Predicción del punto de fusión de indoles con base en la estructura molecular usando redes neuronales artificiales

R. Vivas-Reyes*, R. Valencia y N. Ramírez

Grupo de Química Cuántica y Teórica de la Universidad de Cartagena, Programa de Química, Cartagena-Colombia

Grupo de investigación CIPTEC, Programa de ingeniería industrial, Fundación Universitaria Tecnológico de Comfenalco, Cartagena, Colombia

Prediction of melting point of indols based on using molecular structure artificial neural networks

Predicció del punt de fusió dels indols amb base en l'estructura molecular fent servir xarxes neuronals artificials

Recibido: 10 de septiembre de 2015; aceptado: 27 de noviembre de 2015

RESUMEN

Mediante la aplicación del método de relación cuantitativa de estructura propiedad se determinó un modelo para predecir la temperatura del punto de fusión de indoles a partir de su estructura molecular (n = 86). Usando los programas de computadora Gaussian 98 y PCDM 2.0, se calcularon una serie de descriptores moleculares; descriptores electrónicos, topológicos y geométricos. Para la elaboración del modelo de predicción se empleó la regresión linear múltiple entre los descriptores moleculares y la temperatura de los puntos de fusión de los indoles presentes en la base de datos. Dando como resultados un coeficiente de determinación (R^2) y un error estándar de estimación (EEE) de 0.73 y 27.42°C respectivamente. Por medio de una red neuronal retropropagación (5: 4: 1) se optimizó el modelo de regresión lineal múltiple, pudiéndose incluir relaciones no lineales entre la estructura molecular y la temperatura del punto de fusión de los indoles, obteniendo mejores resultados en la predicción del punto de fusión para el grupo de entrenamiento (R^2 =0.9978) y el grupo de validación (R^2 =0.9987). El error cuadrático promedio (MSE) asociado al grupo de entrenamiento y de validación para el modelo de con la red fue 0.006 y 0.006 respectivamente.

Palabras clave: Relación cuantitativa de estructura propiedad; puntos de fusión; indoles; descriptores moleculares; regresión linear múltiple y red neuronal retropropagación.

SUMMARY

By applying the method of quantitative structure property relationship, was determined a model to predict the temperature of the melting point of indoles from their molecular structure (n = 86). Using computer programs such as Gaussian 98 and PCDM, a series of electronic, topological and geometric descriptors were calculated. For the preparation of the prediction was used a multiple linear regressions model between molecular descriptors and the temperature of the melting points of indols present in the database. Giving as a result a coefficient of determination (R^2) and standard error of estimate (EEA) of 0.73 and 27.42 ° C respectively.

Through a neural network backpropagation (5: 4: 1) model of multiple linear regression was optimized, including possible non-linear relationships between molecular structure and temperature of the melting point of the indols, obtaining better results in predicting melting point for the training group (R^2 = 0.9978) and the validation group (R^2 = 0.9987). The mean square error (MSE) associated with the training group and validation for the network model was 0.006 and 0.006 respectively.

Keywords: quantitative structure property relationship; melting points; indoles; molecular descriptors; Multiple linear regression and back-propagation neural network.

RESUM

Mitjançant l'aplicació del mètode de relació quantitativa d'estructura propietat es va determinar un model per predir la temperatura del punt de fusió dels indols a partir de la seva estructura molecular (n=86). Usant els programes d'ordinador Gaussian 98 i PCDM 2.0, es van calcular una sèrie de descriptors moleculars; descriptors electrònics, topològics i geomètrics. Per a l'elaboració del model de predicció es va emprar la regressió lineal múltiple entre els descriptors moleculars i la temperatura dels punts de fusió dels indols presents a la base de dades. Donant com a resultats un coeficient de determinació (R2) i un error estàndard d'estimació (EEE) de 0.73 i 27.42°C, respectivament. Per mitjà d'una xarxa neuronal retropropagació (5: 4: 1) es va optimitzar el model de regressió lineal múltiple, podentse incloure relacions no lineals entre l'estructura molecular i la temperatura del punt de fusió dels indols, obtenint millors resultats en la predicció del punt de fusió per al

*Autor para la correspondencia: rvivasr@unicartagena.edu.co

grup d'entrenament (R2 = 0,9978) i el grup de validació (R2 = 0,9987). L'error quadràtic mitjà (MSE) associat al grup d'entrenament i de validació per al model de amb la xarxa va ser 0.006 i 0.006 respectivament.

Paraules clau: Relació quantitativa d'estructura propietat; punts de fusió; indols; descriptors moleculars; regressió lineal múltiple i xarxa neuronal retropropagació.

INTRODUCCIÓN

El punto de fusión de un sólido se define como la temperatura a la cual la fase sólida y liquida coexisten en el equilibrio. El punto de fusión de los sólidos refleja la magnitud de las fuerzas intermoleculares que actúan en la molécula; esas fuerzas son las que mantienen unidas a las moléculas, átomos o iones, las cuales son las principales responsables de las propiedades macroscópicas de estos materiales (1).

El punto de fusión es una propiedad fisicoquímica la cual ayuda a comprender propiedades bioquímicas como la toxicidad (2,3) y otras variables fisicoquímicas como la energía reticular y la solubilidad en agua de muchos compuestos. El punto de fusión tiene numerosas aplicaciones en la bioquímica y en las ciencias medioambientales debido a su relación con la solubilidad (4). La solubilidad a su vez es importante en el diseño de fármacos y es un medidor de la toxicidad efectiva de químicos y materiales (5).

Sin embargo, no se ha desarrollado aún un método general basado en la estructura química de los compuestos para predecir el punto de fusión. Algunos de los trabajos previos en relación cuantitativa estructura propiedad (Quantitative Structure-Property Relationships: QSPR) para determinar el punto de fusión a partir de la estructura química de los compuestos están confinados a conjuntos reducidos de hidrocarburos y a familias de aromáticos dando lugar a modelos relativamente satisfactorios. El modelo Needham *et al.*, reportado en la literatura para alcanos normales y ramificados usando índices topológicos, presenta un error estándar de estimación de 23.8 K (6).

Las correlaciones de puntos de fusión de compuestos aromáticos, se han desarrollado con un éxito moderado. Por ejemplo, para el conjunto de 443 mono y di bencenos sustituidos, combinando descriptores tanto cuánticos como tradicionales, Katritzky obtiene un coeficiente de determinación de 0.84 (7,8), y para 72 compuestos análogos de la 1,2,3-diazoborinas, empleando descriptores electrónicos y topológicos, obteniendo un coeficiente de determinación y error estándar de estimación de 0.86 y 16.79°C respectivamente. (9).

En los últimos años el interés por los modelos QSPR basados en redes neuronales se ha incrementado. La principal ventaja de los modelos de redes recae en el hecho que un modelo QSPR puede desarrollarse sin especificar a priori la forma analítica del modelo. Las redes neuronales son especialmente útiles para establecer las complejas relaciones existentes entre la salida del modelo (propiedades fisicoquímicas) y la entrada del modelo (descriptores moleculares). Los modelos QSPR basados en redes neuronales utilizan principalmente algoritmos del tipo retropropagación (back-propagation). La red retopropagación emplea un sistema de aprendizaje por minimización del error.

En este estudio nosotros consideraremos el uso de los descriptores moleculares del mejor modelo estadístico obtenido por la regresión lineal múltiple como capa de entrada de la red neuronal retropropagación para la predicción del punto de fusión de una serie de compuestos que presentan el anillo indolico como estructura base (indoles), pudiendo incluir relaciones no lineales entre su estructura química y el punto de fusión, dando la oportunidad de obtener mejores resultados en la predicción.

METODOLOGÍA

El procedimiento seguido para la elaboración del modelo de predicción del punto de fusión de indoles con base en la estructura molecular usando redes neuronales artificiales, incluye los siguientes pasos: selección de la base de datos, ingreso de las estructuras y modelación molecular, cálculo de los descriptores moleculares, análisis estadístico para obtener el modelo, validación del modelo y por último la optimización del modelo usando redes neuronales artificiales. (9,10).

Este trabajo fue realizado en el Laboratorio del Grupo de Química Cuántica y Computacional de la Universidad de Cartagena, a través de los programas Gaussian 98,(11) PCDM Versión 2.0 (12), Statgraphics (13)

Selección de la base de datos: Los 86 indoles estudiados con sus diferentes puntos de fusión experimentales (puntos de fusión promedios) fueron tomados de los manuales comerciales de las compañías Sigma-Aldrich Chemical Co. y Lancaster Synthesis Inc.

Éstos compuestos poseen un rango de puntos de fusión de 30.5 °C a 247.0 °C, y un valor promedio de 124.9 °C. En general, la diferencia entre los puntos de fusión para esta serie indoles se ve marcada por las sustituciones que se presentan en el anillo indolico (ver Figura 1).

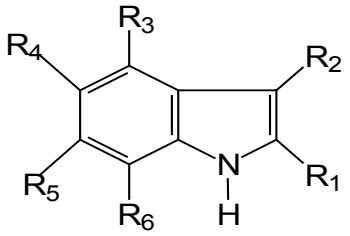

Figura 1. Anillo indolico y sustituciones

Ingreso de las estructuras y modelación molecular: Las estructuras moleculares de cada uno de los indoles fueron dibujadas en el programa GaussView. El cual generó una matriz de conexión, información del anillo, coordenadas atómicas y otras informaciones para cada una de las estructuras dibujadas. Luego se realizó una optimización geométrica para cada estructura utilizando un método de cálculo Hartree-Fock (full optimization) en el programa Gaussian 98 , para encontrar la estructura más estable.

Cálculos de los descriptores moleculares: Mediante el programa Gaussian 98 se calcularon descriptores electrónicos a través de un método de cálculo ab-initio sobre un conjunto de base 6-31G* (14) (HF/6-31G*) usando el programa computacional Gaussian 98 después de ser optimizadas, estos incluyen energía total, energía de orbítales frontera, momento dipolar, momento cuadrapolar, suavidad, dureza, electronegatividad y constante rotacional. Los descriptores topológicos, electrónicos y geométricos se obtuvieron mediante el programa PCDM Versión

2.0; desarrollado en Turbo Pascall por el grupo de investigación de Química ambiental y computacional,, éstos incluyen cargas parciales atómicas, índices de Randic (15), índices de conectividad (16), descriptores de fragmentación (número de átomos de oxígeno, de heteroatomos, de hidrógenos, de carbonos, número de enlaces simples, peso molecular etc.), y la superficie y el volumen molecular.

Se calcularon un total de 96 descriptores moleculares; de los cuales 41 fueron descriptores electrónicos, 29 topológicos, 4 geométricos y 22 descriptores adicionales (se obtuvieron de diferentes combinaciones entre varios descriptores moleculares).

Análisis estadístico para obtener el modelo: La primera etapa para desarrollar modelo de relación cuantitativa estructura propiedad es disminuir el número de variables y seleccionar las más importantes. Para esto se emplearon las técnicas de *Descriptores No Redundantes (DNR)* (17) Esta técnica se basa en que si dos descriptores moleculares se correlacionaban entre ellos con un valor de R mayor que 0.9, y éstos a su vez se correlacionaban por separado con la variable dependiente (PF), y el A*nálisis de Componentes Principales (ACP).* (18). En el análisis de componentes principales se escogieron aquellos descriptores que estaban relacionados con los tres primeros componentes principales, los cuales contenían el 80% de la significancia.

Posteriormente se procedió a realizar una regresión lineal múltiple mediante el método de selección hacia delante o forward (19) entre los descriptores restantes de la eliminación y los puntos de fusión experimentales.

Todos los cálculos estadísticos se realizaron con la ayuda del programa Statgraphics (13)

La calidad del modelo de regresión se midió usando primordialmente cuatro parámetros estadísticos: el coeficiente de determinación (R^2) que indica la fracción o porcentaje en que las variables independientes predicen a la variable dependiente; el error estándar de estimación (EEE) el cual índica el error general del modelo; el valor F del análisis de varianza (F) que indica la variación de las variables en el modelo y el valor del nivel de significancia (P) el cual nos indica si el modelo es significativo o la confianza que se tiene para predecir un dato (18).

Validación del modelo: El modelo deducido estadísticamente se validó mediante una técnica de validación cruzada interna (*cross validación*), que consiste en excluir una molécula de la base de datos y recalcular nuevamente el modelo con las n - 1 moléculas restantes; el dato excluido se calcula con el modelo obtenido, esto se realizó para cada una de las 86 moléculas usadas para calcular el modelo. Luego se correlacionó los datos calculados para cada una de las moléculas obtenidas por el método anterior con los datos experimentales. En este caso el coeficiente de correlación de la validación cruzada debe ser igual o similar al encontrado por el modelo (20).

Optimización con redes neuronales artificiales: La red neuronal usada en este trabajo consistió de un sistema de tres capas conectadas completamente (capa de entrada-oculta-de salida). Cada neurona en una capa dada se conecta totalmente a todas las neuronas en las capas adyacentes.

El número de neuronas en la capa de entrada es igual al número de descriptores moleculares tomados por el modelo obtenido de la regresión lineal múltiple. El número de neuronas en la capa oculta fue determinado por ensayo y error, calculando el error cuadrático promedio (MSE) y el coeficiente de determinación (R^2) para el grupo de entrenamiento y el grupo de validación en el proceso de aprendizaje. La capa de salida contiene una neurona representando el valor del punto de fusión calculado por la red.

La red neuronal artificial fue entrenada usando el algoritmo de retropropagación normal mediante el programa WinNN 32 (21). En la capa de entrada se trasmite una señal a la neurona de la capa oculta, la cual procesa los datos usando una función sigmoidal y los envía a la capa de salida. La función sigmoidal empleada fue la tangente hiperbólico, $tanh(x) = exp(X)-exp(-X)/exp(X)+exp(-X)$, la cual fue usada para calcular salida de la neurona en la capa oculta.

Los valores que alimentaban a la capa de la entrada (valores de los descriptores moleculares) al proceso de las redes neuronales artificiales fueron normalizados para que éstos tuvieran un valor promedio de cero y desviación estándar de uno. Similarmente, los pesos y bias iniciales se generan al azar en el intervalo (-1,1). También se optimizó la velocidad de aprendizaje por ensayo y error, obteniendo como mejor resultado 0.15.

Para entrenar la red neuronal, la base de los datos fue dividida al azar en dos grupos, un grupo de entrenamiento y un grupo de validación consistente en 76 y 10 moléculas respectivamente.

Los datos del grupo de validación fueron utilizados para monitorear la red, de tal forma que se fuera evitando el sobreentrenamiento del grupo de entrenamiento y mejorar así el poder de predicción de la red. La red fue entrenada hasta un punto óptimo en el que el error cuadrático promedio fue el mínimo entrenando y validando simultáneamente.

Con el grupo de entrenamiento se generó el modelo y el grupo de la validación cruzada se empleó para monitorear el proceso de entrenamiento y evaluar el modelo generado (22).

RESULTADOS Y DISCUSIÓN

Modelo con regresión lineal múltiple: Los descriptores moleculares empleados en el modelo son: Sumatoria de las cargas atómicas de los carbonos (CAC), carga total de los heteroatomos con sus hidrógenos (CTHH), el número de átomos de oxigeno presentes en la molécula (OXIG), el momento dipolar (Dipolo) y el índice de conectividad de tercer orden con valencia ramificado (X3Cv).

El modelo posee un coeficiente de determinación (R^2) igual a 73.30%, es decir que el 73.30 de la varianza residual es explicada por el modelo obtenido, lo cual nos dice que existe una correlación moderadamente alta entre los descriptores moleculares utilizados en el modelo y el punto de fusión de los indoles estudiados.

Los resultados del análisis de varianza se presentan en la Tabla 1 y los valores del punto de fusión experimentales contra los calculados por el modelo con la regresión lineal múltiple se muestran en la figura 2. Mostrando que al modelo tiene una alta significancia ($P < 0.0001$), lo cual nos quiere decir que hay una relación estadísticamente significativa entre las variable del 99% o el modelo fue calculado con nivel de confianza del 99%.

De acuerdo con los valores del T estadístico obtenidos para cada descriptor molecular, tenemos que el descriptor más importante en la regresión lineal múltiple es el *X3Cv*.

Tabla 1. *Modelo seleccionado por la RLM entre los puntos de fusión de los indoles estudiados y los descriptores moleculares.*

Descriptores Moleculares	Coeficiente de Regresión	Error Estándar del Coeficiente	T estadístico	Valor P
Constante	-88.416	24.982	-3,539	0,0007
CAC	1355.69	247.028	5,488	0,0000
CTHH	73,273	17.245	4.249	0,0001
Dipolo	9.240	1.764	5.237	0,0000
OXIG	15.341	4.314	3.556	0,0006
X3Cv	227.84	39.440	5,777	0,0000

*PF = -88.416 + 1355.69*CAC + 73,273*CTHH + 9.240*Dipolo + 15,341*OXIG + 227,84*X3Cv*

R^2 = 73.30; R = 0.84; EEE = 27.40; Valor P < 0.0001;
Razón F = 43.93; $R_{crossval}$ = 0.813; n = 86

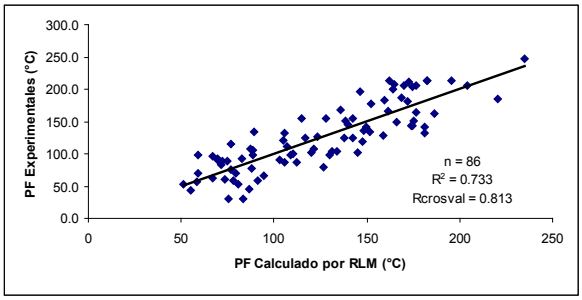

Figura 2. *Valores del punto de fusión experimentales contra los calculados por el modelo de predicción con la regresión lineal múltiple*

Por medio de la matriz de correlación de los descriptores moleculares involucrados en el modelo (Tabla 2), notamos el carácter no redundante de cada uno de los descriptores presentes (R < 0.5). Pudiendo afirmar que no hay problemas os de multicolinealidad entre los descriptores moleculares.

Tabla 2. *Matriz de correlación de los descriptores moleculares utilizados en el modelo de predicción.*

	CAC	CTHH	Dipolo	OXIG	X3Cv
CAC	1.000				
CTHH	0.286	1.000			
Dipolo	-0.399	-0.324	1.000		
OXIG	-0.089	-0.492	-0.292	1.000	
X3Cv	-0.534	-0.368	-0.331	0.195	1.000

Con el objeto de detectar intercorrelaciones entre las variables seleccionadas por el modelo, se empleó el método multivariado de Análisis Factor (23), que permite combinar los descriptores moleculares existentes en el modelo para obtener un número reducido de nuevas variables (factores), con las cuales se puede explicar la información contenida en todas las variables inicialmente incluidas. Para los descriptores del modelo de predicción fueron extraídos tres factores, es decir que las cinco variables se combinaron en tres nuevas variables o factores. El primer factor explica el 44.708%, el segundo el 20.4943% y el tercero 20.669% de la variabilidad total de los datos, para un acumulado del 86.320% (Tabla 3).
En la Tabla 3 se ilustra que hay cinco variables y así la suma de las varianza para los datos es cinco para las variables ya estandarizadas (descartando los errores de redondeo).

La primera columna de los autovalores muestra cómo la varianza se comparte entre los cinco factores, teniendo el factor 1 una varianza de 2.235, el factor 2 de 1.047, así sucesivamente. Nótese que, como se esperaría, el factor 1 tiene la varianza más grande, el factor 2 la siguiente más alta, y así sucesivamente. Los factores con un autovalor mayor que 1 contribuyen más a la varianza que las otras variables originales (18).

Tabla 3. Análisis Factor para los descriptores moleculares usados en el modelo estadístico por el método Varimax Rotado.

Número de Factores	Autovalores	Porcentaje de la Varianza	Porcentaje Acumulativo de la Varianza
1	2.235	44.708	44.708
2	1.047	20.943	65.651
3	1.033	20.669	86.320
4	0.410	8.198	94.518
5	0.274	5.482	100.000

Con la finalidad de obtener una mejor interpretación del Análisis Factor, los principales factores extraídos se rotaron aplicando el método Varimax con lo cual se maximiza la varianza de cada factor y de esta forma los valores aparecerán más altos en unos y bajos en otros, mostrando la mejor relación entre cada factor y las variables que lo conformaron. Los resultados se presentan en la Tabla 4 muestran que el primer factor esta relacionado con OXIG, CAC y CTHH, el segundo factor con X3Cv y el tercer factor con Dipolo.

Tabla 4. *Matriz del Análisis Factor para los descriptores moleculares usados para predecir el punto de fusión de los indoles estudiados.*

Descriptor Molecular	Factor		
	1	2	3
CAC	0.879	-0.199	0.129
CTHH	-0.691	-0.497	0.300
Dipolo	0.134	0.096	0.956
OXIG	0.834	0.269	0.247
X3Cv	0.035	0.940	0.134

OPTIMIZACIÓN DEL MODELO UTILIZANDO REDES NEURONALES ARTIFICIALES (RNA)

La arquitectura de la red neuronal de retropropagación fue obtenida ajustando los parámetros de optimización por error y ensayo. El parámetro de la velocidad de aprendizaje fue proporcionado antes del entrenamiento para reducir el error de la red neuronal. El número de neuronas en la capa de entrada fue de cinco, una por cada descriptor obtenido en el modelo de regresión lineal múltiple. La capa de salida utilizó una neurona, la cual corresponde al vector de los valores del punto de fusión para los indoles estudiados. El número de neuronas en la capa oculta fue determinado corriendo la red neuronal desde 5:1:1 (cinco neuronas de entrada, una neurona en la capa oculta y una neurona en la capa de salida) hasta 5:8:1 (cinco neuronas de entrada, ocho neuronas en la capa oculta y una neurona en la capa de salida).
Los valores del coeficiente de determinación (R^2) para el grupo de entrenamiento y de validación muestran mejores resultados cuando se tienen cuatro neuronas en la capa oculta y una disminución significativa en el error

cuadrático promedio (MSE) en el grupo de entrenamiento y validación, razón por la cual la arquitectura considerada como la mejor fue la **5: 4: 1** (Tabla 5). Los valores del punto de fusión experimental y los calculados por la red se muestran en la Tabla 6 y la grafica de los valores del punto de fusión experimental y los calculados por la red para el grupo de entrenamiento se representa en la Figura 3.

Tabla 5. *Resultados con la red retropropagación incrementando el número en las capas ocultas de la red.*

Neuronas	MSE de Entrenamiento	MSE de Validación	R^2 de Entrenamiento	R^2 de Validación
1	0,0123	0,0158	0,814	0,782
2	0,0102	0,0135	0,889	0,846
3	0,0085	0,0092	0,945	0,916
4	***0,0066***	***0,0065***	***0,983***	***0,964***
5	0,0084	0,0097	0,932	0,915
6	0,0116	0,0124	0,895	0,872
7	0,0125	0,0156	0,838	0,816
8	0,0152	0,0167	0,791	0,770

Tabla 6. *Valores del punto de fusión experimentales, calculados por el modelo de predicción con la regresión lineal múltiple y con la red para el conjunto de entrenamiento.*

Moléculas		Punto de Fusion (°C)		
No	CAS No[1]	Experimental	**Calculado por la RLM**	**Calculado por la RNA**
1	61-49-4	88.0	74.819	89,130
2	61-54-1	114.5	76.8493	116,370
3	83-34-1	96.0	67.0743	95,214
4	87-51-4[2]	165.5	161.345	165,126
5	87-52-5	133.0	105.773	133,125
7	95-20-5	59.0	77.8840	61,422
8	120-72-9	53.0	51.1105	52,432
9	133-32-4	124.5	142.081	125,483
10	299-26-3	99.0	129.996	100,725
11	348-36-7	146.0	139.601	145,422
12	387-43-9	31.0	83.1839	33,122
13	387-44-0	62.0	66.9122	65,130
14	399-51-9	75.5	76.7561	76,498
15	399-52-0	45.0	86.5111	48,125
16	399-72-4[2]	99.5	110.340	101,430
17	487-89-8	196.5	146.0230	198,435
18	526-55-6[2]	58.5	91.0341	61,124
19	608-08-2	129.0	158.679	133,428
21	700-06-1	97.5	88.8186	95,136
22	827-01-0	211.0	172.808	214,589
23	830-96-6	134.5	151.635	139,525
24	877-03-2	205.0	170.077	209,134
25	933-67-5	83.5	71.3669	82,244
26	942-24-5	150.5	138.25	149,435
27	1006-94-6	53.5	80.9149	55,364
28	1011-65-0[2]	124.5	137.472	124,359
30	1075-35-0	112.0	107.099	111,223
31	1076-74-0[2]	87.0	105.553	89,142
32	1196-69-6	101.5	144.968	103,679

Moléculas		**Punto de Fusion (°C)**		
No	**CAS No**	**Experimental**	**Calculado por la RLM**	**Calculado por la RNA**
34	1477-50-5	204.5	174.399	204,246
35	1670-81-1	212.5	182.66	214,249
36	1953-54-4[2]	107.5	121.471	110,236
37	2124-55-2	213.5	182.573	212,210
38	2338-71-8	164.0	176.274	165,435
39	2380-94-1	98.0	109.232	99,436
40	2882-15-7	162.0	186.119	162,361
41	3189-13-7[2]	92.0	82.819	92,003
42	3389-21-7	98.0	58.9165	98,312
44	3770-50-1	124.0	116.861	122,370
45	4769-96-4	143.5	174.006	149,365
46	4769-97-5	206.0	176.694	206,320
47	4771-49-7	187.0	168.371	189,431
48	4771-50-0	207.0	164.695	219,014
49	4792-58-9	155.5	142.498	155,131
50	4792-67-0	167.5	135.634	169,251
51	4837-90-5	70.0	59.1212	70,205
52	5192-04-1	90.5	102.693	92,325
54	5192-03-0	134.5	89.3256	134,003
55	5318-27-4	65.5	94.2652	70,125
56	5416-80-8	200.5	163.83	200,036
57	5448-47-5	143.5	174.82	149,136
58	5585-96-6	101.0	120.375	104,125
59	6146-52-7	141.5	181.36	140,421
61	7598-91-6	206.5	204.093	205,842
62	10075-50-0	91.0	69.7576	86,312
63	10601-19-1	181.0	171.613	188,859
64	13544-43-9	103.0	131.098	108,364

Moléculas		**Punto de Fusion (°C)**		
No	**CAS No**	**Experimental**	**Calculado por la RLM**	**Calculado por la RNA**
66	14618-45-2[2]	212.5	195.702	215,231
67	16382-18-6[2]	177.0	152.4	178,312
68	16620-52-3	126.0	123.147	125,134
70	17422-32-1	70.0	79.656	73,252
71	17422-33-2	89.5	72.2708	86,346
72	21296-94-6	185.0	220.332	185,435
73	21598-06-1	247.0	235.117	247,223
75	50820-65-0	79.0	126.798	79,214

76	51417-51-7	43.0	55.0452	45,124
77	52415-29-9	92.0	69.5683	93,124
78	52562-50-2	149.5	166.125	150,286
79	53855-47-3	135.0	148.224	135,456
80	53924-05-3	57.5	58.5276	57,946
81	98081-83-5	118.0	147.577	119,370
82	105776-13-4[2]	182.5	159.159	182,134
83	111258-23-2	141.0	149.777	145,141
84	165669-16-9	214.0	162.18	214,311
85	302912-21-6	77.5	88.0095	79,122
86	374537-99-2	104.5	133.548	106,245

1. CAS No: Número del Chemical Abstracts Service.
2. Indoles que conforman el grupo de validación.

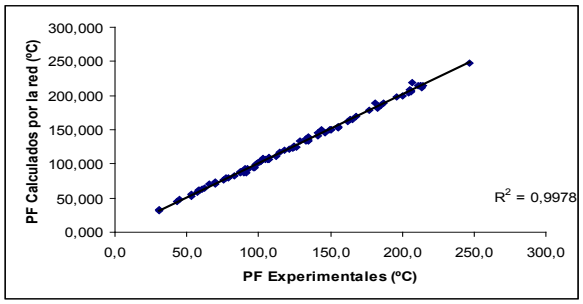

Figura 3. *Valores del punto de fusión experimentales contra los calculados por la red para el grupo de entrenamiento.*

Tabla 7. *Valores del punto de fusión experimentales, calculados por el modelo de predicción con la regresión lineal múltiple y con la red para el conjunto de validación*

Moléculas		Punto de Fusion (°C)		
No	CAS No[1]	Experimental	**Calculado por la RLM**	**Calculado por la RNA**
6	91-55-4	106.0	88.8114	106,440
20	614-96-0	61.0	73.5201	62,425
29	1074-88-0	87.5	112.219	87,013
33	1210-83-9	121.0	105.009	121,426
43	3420-08-2	30.5	75.352	32,124
53	5192-23-4	107.5	87.5943	106,316
60	7570-49-2[2]	155.0	127.951	153,021
65	14430-23-0	155.5	114.559	153,131
69	17357-14-1	133.0	180.915	136,135
74	24985-85-1	150.5	175.395	150,134

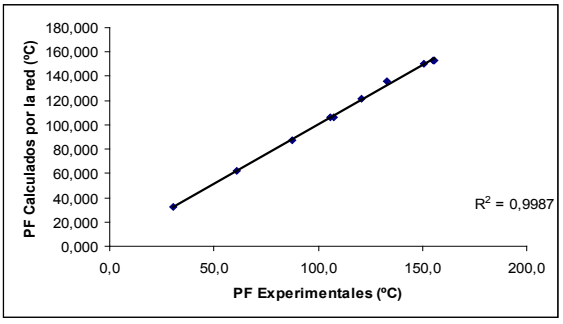

Figura 4. *Valores del punto de fusión experimentales contra los calculados por la red para el conjunto de validación.*

INTERPRETACIÓN DE LOS DESCRIPTORES MOLECULARES

En el modelo propuesto aparecen 3 descriptores electrónicos y 2 topológicos. Cada descriptor en el modelo codifica diferentes aspectos de la estructura molecular de los indoles. Entre los descriptores electrónicos están: El momento dipolar (Dipolo), sumatoria de cargas atómicas de los carbonos (CAC), y carga total de los heteroatomos con sus hidrógenos (CTHH).*El momento dipolar (Dipolo):* es un ejemplo típico de descriptores electrónicos globales que relaciona la distribución de cargas en el espacio y este es responsable del campo electrónico externo Se ha demostrado que esas propiedades están relacionadas con las interacciones intermoleculares tipo dipolo-dipolo (24), las cuales afectan considerablemente los puntos de fusión de los compuestos químicos. El momento dipolar tiene que ver con la polaridad, mientras más polares sea la molécula, presenta mayor momento dipolar (25) En los indoles el carácter polar esta asociado con la presencia de heteroatomos, en particular los átomos de nitrógeno y oxigeno, contribuyendo a que las interacciones dipolo-dipolo sean más fuertes en la molécula.

La carga total de los heteroatomos con sus hidrógenos (CTHH): es un descriptor electrónico determinado básicamente por las cargas atómicas de los átomos de nitrógeno, oxígeno y halógenos presentes en la estructura molecular. Como la mayoría de los heteroatomos presentes en los indoles contienen carga nuclear mayor que el carbono y pares de electrones libres disponibles para ser atraídos, esto hace que las interacciones electrostáticas se den con mayor frecuencia entre los átomos. Estas cargas sobre los heteroatomos y sus hidrógenos puede reflejarse en la capacidad de formar puentes de hidrógeno intermolecularmente en los indoles, este tipo de enlace afectan a los puntos de fusión debido a que son fuerzas intraatómicas fuertes (26). Este descriptor también fue importante en la predicción del punto de fusión de las anilinas (27)

Sumatoria de cargas atómicas de los carbonos (CAC): La distribución de las cargas atómicas de lo carbonos facilita la formación de interacciones electrostática tipo dipolo-dipolo, esta distribución afectaría directamente a la fuerzas de atracción de los átomos entre si y así a los puntos de fusión (28).

Los descriptores topológicos empleados en el modelo fueron: El número de átomos de oxigeno presentes en la estructura (OXIG) y el índice de conectividad molecular de tercer orden con valencia ramificado (X3Cv).

El número de átomos de oxigeno presentes en la estructura (OXIG): este descriptor topológico esta relacionado con el momento dipolar, ya que al aumentar el número de átomos de oxigeno aumenta el momento dipolar, esto debido a los electrones libes del átomo de oxigeno forma un dipolo cargado negativamente. En los indoles cada átomo de oxigeno tiene una carga parcialmente y estos en algunos casos se encuentran unido a un átomo de hidrógeno pudiéndose formar puentes de hidrógeno, los cuales afectan considerablemente a los puntos de fusión por ser fuerzas de atracción muy fuertes contribuyendo al aumento de la temperatura del punto de fusión.

El índice de conectividad molecular tres con valencia ramificado (X3Cv): (16): los términos o índices de conectividad molecular son descriptores topológicos que depende del arreglo de los átomos en la molécula. El X3Cv es un índice de agrupación que describe más específicamente las

ramificaciones de la molécula, este se refiere a la distribución de tres enlaces sobre átomos cualesquiera. Como este descriptor codifica el tamaño y el grado de las ramificaciones de la molécula, lo cual es un factor que hay que tener en cuenta debido a que las ramificaciones dificultan que una molécula se sobreponga con otra, afectando el empaquetamiento y así a los puntos de fusión de los diferentes compuestos (29).

CONCLUSIONES

La temperatura del punto de fusión de los estudiados esta asociada con los factores moleculares que describen las interacciones de tipo electrostático, forma molécula y número de átomos de oxigeno presente en la estructura. En general los descriptores moleculares del modelo encontrados son los que mejor están relacionados con el punto de fusión de los indoles objeto de estudio.

La red neuronal retropropagación se vale del resultado de análisis estadístico exhaustivo, realizado por las técnicas de descriptores no redundantes, análisis de componentes principales y la regresión lineal múltiple para crear su propio modelo. Refinando así el modelo obtenido por la regresión lineal múltiple de un coeficiente de determinación de 0.733 a un coeficiente de determinación de 0.983 dado por la red. Como también mejorando la predicción, el error de predicción promedio de la red es de 1.18%.

La consideración de todo lo dicho nos lleva a sugerir, que las técnicas estadísticas y las redes neuronales deben comenzar a ser usadas conjuntamente. De este modo, la estadística, centrada tradicionalmente en funciones lineales, y las redes neuronales, más acostumbradas a tratar con problemas mal definidos o no lineales, se verán mutuamente enriquecidas.

BIBLIOGRAFIA

1. Abramowitz, R.; Yalkowsky, S.H. Melting point, boiling point and symmetry. *Pharm. Res.* **1990,** 7, 942-947.
2. Afifi, A. A; Azon, S. P. *Statical Analysis: a computer oriented approach* Segunda Edición, Academic Press, New York. **1979.**
3. Álvarez, R. *Estadística multivariante y no paramétrica con SPSS: Aplicaciones a las ciencias de la salud*. Diaz de Santos Ediciones, S.A. España **1995**.
4. Andersson, P., Haglund, P., Rappe, C., y Tysklind. M. Ultraviolet Adsorption Characteristies and Calculated Semi-empirical Parameters as Chemical Descriptors in Multivariate Modeling of Polychlorinated Biphenyls. *Chemometrics.* **1996**, 10, 171-185.
5. Argese, E.; Bettiol, C.; Giurin, G.; Miana, P. Quantitative structure-activity relationships for the toxicity of chlorophenols to mammalian submitochondrial particles. *Chemosphere.* **1999**, 38, 2281-2292.
6. Atkins, P. W. In physical chemistry. Fifth Edition, W.H. Freeman and Co., New York. **1994.**
7. Benoit-Guyod, J. L.; Andre, C.; Taillandier, G.; Rochat, J.; Boucherle, A. Toxicity and QSAR of chlorophenols on Lebistes reticulatus. *Ecotoxicol. Environ. Saf*. **1984**, 8, 227-235.
8. Binkley, J. S.; Defrees, D. J.; Baker, J.; J. Stewart, P.; Head-Gordon, M.; Gonzalez, C.; Pople, J. A.;.*Gaussian 98*, *Revision D.3,* Gaussian, Inc., Pittsburgh PA. **1995**.
9. Castellan, G. W. *Fisicoquímica*. Segunda Edición, Addison Wesley Iberoamericana, Argentina. **1987.**
10. Dearden, J. C. The QSAR prediction of melting point, a property of environmental relevance. *Sci. Total Environ.* **1991**, 110, 59-68.
11. Dearden, J. C; Ghafourian, T. Hydrogen bonding parameters for QSAR: comparison of indicador variables, hydrogen bond counts, molecular orbital and other parameters. *J. Chem. Inf. Comput. Sci*. **1999**, 39, 231-235
12. Frisch, *M.* J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M.; Johnson, B. G.; Robb, M. A.; Cheeseman, J. R.; Keith, T.; Petersson, G. A.; Montgomery, J. A.; Raghavachari, K.; Al-Laham, M. A.; Zakrzewski, V. G.; Ortiz, J. V.; Foresman, J. B.; Cioslowski, J.; Stefanov, B. B.; Nanayakkara, A.; Challacombe, M.; Peng, C. Y.; Ayala, P. Y.; Chen, W. M.; Wong, W.; Andres, J. L.; Replogle, E. S.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Binkley, J. S.; Defrees, D. J.; Baker, J.; J. Stewart, P.; Head-Gordon, M.; Gonzalez, C.; Pople, J. A.;.*Gaussian 98*, *Revision D.3,* Gaussian, Inc., Pittsburgh PA. **1995**.
13. Hehre, W. J., Radom, L., Van, P., Schleyer, R., y Pople, J.A. “Ab Initio Molecular Orbital Theory”, J. Wiley, New York **1986**.
14. Hilera J. R.; Martínez V. J. *Redes neuronales artificiales:* fundamentos, modelos y aplicaciones. Editorial Alfa omega RA-MA, España. **2002**.
15. Gasteiger, T; Marsili, M. Iterate partial equalization of orbital electronegativity a rapic access to atomic charges. *Tethraedron*, **1980**, 36, 3219-3227.
16. Goll, E. S.; Jurs, P. C. Prediction of vapor pressures of hydrocarbons and halohydrocarbons from molecular structures with a computational neural network model. *J. Chem. Inf. Comput. Sci.* **1999**, 39, 1081-1089
17. Jansson, P.A. Neural networks: An overview. *Anal. Chem.* **1991,** 63, 357A -362A.
18. Johnson-Restrepo, B.; Pacheco-Londoño, L.; Olivero-Verbel, J. Molecular parameters responsible for the melting point of 1,2,3-diazoborine compounds. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 1513-1519
19. Katritzky, A. R.; Murugan, R.; Grendze, M. P.; Toomey, J.E.; Karelson, Jr. M.; Lobanov, V.; Rachwal. P. Predicting physical properties from molecular structure. *Chem. Tech.*, **1994**, *24*, 17-24
20. Katritzky A. Karelson M. and Lobanov V. QSPR as Means of Predicting and Understanding Chemical and Physical Properties in Terms of Structure. *Pure & Appl. Chem*. **1997**, 69, 2, 245.
21. Katritzky A. Maran U. Karelson M. and Lobanov V, Prediction of Melting points for the Substituted Benzenes: A QSPR Approach. *J. Chem. Inf. Comput. Sci*., **1997,** 37, 913.
22. Katritzky, A. R., Lovanov, V.S., Karelson, M., Murugan, R., Grendze, M.P., Toomey, J.E. Jr. Comprehensive descriptors for structural and statistical analysis. 1. Correlations between structure and physical properties of sudstituted pyridine. Rev. Roum. Chim. **1996,** 41, 851-867.
23. Kier, L. B.; Hall, L. H. *In Molecular Connectivity in Chemistry and Drug Research*; Bawden, D., Editorial Research Studies Press LTD, Letcworth, Hertfordshire, England. **1986**.

24. Kier, l.; May, H. l. *Cheometrics Series.* Research Studies Press Ltda. New York. **1986**.
25. Levine, I.N: *Química Cuántica*. Quinta Edición. Editorial Préntice Hall. España, **2001**.
26.
27. Levine, I.N. *Fisicoquímica*. Cuarta Edición, Editorial McGraw-Hill, España. **1996**.
28. Miller, J. N; Miller, J.C. *Estadística y Quimiometria para Química Analítica,* Cuarta Edición, Editorial Prentice-Hall, España. **2002**.
29. Morrison, R.T.; Boyd, R.N. *Química orgánica*. Quinta Edición, Addison Wesley Longman, México D. F. **1990**.
30. Needham D. Wei I. Seybold P. Molecular Modeling of the Physical Properties of the Alkanes. *J. Am. Chem. Soc*. **1988**, 110, 4186
31. Olivero, J.; Payares, P.; Díaz, D.; Vivas, R.; Mercado, J. *PCDM V.2.1995.* Modificado por Pacheco, L. y Johnson, B. Universidad de Cartagena, Colombia. **2000**.
32. Randic, M. On characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, 97, 6609-6614.
33. Stanton, D. T. Development of a Quantitative Structure-Property Relationship model for estimating normal boiling point of small multifunctional organic molecules. *J. Chem. Inf. Comput. Sci*. **2000**, 40, 81-90
34. Statgraphics Plus for Windows. Statistical graphics System, User's Guide version 3.0. Statistical Graphics Corporation. **1999**.
35. Randic, M. On characterization of molecular branching. *Journal of American Chem. Soc.* **1975**, 97, 23-38.
36. Walpole, M. *Probabilidad y Estadística*. Cuarta Edición, McGraw - Hill/ interamericana, México D.F. **1992.**
37. Well, A. F. *Química orgánica estructural*. Cuarta Edición, Editorial Reverte, España. **1992**.
38. WinNN32_ Windows. Neural Network for Windows 95/NT. **2001.**