de desorción de las moléculas de la fase estacionaria no polar OV-101, en la fase vapor. $PJI2$ representa interacciones energéticas que se establecen entre moléculas adsorbidas de tipo dispersivo, las mismas que son muy débiles. Evidentemente las moléculas presentan su lado no polar para la absorción. Al aumentar el tamaño hay más partes con orden de enlaces más altos para formar las interacciones de dispersión y por lo tanto se necesita una temperatura más alta para su paso a la fase vapor. $SCBO$ es un índice de forma que está relacionado con la complejidad de la molécula, si la parte no polar que se adhiere a la superficie del OV-101 tiene forma compleja, la variación de entropía en la adsorción es positiva y alta por lo que, para invertir el proceso se necesita una temperatura mayor para que pase en fase vapor.

Agradecimientos

CR agradece la beca doctoral brindada por la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación de la República del Ecuador (SENESCYT).

Referencias Bibliográficas

1) Todeschini R, Data Correlation of Significant Component and Shape of Molecules. The K Correlation Index. Anal.Chim.Acta,1997,348,419-430

2) Todeschini R. Consonni V, and Majocchi A, The K correlation Index: Theory Development and its Aplications in Chemometrics. Chemometrics & Intell.Lab.Syst.1998,46,13-29

3) P. R. Duchowicz, E. A. Castro, and F. M. Fernández, MATCH Commun. Math. Comput. Chem. 55 (2006) 179.
4) P. R. Duchowicz, E. A. Castro, F. M. Fernández, and M. P. González, Chem Phys Lett 412 (2005) 376.
5) Miller A. Subset selection in regression, 2002, C&H/CRC, Monographs on Statistics & Applied Probability, Chapman and Hall/CRC
6) Bledsoe WW, The use of biological concepts in the analytical study of systems,1961, Proceedings of the ORSA-TIMS National Meeting
7) Holland J H, Adaptation in Natural and Artificial Systems. An Introductory Analysis with Applications to Biology, Control, and Artifical intelligence,1992, The MIT Press
8) Goldberg D E, Holland J H, Genetic Algoritms and Machine Learning, 1988, Machine Learning 3, 95-99
9) Davis L Handbook of Genetic Algoritms, 1991, Van Nostrand Reinhold, New York

10) Todeschini R, Introduzione alla chemiometria,1998, EdiSES, Napoli
11) Pavan M, Todeschini R, Total Order Ranking Methods, 51-57 en Pavan M, Todeschini R (Editors) , “Data Handling in Science and Tecnology, Scientific Data Ranking Methods: Theory and Aplications”, 2008, Elsevier, Amsterdam
12) IUPAC Compendium of Chemical Terminology - the Gold Book, 2012.
13) W. Jennings and T. Shibamoto, Qualitative Analysis of Flavor and Fragrance Volatiles by Glass Capillary Gas Chromatography, ACADEMIC PRESS, INC, London, 1980.
14) Hypercube, Inc., http://www.hyper.com.
15) Dragon, Software for Molecular Descriptor Calculation, Talete, SRL., http://www.talete.mi.it/, 2014.
16) Leardi R, Boggia R, Terrile M, Genetic algoritms as a strategy for feature selection. 1992, Jourrnal of Chemometrics, 6, 267-281
17) Dragon, Software for Molecular Descriptor Calculation, Talete, SRL., http://www.talete.mi.it/, 2014.