Algoritmos basados en álgebra tensorial para la caracterización geométrica de moléculas orgánicas. Aplicación a la predicción de actividad biológica.

Autor principal
César Raúl García Jacas[1].

Otros autores
Yovani Marrero Ponce[2], Stephen J. Barigye[3], Liesner Acevedo Martínez[1], Ernesto Contreras Torres[1]

Colaboradores
DrC. Néstor Cubillán[4], DrC. Oscar Miguel Rivera Borroto[5], DrC. Harold Ariza Rico[4], DrC. Ysaías J. Alvarado[6], DrC. Huong Le-Thi-Thu[7], MSc. Longendri Aguilera Mendoza[1], MSc. José Ricardo Valdés Martiní[2], MSc. Mario Pupo Meriño[1], MSc. Lisset Cabrera Leyva[8], MSc. Taymara Hernández Ortega[1], Ing. Ricardo W. Pino Urias[9].

Entidades ejecutoras principales
[1]Universidad de las Ciencias Informáticas, La Habana, Cuba.

Entidades participantes
[2]Universidad Tecnológica de Bolívar, Cartagena de Indias, Colombia.
[3]Universidad Federal de Lavras, Brasil.
[4]Universidad del Zulia, Maracaibo, República Bolivariana de Venezuela.
[5]Pontificia Universidad Católica del Ecuador, Ecuador.
[6]Instituto Venezolano de Investigaciones Científicas (IVIC), Maracaibo, República Bolivariana de Venezuela.
[7]Escuela de Medicina y Farmacia, Universidad Nacional de Vietnam, Hanoi, Vietnam.
[8]Universidad de Camagüey, Camagüey.
[9]Universidad Central “Marta Abreu” de Las Villas, Santa Clara, Villa Clara, Cuba.

Autor para correspondencia
DrC. César Raúl García Jacas
Grupo de Investigación de Bioinformática,
Centro de Estudios de Matemática Computacional
Facultad 6, Universidad de las Ciencias Informáticas
La Habana, Cuba
crjacas@uci.cu

Resumen

La caracterización geométrica de las estructuras moleculares constituye un enfoque necesario en el diseño de fármacos asistido por computadora para establecer una relación entre las características de las moléculas y su correspondiente propiedad o actividad biológica. Con este propósito son utilizados varios algoritmos reportados en la literatura que extraen representaciones numéricas (descriptores moleculares, DMs) a partir de la información geométrica de las moléculas. Sin embargo estos procedimientos definidos hasta la fecha solo codifican información para relaciones entre pares átomos y/o consideran únicamente la distancia Euclidiana para este fin, a pesar de que la actividad o propiedad de los compuestos puede depender de las relaciones entre más de dos átomos y que no existe postulado teórico donde se demuestre que la distancia Euclidiana es la más adecuada para relacionar dos átomos de una molécula. Por lo tanto en la presente investigación se proponen por primera vez nuevos algoritmos para obtener descriptores moleculares geométricos independientes de alineamiento que codifiquen información para relaciones entre dos, tres y cuatro átomos mediante el uso de diferentes métricas para relaciones entre pares de átomos (e.g. Canberra, Soergel, Separación Angular), así como métricas ternarias (e.g. Ángulo de enlace) y cuaternarias (e.g. Ángulo diedro) para codificar relaciones entre tres y cuatro átomos, respectivamente. Estos nuevos algoritmos están basados en las formas algebraicas 2-lineales, 3-lineales y 4-lineales como casos específicos de las formas algebraicas N-lineales y emplean las kth matrices espaciales 2-tuplas, 3-tuplas y 4-tuplas de similitud-disimilitud, definidas en esta investigación, para representar la información química para las relaciones entre dos, tres y cuatro átomos de una molécula. Además se proponen varias transformaciones para normalizar las representaciones matriciales definidas y se introducen nuevas estrategias para considerar relaciones inter-atómicas de interés. Por último se define un procedimiento que calcula los DMs a partir de su descomposición a nivel atómico utilizando varios operadores de agregación. Para calcular estos DMs se desarrolló el software QuBiLS-MIDAS el cual aprovecha las arquitecturas multi-núcleos actuales y utiliza el sistema de cómputo distribuido T-arenal (introducido en esta investigación) para disminuir el tiempo de procesamiento. Diferentes estudios basados en Análisis de Variabilidad y Análisis de Componente Principales demostraron que los nuevos algoritmos calculan DMs que caracterizan mejor compuestos estructuralmente diferentes y codifican información ortogonal con respecto a otros enfoques definidos. Por último los algoritmos propuestos se utilizaron para determinar DMs con el propósito de evaluar su utilidad en la predicción de actividad biológica. Para este fin se utilizaron ocho bases de compuestos químicos y se construyeron modelos predictivos basados en la técnica Regresión Lineal Múltiple (RLM). Los resultados alcanzados son estadísticamente superiores a los reportados en la literatura donde se consideraron modelos basados en técnicas más complejas que RLM. Por lo tanto puede concluirse que los nuevos algoritmos constituyen un valioso aporte al conocimiento científico en el campo de la informática-química para ser aplicados en el diseño de nuevos fármacos. Entre los principales avales están la publicación de 5 artículos en revistas científicas de alto factor de impacto y la presentación en congresos nacionales e internacionales.

Comunicación corta

1. Introducción

La Informática-Química es la disciplina que se dedica a la extracción y análisis de la información contenida en las estructuras químicas con el propósito de obtener conocimiento que pueda ser utilizado para guiar la identificación, diseño y/o optimización de fármacos [1,2]. Para extraer dicha información se emplean algoritmos que codifican las propiedades o características moleculares en valores numéricos, conocidos como descriptores moleculares (DMs) [3]. Entre los diferentes tipos de DMs se encuentran los geométricos (3D-DMs), los cuales brindan información relacionada con la representación tridimensional de la estructura molecular [4–7].

Los 3D-DMs pueden clasificarse en dependientes y libres de alineamiento [8]. Los primeros consideran la geometría del receptor o la información relacionada con otro compuesto de referencia o un farmacóforo [9–15], mientras los segundos al no considerar esta cualidad son invariantes a la rotación y traslación de las estructuras moleculares [16–18]. En su generalidad estos últimos 3D-DMs se definen a partir de la métrica Euclidiana para calcular la distancia entre pares de átomos, o a partir de la matriz de distancia geométrica y sus derivaciones. Otra característica es que únicamente establecen relaciones entre pares de átomos y no entre N de ellos para codificar información química.

Los aspectos antes mencionados pueden ser considerados deficiencias de los actuales 3D-DMs debido a: 1) que el proceso de alineamiento por lo general no es una característica deseable por no contarse siempre con la información del compuesto de referencia o un farmacóforo, y que en su mayoría solo se aplica en compuestos cogenéricos; 2) que en campos investigativos actuales una de las tareas primarias es la identificación de la función de distancia más adecuada para el problema en cuestión [19–23] y por consiguiente otras métricas distintas a la Euclidiana pueden ser usadas para el cálculo de la distancia inter-atómica; y 3) que solo se consideran relaciones entre pares de átomos a pesar de que la actividad o propiedad de los compuestos puede depender de las interacciones entre varios de ellos [24].

Por lo tanto esta investigación tiene como objetivo proponer nuevos algoritmos para calcular 3DDMs independientes de alineamiento, a partir de la utilización de varias métricas para el cálculo de la distancia inter-atómica y del establecimiento de relaciones entre dos o más átomos, que permitan codificar mayor información relevante de las estructuras químicas y desarrollar modelos con mejor poder predictivo.

2. Resultados

2.1. Definición de los 3D-DMs basados en álgebra lineal y multi-lineal

2.1.1. Vector molecular

El enfoque de vector molecular (tensor de orden 1) basado en átomos como representación de las estructuras químicas orgánicas de pequeño y mediano tamaño se ha explicado en detalle en varios reportes [25–27]. Los componentes de un vector molecular son valores numéricos que representan cierta propiedad atómica. En este trabajo las propiedades químicas utilizadas son: 1) masa atómica, volumen de van der Waals, polarizabilidad, electronegatividad en la escala de Pauling, Ghose-Crippen LogP, carga atómica de Gasteiger-Marsili, superficie de área polar, refractividad, dureza (hardness) y suavidad (softness).

2.1.2. Matriz espacial N-tuplas de similitud-disimilitud: nueva representación geométrica de moléculas orgánicas

En el presente trabajo para la codificación de la información 3D de las estructuras químicas de moléculas orgánicas se proponen la k^{th} matriz espacial total 2-tuplas de similitud-disimilitud (GB^k), la k^{th} matriz espacial total 3-tuplas de similitud-disimilitud (GT^k) y la k^{th} matriz espacial total 4-tuplas de similituddisimilitud (GQ^k) para las relaciones entre dos, tres y cuatro átomos de una molécula, respectivamente [28–30]. El índice superior k indica la potencia a la cual se elevan las matrices. De esta manera, para k = 1 los coeficientes gb^1_{ij} , gt^1_{ij} y gq^1_{ij} correspondientes a las matrices GB^1, GT^1 y GQ^1 representan la información de todas las interacciones entre dos, tres y cuatro átomos, respectivamente. Los coeficientes de la diagonal pueden tener asignados dos valores: 1) el número de electrones no pareados del átomo en cuestión, o 2) la distancia, D_{io}, para cada átomo i y el centro de la molécula, *o*.

La información química representada en los enfoques matriciales anteriores es obtenida para relaciones entre pares de átomos mediante métricas como la de Soergel, Canberra, Clark, Lance-Williams, Separación Angular, entre otras. De esta forma la matriz GB^1 constituye una generalización de la matriz de distancia geométrica empleada hasta la fecha y la cual solo es calculada a partir de la distancia Euclidiana. Por otro lado con el uso de multi-métricas ternarias como Área y Ángulo de enlace y multimétricas cuaternarias como Volumen y Ángulo diedro se obtiene información para relaciones entre tres y cuatro átomos, respectivamente.

Las matrices GB^k, GT^k y GQ^k para k ≥ 2, se calculan mediante el producto Hadamard. El exponente k puede tener tanto valores positivos como negativos. Esto significa que cuando el parámetro k es un número negativo el recíproco se calcula para cada una de las entradas de las matrices espaciales totales N-tuplas. El máximo valor de k, ±12, está en correspondencia con las interacciones asociadas con la forma funcional del potencial 6 - 12 de Lennard-Jones.

Por otro lado, en este trabajo para considerar la representación de la información correspondiente a determinados tipos de átomos (o fragmentos químicos) se proponen las k^{th} matrices espaciales de fragmento-local 2-tuplas, 3-tuplas 4-tuplas de similitud-disimilitud, representadas por GB^k_F , GT^k_F y GQ^k_F , respectivamente, donde F es el fragmento en cuestión. Los tipos de átomos o fragmentos, F, considerados en este trabajo son: aceptores de enlaces de hidrógeno (A), átomos de carbono en cadenas alifáticas (C), donadores de enlaces de hidrógeno (D), halógenos (G), grupos metilos terminales (M), átomos de carbono en porciones aromáticas (P) y heteroátomos (X).

También con el objetivo de considerar las interacciones más importantes acorde un criterio seguido se propone el Corte molecular N-tuplas grafo-teórico basado en distancia topológica y el Corte molecular N-tuplas geométrico basado en distancia Euclidiana, respectivamente. La aplicación de uno o ambos cortes moleculares permite la creación de las k^{th} matrices espaciales cociente de vecindad totales (o de fragmento-local) 2-tuplas, 3-tuplas y 4-tuplas de similitud-disimilitud, representadas por $V\ GB^k_{(F)}$, $V\ GT^k_{(F)}$ y $V\ GQ^k_{(F)}$, respectivamente. También, se definen un conjunto de Cortes moleculares N-tuplas geométricos basados en medidas ternarias o cuaternarias para considerar solamente aquellas relaciones entre tres (N = 3) y cuatro (N = 4) átomos cuyos valores se encuentran en un intervalo determinado.

2.1.3. Descriptores moleculares 3D N-lineales

A partir de la representación tensorial (secciones 2.1.1 y 2.1.2) de la estructura química de moléculas orgánicas y la definición matemática de formas algebraicas, entonces se pueden definir los nuevos 3DDMs algebraicos libres de alineamiento (3D-DMs N-lineales). Por lo tanto, si una molécula contiene n átomos, entonces, los k^{th} descriptores 2-lineales (relaciones entre 2 átomos), 3-lineales (relaciones entre 3 átomos) y 4-lineales (relaciones entre 4 átomos) se calculan como formas algebraicas N-lineales en R^n sobre un conjunto de vectores base canónicos, y se expresan por las siguientes ecuaciones, respectivamente:

$$
\begin{aligned}
{}_bL_a = {}^{(V)}b^{a,k}_{(F)}(\bar{x},\bar{y}) &= \sum_{i=1}^{n}\sum_{j=1}^{n}{}^{(V)}_{ns[ss,ds,mp]}gb^{a,k}_{ij(F)}x^i y^j \\
&= [X]^T\,{}^{(V)}_{ns[ss,ds,mp]}GB^{a,k}_{(F)}\,[Y]
\end{aligned}
\tag{1}
$$

$$ {}_{tr}L_a = \overset{(V)}{}_{ns[ss,mp]} tr^{a,k}_{(F)}(\bar{x},\bar{y},\bar{z}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} \overset{(V)}{}_{ns[ss,mp]} gt^{a,k}_{ijl(F)} x^i y^j z^l = \overset{(V)}{}_{ns[ss,mp]} GT^{a,k}_{(F)} \cdot \bar{x} \cdot \bar{y} \cdot \bar{z} \quad (2) $$

$$ {}_{qu}L_a = \overset{(V)}{}_{ns[ss,mp]} qu^{a,k}_{(F)}(\bar{x},\bar{y},\bar{z},\bar{w}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{h=1}^{n} \overset{(V)}{}_{ns[ss,mp]} gq^{a,k}_{ijlh(F)} x^i y^j z^l w^h = \overset{(V)}{}_{ns[ss,mp]} GQ^{a,k}_{(F)} \cdot \bar{x} \cdot \bar{y} \cdot \bar{z} \cdot \bar{w} \quad (3) $$

donde, n es el número de átomos de la molécula, "a" indica el átomo analizado ($a = 1, \ldots, n$), y los coeficientes $\overset{(V)}{}_{ns[ss,ds,mp]} gb^{a,k}_{ij(F)}$, $\overset{(V)}{}_{ns[ss,mp]} gt^{a,k}_{ijl(F)}$ y $\overset{(V)}{}_{ns[ss,mp]} gq^{a,k}_{ijlh(F)}$ son las entradas pertenecientes a las matrices de *nivel atómico* $\overset{(V)}{}_{ns[ss,ds,mp]} GB^{a,k}_{(F)}$, $\overset{(V)}{}_{ns[ss,mp]} GT^{a,k}_{(F)}$ y $\overset{(V)}{}_{ns[ss,mp]} GQ^{a,k}_{(F)}$, respectivamente. Por otro lado, $x^1, \ldots, x^n$, $y^1, \ldots y^n$, $z^1, \ldots z^n$ y $w^1, \ldots w^n$ son las componentes de los *vectores moleculares* $\bar{x}$, $\bar{y}$, $\bar{z}$ y $\bar{w}$. Entonces, la obtención de los *3D-DMs N-lineales* a partir de la contribución de cada átomo es generalizada como se muestra en las siguientes ecuaciones:

$$ \overset{(V)}{}_{ns[ss,ds,mp]} b^k_{(F)}(\bar{x},\bar{y})[\overset{(V)}{}_{ns[ss,mp]} tr^k_{(F)}(\bar{x},\bar{y},\bar{z}), \overset{(V)}{}_{ns[ss,mp]} qu^k_{(F)}(\bar{x},\bar{y},\bar{z},\bar{w})] = OPER({}_bL_a[{}_{tr}L_a, {}_{qu}L_a]) \quad (4) $$

donde, $OPER$ es el *operador de agregación* utilizado sobre el vector $\bar{L}$ de *descriptores por nivel atómico*.

2.2. Software para el cálculo de los 3D-DMs N-lineales

Para el cálculo de los 3D-DMs N-lineales se desarrolló el software QuBiLS MIDAS [31] utilizando el lenguaje de programación Java en su versión 1.7. Esta aplicación consiste de dos componentes principales: el front-end y el back-end. En el front-end se encuentran implementadas las interfaces gráficas de usuario para la configuración de los DMs, mientras en el back-end están definidas las clases responsables de realizar los respectivos cálculos. Los algoritmos propuestos se implementaron para aprovechar arquitecturas multi-núcleos y distribuidas para mejorar la velocidad durante el procesamiento.

Para analizar el comportamiento durante el cálculo multi-núcleo un total de 20 280 3D-DMs 2-lineales, 3-lineales y 4-lineales se calcularon para el conjunto PrimScreen1 (http://www.otavachemicals.com/) compuesto por 1 000 estructuras. Como resultado se tiene que el tiempo total de procesamiento siempre disminuye en la medida en que se incrementa el número de procesadores y por lo tanto se utiliza apropiadamente la arquitectura multi-núcleo (ver Tabla 1).

No obstante a los resultados anteriores el enfoque multi-núcleo no es suficiente para satisfacer la demanda computacional para el cálculo de 3D-DMs N-lineales sobre un gran número de compuestos. Por tal motivo se empleó la computación distribuida como otra alternativa de cálculo haciendo uso de la plataforma T-arenal [32], la cual se desplegó en 337 estaciones de trabajo pertenecientes a los laboratorios docentes de la Universidad de las Ciencias Informática (UCI). La experimentación realizada está basada en el cálculo de 12 480 y 7 488 3D-DMs 2-lineales y 3-lineales

respectivamente. Para este fin se utilizó la base PrimScreen15 (http://www.otavachemicals.com/) conformada por 15 000 estructuras químicas. Como resultado de esta experimentación se tiene que el tiempo secuencial es reducido desde 49 349 segundos (13 horas) a 950 segundos (16 minutos) y desde 166 017 segundos (46 horas) a 2 783 segundos (46 minutos) en el cálculo de los 3D-DMs 2-lineales y 3-lineales, usando como máximo 265 y 282 clientes respectivamente.

Tabla 1: Resultado del cálculo multi-núcleo de los descriptores 3D N-lineales.

Procesadores	**Tiempo de cálculo promedio (TCP) [seg]**	**Speedup**	**Eficiencia**
	QuBiLS MIDAS 2-lineales		
1	2139	1.000	1.000
4	932	2.295	0.574
16	383	5.584	0.349
	QuBiLS MIDAS 3-lineales		
1	31847	1.000	1.000
4	9079	3.508	0.877
16	4633	6.874	0.430
	QuBiLS MIDAS 4-lineales		
1	639006	1.000	1.000
4	213830	2.988	0.747
16	106370	6.007	0.375

2.3. Predicción de actividad biológica de moléculas orgánicas

Los algoritmos propuestos se aplicaron en la predicción de actividad biológica con el propósito de evaluar la utilidad de los 3D-DMs que calculan. Para este fin se utilizaron ocho bases de compuestos químicos conformadas entre 66 y 397 estructuras. Las coordenadas 3D de las moléculas y los conjuntos de entrenamiento y prueba considerados están en correspondencia con los usados en la literatura con el propósito de garantizar comparabilidad de resultados [33–38].

Para el desarrollo de los modelos predictivos (QSAR) se utilizó la técnica estadística Regresión Lineal Múltiple (RLM) [39], la cual se acopló con la metaheurística Algoritmo Genético (AG) como estrategia de búsqueda y optimización de los correspondientes modelos QSAR. Este procedimiento (RLM + AG) está implementado en el software MobyDigs (versión 1.0) [40] el cual se utilizó para realizar este estudio.

Para cada uno de los conjuntos químicos considerados se construyeron los mejores modelos de 3 a 9 variables para la correspondiente actividad biológica usando como función objetivo el parámetro Q^2_{loo} (validación cruzada dejando-uno-afuera).

Los resultados de los mejores modelos QSAR construidos se compararon con respecto a 12 procedimientos reportados en la literatura [33–38]. En la Tabla 2 se muestran las comparaciones acorde a la predicción externa lograda por cada enfoque considerado, donde los resultados obtenidos en esta investigación son estadísticamente superiores a los previamente reportados.

Tabla 2: Comparación de los resultados obtenidos en predicción externa de los modelos QuBiLS MIDAS con respecto a otros enfoques reportados en la literatura.

	ACE	**AchE**	**BZR**	**COX2**	**DHFR**	**GPB**	**THER**	**THR**
QuBiLS MIDAS[1]	0.7422	0.6309	0.5692	0.4932	0.6405	0.8283	0.7248	0.7674
CoMFA[a,c]	0.49	0.47	0.00	0.29	0.59	0.42	0.54	0.63
COMSIA basic[a,c]	0.52	0.44	0.08	0.03	0.52	0.46	0.36	0.55
COMSIA extra[a,c]	0.49	0.44	0.12	0.37	0.53	0.59	0.53	0.63
EVA[b,c]	0.36	0.28	0.16	0.17	0.57	0.49	0.36	0.11
HQSAR[b,c]	0.30	0.37	0.17	0.27	0.63	0.58	0.53	-0.25
2D[b,c]	0.47	0.16	0.14	0.25	0.47	-0.06	0.14	0.04
2.5D[b,c]	0.51	0.16	0.20	0.27	0.49	0.04	0.07	0.28
O3Q[a,d]	0.69	0.67	0.17	0.32	0.60	0.50	0.51	0.67
O3A/O3Q[a,d]	0.54	0.65	0.24	0.28	0.53	0.41	-0.18	0.30
O3QMFA[a,e]	0.45	0.61	0.13	0.37	0.59	0.29	0.49	0.60
COSMOsar3D[a,e]	0.62	0.61	0.13	*0.43*	0.58	0.63	0.59	0.66
2D-FPT[b,f]	*0.713*[l]	*0.714*[n]	*0.378*[l]	0.329[n]	*0.683*[n]	*0.667*[l]	*0.649*[l]	*0.737*[n]

[1] : modelos QuBiLS MIDAS basados en RLM

[a] : métodos QSAR dependientes de alineamiento; [b] : métodos QSAR libres de alineamiento; [c] : referencia [33]; [d] : referencia [34]; [e] : referencia [38]; [f] : referencia [35]; [l] : modelos 2D-FPT de términos lineales; [n] : modelos 2D-FPT de términos no lineales

3. Conclusiones

Atendiendo a los resultados obtenidos se llegó a las siguientes conclusiones:

1. Se definieron:
   a. nuevos 3D-DMs basados en conceptos del álgebra lineal haciendo uso de métricas diferentes a la Euclidiana para el cálculo de distancias inter-atómicas;
   b. nuevos 3D-DMs basados en relaciones entre N (N = 3; 4) átomos mediante la aplicación de conceptos relacionados con el álgebra multi-lineal y la utilización de multi-métricas para codificar información química; y
   c. nuevos procedimientos de cortes moleculares basados en multi-métricas para el análisis de relaciones inter-atómicas específicas.
2. Se introdujo la Plataforma de Cómputo T-arenal para la gestión de los recursos computacionales disponibles en una red institucional con el propósito de realizar cálculos distribuidos.
3. Se implementaron eficientemente los algoritmos propuestos en el software QuBiLs MIDAS, el cual durante el procesamiento aprovecha las arquitecturas multi-núcleos actuales y puede utilizar varias estaciones de trabajo conectadas a una red local mediante el sistema T-arenal.
4. Se demostró con la aplicación de los nuevos descriptores en el problema de predicción de actividad biológica que estos permiten extraer mayor información

estructural de relevancia de moléculas orgánicas y por lo tanto posibilitan el desarrollo de modelos QSAR con mejor poder predictivo.

Referencias Bibliográficas

[1] Vogt M, Bajorath J. Chemoinformatics: A view of the field and current trends in method development. Bioorganic & Medicinal Chemistry. 2012;20(18):5317 – 5323.

[2] Varnek A, Baskin II. Chemoinformatics as a Theoretical Chemistry Discipline. Molecular Informatics. 2011;30(1):20–32.

[3] Todeschini R, Consonni V. Molecular Descriptors for Chemoinformatics. 2nd ed. WILEY - VCH; 2009.

[4] Cross S, Baroni M, Ortuso F, Alcaro S, Cruciani G. Disrupting ProteinProtein Interfaces Using GRID Molecular Interaction Fields. In: Disruption of ProteinProtein Interfaces. Springer Berlin Heidelberg; 2013. p. 61–82.

[5] Singh M, K Singh S, T Chhabria M, Vasu K, Pandya D. CoMFA and CoMSIA 3D QSAR Models for a Series of Some Condensed Thieno[2,3-d]pyrimidin-4(3H)-ones with Antihistaminic (H1) Activity. Medicinal Chemistry. 2013;9(3):389–401.

[6] Malla P, Kumar M. 3D-QSAR Studies on a Series of 2,4-Thiazolidinedione Derivatives: A Self-Organizing Molecular Field Analysis Approach to Design Novel PTP 1B Inhibitors. Medicinal Chemistry. 2013;9(6):828–845.

[7] Cruz VL, Martinez S, Ramos J, Martinez-Salazar J. 3D-QSAR as a Tool for Understanding and Improving Single-Site Polymerization Catalysts. A Review. Organometallics. 2014;33(12):2944–2959.

[8] Hopfinger A, Tokarski J. 3D-QSAR analysis. Practical Applications of Computer-Aided Design (Charifson PS, ed) New York: Marcel Dekker. 1997;105:164.

[9] Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. Journal of Medicinal Chemistry. 1985;28(7):849–857.

[10] Todeschini R, Moro G, Boggia R, Bonati L, Cosentino U, Lasagni M, et al. Modeling and prediction of molecular properties. Theory of grid-weighted holistic invariant molecular (G-WHIM) descriptors. Chemometrics and Intelligent Laboratory Systems. 1997;36(1):65 – 73.

[11] Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. Journal of the American Chemical Society. 1988;110(18):5959–5967.

[12] Robinson DD, Winn PJ, Lyne PD, Richards WG. Self-Organizing Molecular Field Analysis: A Tool for Structure Activity Studies. Journal of Medicinal Chemistry. 1999;42(4):573–583.

[13] Silverman BD, Platt D, Pitman M, Rigoutsos I. Comparative molecular moment analysis (CoMMA). Perspectives in Drug Discovery and Design. 1998;12-14(0):183–196.

[14] Klebe G. Comparative Molecular Similarity Indices Analysis: CoMSIA. In: 3D QSAR in Drug Design. vol. 3 of Three-Dimensional Quantitative Structure Activity Relationships. Springer Netherlands; 1998. p. 87–104.

[15] Jain AN, Koile K, Chapman D. Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark. Journal of Medicinal Chemistry. 1994;37(15):2315–2327.

[16] Mihalic Z, Nikolic S, Trinajstic N. Comparative study of molecular descriptors derived from the distance matrix. Journal of Chemical Information and Computer Sciences. 1992;32(1):28–37.

[17] Bogdanov B, Nikolic S, Trinajstic N. On the three-dimensional wiener number. Journal of Mathematical Chemistry. 1989;3(3):299–309.

[18] Bath PA, Poirrette AR, Willett P, Allen FH. The Extent of the Relationship between the Graph-Theoretical and the Geometrical Shape Coefficients of Chemical Compounds. Journal of Chemical Information and Computer Sciences. 1995;35(4):714–716.

[19] Kumar V, Chhabra J, Kumar D. Impact of Distance Measures on the Performance of Clustering Algorithms. In: Intelligent Computing, Networking, and Informatics. vol. 243 of Advances in Intelligent Systems and Computing. Springer India; 2014. p. 183–190.

[20] Xu Z, Xia M. Distance and similarity measures for hesitant fuzzy sets. Information Sciences. 2011;181(11):2128 – 2138.

[21] Mironica I, Ionescu B, Vertan C. The influence of the similarity measure to relevance feedback. In: Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European. IEEE; 2012. p. 1573–1577.

[22] Krawczyk B, Wozniak M. Influence of Distance Measures on the Effectiveness of One-Class Classification Ensembles. Applied Artificial Intelligence. 2014;28(3):258–271.

[23] Talevi A, Castro EA, Bruno-Blanch LE. Recent Studies on Similarity Measures and its Applications to Chemoinformatics and Drug Design. In: Recent Trends on QSAR in the Pharmaeutical Perceptions. Bentham Science Publishers; 2012. p. 272–297.

[24] Muller-Dethlefs K, Hobza P. Noncovalent Interactions: A Challenge for Experiment and Theory. Chemical Reviews. 2000;100(1):143–168.

[25] Garit JAC, Ponce YM, Torrens F. Atom-based 3D-chiral quadratic indices. Part 2: Prediction of the corticosteroid-binding globulinbinding affinity of the 31 benchmark steroids data set. Bioorganic & Medicinal Chemistry. 2006;14(7):2398 – 2408.

[26] Castillo JA Garit, Martinez O Santiago, Marrero Y Ponce, Casanola GM Martin, Torrens F. Atom-based non-stochastic and stochastic bilinear indices: Application to QSPR/QSAR studies of organic compounds. Chemical Physics Letters. 2008;464(1):107–112.

[27] Marrero Ponce Y, Torrens F, Garcia Domenech R, Ortega Broche SE, Zaldivar VR. Novel 2D TOMOCOMD-CARDD molecular descriptors: atom-based stochastic and non-stochastic bilinear indices and their QSPR applications. Journal of Mathematical Chemistry. 2008;44(3):650–673.

[28] Cubillán N, Marrero-Ponce Y, Ariza-Rico H, Barigye S, García-Jacas C, Valdes-Martini J, et al. Novel global and local 3D atom-based linear descriptors of the Minkowski distance matrix: theory, diversity variability analysis and QSPR applications. Journal of Mathematical Chemistry. 2015;53(9):2028–2064.

[29] Marrero-Ponce Y, García-Jacas CR, Barigye SJ, Valdés-Martiní JR, Rivera-Borroto OM, Pino-Urias RW, et al. Optimum Search Strategies or Novel 3D Molecular Descriptors: is there a Stalemate? Current Bioinformatics, Aceptado para publicación. 2015;.

[30] García-Jacas CR, Marrero-Ponce Y, Barigye SJ, Valdés-Martiní JR, Rivera-Borroto OM, Olivero-Verbel J. N-Linear Algebraic Maps for Chemical Structure Codification: A Suitable Generalization for Atom-pair Approaches? Current Drug Metabolism. 2014;15(4):441–469.

[31] García-Jacas CR, Marrero-Ponce Y, Acevedo-Martínez L, Barigye SJ, Valdés-Martiní JR, Contreras-Torres E. QuBiLS-MIDAS: A parallel free-software for molecular descriptors computation based on multilinear algebraic maps. Journal of Computational Chemistry. 2014;35(18):1395–1409.

[32] García-Jacas CR, Aguilera-Mendoza L, González-Pérez R, Marrero-Ponce Y, Acevedo-Martínez L, Barigye SJ, et al. Multi-Server Approach for High-Throughput Molecular Descriptors Calculation based on Multi-Linear Algebraic Maps. Molecular Informatics. 2015;34(1):60–69.

[33] Sutherland JJ, O'Brien LA, Weaver DF. A Comparison of Methods for Modeling Quantitative Structure-Activity Relationships. Journal of Medicinal Chemistry. 2004;47(22):5541–5554.

[34] Tosco P, Balle T. A 3D-QSAR-Driven Approach to Binding Mode and Affinity Prediction. Journal of Chemical Information and Modeling. 2011;52(2):302–307.

[35] Bonachera F, Horvath D. Fuzzy Tricentric Pharmacophore Fingerprints. 2. Application of Topological Fuzzy Pharmacophore Triplets in Quantitative Structure-Activity Relationships. Journal of Chemical Information and Modeling. 2008;48(2):409–425.

[36] Hinselmann G, Rosenbaum L, Jahn A, Fechner N, Zell A, et al. jCompoundMapper: An open source Java library and command-line tool for chemical fingerprints. Journal of Cheminformatics. 2011;3(1):3–17.

[37] Manchester J, Czerminski R. SAMFA: Simplifying Molecular Description for 3D-QSAR. Journal of Chemical Information and Modeling. 2008;48(6):1167–1173.

[38] Klamt A, Thormann M, Wichmann K, Tosco P. COSMOsar3D: Molecular Field Analysis Based on Local COSMO s-Profiles. Journal of Chemical Information and Modeling. 2012;52(8):2157–2164.

[39] Chatterjee S, Hadi AS. Regression analysis by example. 5th ed. John Wiley & Sons; 2013.

[40] Todeschini R, Consonni V, Mauri A, Pavan M. MobyDigs: software for regression and classification models by genetic algorithms. In: Nature-insprired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks. vol. 23 of Data Handling in Science and Technology. Elsevier; 2003. p. 141 – 167.

**Las publicaciones que avalan el presente trabajo son: [28], [29], [30], [31] y [32].