CIENCIA 15(3), 406 - 420, 2007
Maracaibo, Venezuela

SIMILARITY: Un programa para el cálculo de la similitud molecular cuántica

Luis Rincón

Grupo de Química Teórica, Departamento de Química. Universidad de Los Andes, La Hechicera, Mérida-5101, Venezuela

Recibido: 23-03-06 Ac eptado: 02-05-07

Resumen

En este artículo se describe SIMILARITY, un programa para evaluar la similitud entre dos moléculas, así como para obtener la orientación relativa que maximiza la similitud. En este trabajo se presentan las bases metodológicas y una descripción detallada de la manera en que fue implementada la similitud molecular. La versión actual del programa fue usada en la estimación de la constante σ de Hammett y el famoso problema de los 31 esteroides de Cramer.

Palabras clave: Alineamiento molecular; similaridad molecular; QSAR.

SIMILARITY: A program for calculation of the quantum molecular similarity

Abstract

This paper presents SIMILARITY, a program for quantitatively evaluating the similarity between molecules and assesses the relative orientation that maximizes their similarity. The methodological bases and a detailed description of how quantum molecular similarity was implemented in this program to handle molecular matching are presented. All current features of the program were applied to the estimation of the Hammett σ constant and the well-known Crammer 31 steroid set.

Key words: Molecular alignment; molecular similarity; QSAR.

Introducción

Este artículo describe la implementación y validación de un programa para el cálculo de la similitud molecular cuántica desarrollado en el Grupo de Química Teórica de la ULA y que lleva por nombre SIMILARITY. En términos muy generales, el objetivo de las medidas de similitud molecular es cuantificar cuan parecidas son dos moléculas, *A* y *B*. Las medidas de similitud molecular son de gran importancia en química farmacéutica donde se usan en el diseño racional de drogas para el cálculo de relaciones cuantitativas de estructura actividad (*QSAR: Quantitative Structure Activity Relationships*). El programa SIMILARITY forma parte de una estrategia computacional para cálculos de QSAR en 3-dimensiones (3D-QSAR) que actualmente se encuentra en desarrollo.

Debido a la naturaleza de lo que se trata de cuantificar, la similitud entre dos moléculas de ninguna forma puede considerarse como un observable que pueda ser contrasta-

Autor para la correspondencia. E-mail: lrincon@ula.ve

do empíricamente. Por esta razón, existe cierta flexibilidad (o arbitrariedad) en el momento de definirla y cuantificarla. Varias definiciones para el concepto de similitud molecular han aparecido en la literatura (1-8).

El cálculo de la similitud molecular depende crucialmente de la definición que sea usada. En este trabajo se uso la definición de similitud molecular cuántica (*SMC*) desarrollada por Carbo-Dorca y colaboradores (1, 2). En esta definición, la similitud molecular entre dos moléculas A y B, denotada como Z_{AB}, es calculada a partir de una integral que involucra la densidad electrónica de las dos moléculas que son compradas, $\rho_A(r_1)$ y $\rho_B(r_2)$, pesadas por un operador positivo de dos electrones, $\Omega(r_1, r_2)$,

$$Z_{AB} = \int\int \rho_A(r_1)\Omega(r_1, r_2)\rho_B(r_2)dr_1dr_2 \qquad [1]$$

Por construcción la similitud molecular cuántica de la ecuación [1] es una cantidad positiva, $Z_{AB} \in R^+$. Para un conjunto de n moléculas, los valores de la similitud molecular cuántica pueden ser representados en una matriz cuadrada y simétrica de orden n, $Z = [Z_{AB}]$. La integral de la ecuación [1], depende del operador Ω utilizado. Para el presente programa se han usado dos operadores: el de superposición y el Coulombico. A continuación se presentan algunas definiciones básicas de similitud molecular que serán usadas en este trabajo.

1. El Operador de Superposición

El operador mas usado en el calculo de la SMC es la función delta de Dirac, $\Omega = \delta(r_1 - r_2)$. Usando este operador, la ecuación [1] toma la forma:

$$Z_{AB} = \int\int \rho_A(r_1)\delta(r_1 - r_2)\rho_B(r_2)dr_1dr_2 \qquad [2]$$

$$= \int \rho_A(r)\rho_B(r)dr$$

Esta cantidad es una medida de la superposición entre dos densidades electrónicas.

2. El Operador Coulombico

Además del operador de superposición, el otro operador ampliamente usado es el operador de repulsión electrónica, $\Omega = |r_2 - r_1|^{-1}$. Usando este operador se tiene una medida electrostática de la similitud molecular,

$$Z_{AB} = \int\int \rho_A(r_1)|r_1 - _2|^{-1}\rho_B(r_2)dr_1dr_2 \qquad [3]$$

3. Medidas de Autosimilitud

La autosimilitud resulta de la comparación de dos densidades electrónicas idénticas,

$$Z_{AA} = \int\int \rho_A(r_1)\Omega(r_1, r_2)\rho_A(r_2)dr_1dr_2 \qquad [4]$$

Se puede notar que las medidas de autosimilitud corresponden a los elementos diagonales de la matriz de similitud molecular. Las medidas de autosimilitud han sido usadas recientemente en la identificación de sitios activos en farmacología (10).

4. Índices de Similitud

Una medida normalizada de la similitud molecular son los índices de similitud. En el programa desarrollado se implemento el índice de Carbo (1), C_{AB}, el cual se define como,

$$C_{AB} = \frac{Z_{AB}}{(Z_{AA}Z_{BB})^{1/2}} \qquad [5]$$

Para la definición de similitud molecular de la ecuación [1], este índice esta comprendido entre cero y uno, $C_{AB} \in (0,1)$. Al igual que las medidas de similitud molecular, los índices de similitud para un conjunto de n moléculas pueden ser representados en una matriz cuadrada y simétrica de orden n, $C = [C_{AB}]$. En este caso, la matriz *C* tendría unos en la diagonal, y los elementos no diagonales estarían entre uno y cero.

Dados que las densidades electrónicas pueden ser interpretadas como vectores de un subespacio de Hilbert definido positivo,

el índice de Carbo puede interpretarse como el coseno del ángulo entre dos densidades electrónicas, cos(α_{AB}).

5. Transformaciones Estocásticas

Adicionalmente a los índices de similitud, otra forma de normalizar las medidas de similitud molecular es usando una transformación estocástica (11, 12). Estas transformaciones son definidas de la forma,

$$S_{AB} = Z_{AB}\left(\sum_{C=1}^{N} Z_{AC}\right)^{-1} \quad [6]$$

A partir de esta definición se construye la matriz de transformación estocástica, *sincerely= [SAB]*. En esta matriz la suma de los elementos de cada columna es igual a 1. La transformación [6] produce un índice alternativo al descrito en la sección 4, que no es simétrico, $S_{AB} \neq S_{BA}$, y puede ser interpretada como una distribución discreta de probabilidades.

Descripción del programa

SIMILARITY es un programa para el calculo de las medidas de similitud molecular y los índices de similitud asociados, este programa esta dividido en dos rutinas: INITIALIZATION y ALIGNMENT.

1. Initialization

Como su nombre lo indica, esta es la rutina de inicialización, donde se leen las coordenadas de las moléculas que serán comparadas, y los parámetros para el cálculo de la densidad electrónica. En principio, las coordenadas pueden ser obtenidas a partir de datos de difracción de Rayos-X, o computacionalmente usando métodos *ab-initio*, semiempíricos o de mecánica molecular. Además de las coordenadas atómicas, el programa requiere que se suministren las cargas atómicas que serán usadas en el cálculo de la densidad.

En general, la evaluación de las integrales involucradas en la similitud molecular usando densidades electrónicas provenientes de cálculos de estructura electrónica (*ab-initio* o semiempiricos) resulta computacionalmente prohibitivas en un estudio típico de QSAR, donde el tiempo de calculo es un parámetro critico y usualmente involucran numerosas moléculas con el objeto de hacer un análisis estadístico. Por esta razón, para simplificar la evaluación de la densidad electrónica, y las integrales de la similitud molecular, resulta conveniente usar densidades aproximadas. El grupo de Girona ha desarrollado la aproximación ASA (*Atomic Shell Approximation*) (1), la cual fue implementada en el programa SIMILARITY. En la aproximación ASA, las densidades electrónicas moleculares son ajustadas mediante una combinación lineal de funciones gaussianas de simetría s (esférica). Específicamente, se usara la aproximación *promolecular*, que consiste en escribir la densidad como una suma de densidades atómicas,

$$\rho_A^{ASA}(r) = \sum_{a\in A} Z_a \rho_a^{ASA}(r) \quad [7]$$

En esta ecuación el índice a se refiere a todos los átomos que pertenecen a la molécula A, Z_a es el numero de electrones en el átomo a, el cual es calculado a partir de la carga de este átomo en la molécula A. La densidad de la ecuación [7] esta normalizada al numero de electrones de la molécula A,

$$\int \rho_A^{ASA}(r)dr = N_A \quad [8]$$

Las densidades atómicas de la ecuación [7] se escriben como una combinación lineal de gaussianas esféricas normalizadas,

$$\rho_a^{ASA}(r) = \sum w_i \left| g_i(r - r_a); \zeta_i \right|^2 \quad [9]$$

$$g_i(r - r_a; \zeta_i) = \left(\frac{2\zeta_i}{\pi}\right)^{3/4} \exp[-\zeta_i (r - r_a)^2] \quad [10]$$

Los pesos w_i, y los exponentes orbitales ξ_i, así como el numero de funciones gaussianas usadas en las expansión de la ecuación [9], son ajustadas tomando como referencia densidades atómicas obtenidas de cálculos

de estructura electrónica *ab-initio* de manera de minimizar el error cuadrático medio,

$$\varepsilon = \int \left|\rho_A(r) - \rho_A^{ASA}(r)\right|^2 dr \qquad [11]$$

Para asegurar que la densidad electrónica sea positiva en todo el espacio, los pesos w_i se restringen a que sean positivos. Los valores de los pesos y exponentes orbitales han sido tomados de los reportados en la literatura, y que están disponibles en la WEB (13). Estos parámetros han sido ajustados usando densidades atómicas *ab-initio* de diferente calidad, por ejemplo, se tiene que cuando se uso la base 3-21G se ajustaron los átomos desde el H hasta el Kr, para la base 6-311G se ajustaron desde el H hasta el Ar, y para la base de Huzinaga, que es intermedia a las dos anteriores, se tiene desde el H hasta el Rn. Estos parámetros son guardados en los archivos *3-21g.arc*, *6-311g.arc* y *huzinaga.arc*. Al principio del cálculo se debe indicar conjunto de parámetros serán usados.

Una de las ventajas de usar densidades aproximadas tipo ASA es que el calculo de las integrales de similitud molecular se simplifica considerablemente, ya que solo involucra funciones gaussianas esféricas. La similitud molecular toma la forma,

$$Z_{AB} = \sum_{a\in A}\sum_{b\in B}\sum_{i\in a}\sum_{j\in b} Z_a Z_b w_i w_j z_{ij} \qquad [12]$$

$$z_{ij} = \int\int \left|g_i(r_1 - r_i, \zeta_i)\right|^2 \Omega(r_1 - r_2) \left|g_j(r_2 - r_j, \zeta_j)\right|^2 dr_1 dr_2 \qquad [13]$$

Para el caso del operador de superposición se obtiene,

$$z_{ij} = \left(\frac{2\zeta_i\zeta_j}{\pi(\zeta_i + \zeta_j)}\right)^{3/2} \exp\left(\frac{-2\zeta_i\zeta_j\left|r_a - r_b\right|^2}{\zeta_i + \zeta_j}\right) \qquad [14]$$

mientras que para el operador Coulombico se obtiene,

$$z_{ij} = 4\left(\frac{\zeta_i\zeta_j}{2\pi(\zeta_i + \zeta_j)}\right)^{1/2} F_o[T] \qquad [15]$$

$$F_o[T] = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(2k+1)} T^k \qquad [16]$$

$$T = \frac{2\zeta_i\zeta_j}{\zeta_i + \zeta_j}\left|r_a - r_b\right| \qquad [17]$$

En general la expansión de la ecuación [16] converge rápidamente si el argumento **T** es pequeño. Sin embargo para valores de **T** mayores de 30au tiende a converger lentamente producto de una pobre cancelación de los términos de la sumatoria. Por esta razón, para valores de T mayores que 30 u.a. se aproximo con una función Lorentziana de la forma,

$$F_0[T] = \frac{F_0[0]}{1 + bT^2} \qquad [18]$$

donde $F_0[0]$ es el valor de la función en T=0, y el parámetro b se calcula para que la función F_0 sea continua en T=30 u.a.

2. Alignment

La similitud molecular entre dos moléculas diferentes, así como los índices de similitud, depende crucialmente de la orientación relativa de estas moléculas en el espacio. El problema de encontrar la orientación relativa óptima entre dos moléculas en el espacio, es conocido en la literatura como el problema del *ALINEAMIENTO*, y es un problema crucial en cualquier metodología de 3D-QSAR, como por ejemplo en el método CoMFA, el más popular de los métodos de 3D-QSAR (14).

Desde el punto de vista de la similitud molecular, un criterio natural para resolver el problema del alineamiento consiste en suponer que *dos moléculas están perfectamente alineadas si su similitud es máxima*. En otras palabras, si denotamos la orientación relativa entre dos moléculas como Θ, la

orientación óptima es aquella que maximiza la integral de similitud moléculas,

$$Z^{*}_{AB}(\Omega,\Theta) = \max_{\Theta} \int\int \rho_A(r_1)\Omega(r_1;r_2)\rho_B(r_2;\Theta)dr_1dr_2 \quad [19]$$

Para maximizar la similitud molecular entre dos moléculas (para alinearlas), se implemento el Algoritmo de la Tabla 1. Este algoritmo esta basado en dos metodologías discutidas y probadas en la literatura: el método de maximización global de Constans, Amat y Carbo (15) y el método de descenso rápido de gradientes de McMahon y King (16). Sin embargo, además de usar estas estrategias se han añadido algunas modificaciones, por esta razón, se explica brevemente el algoritmo de la Tabla 1.

En este algoritmo, las moléculas son alineadas a través de varias traslaciones y rotaciones sucesivas de manera de maximizar Z_{AB}. A partir de tres pares de "do-loop" anidados, se consideran dos conjuntos de tres átomos, {a,a',a"} y {b,b',b"}, en la molécula A y B, respectivamente. En el primer paso, que corresponde al par de lazos más externo, se superponen en el origen de coordenadas los átomos a y b, a partir de la traslación de las moléculas A y B. En un segundo paso, que corresponde al segundo par de lazos, los ejes {a a'} y {b b'}, de las moléculas A y B, respectivamente, son alineados a lo largo del eje X, a través de rotaciones sucesivas en los ejes Y y Z. Este segundo paso es ejecutados únicamente si la deferencia entre la distancia del segmento (a a'), $d_{aa'}$, y el segmento (b b'), $d_{bb'}$, es menor que cierto valor critico ε_1,

$$\varepsilon_1 \rangle \left| d_{aa'} - d_{bb'} \right| \quad [20]$$

En el par de lazos mas interno, a través de una rotación a lo largo del eje X, los átomos a" y b", en la molécula A y B, son alineados de manera que los tres átomos de cada molécula, {a,a',a"} y {b,b',b"}, estén en el plano XY. Este ultimo paso es ejecutado únicamente si la distancia final entre los átomos a" y b", $d_{a"b"}$, la cual viene dada por la ecuación,

$$d_{a"b"} = \sqrt{(d^x_a - d^x_b) + (d^y_a - d^y_b)^2} \quad [21]$$

$$d^x_a = \frac{d^2_{aa'} + d^2_{aa'} - d_{a'a"}}{2d_{aa'}} \quad [22]$$

$$d^y_a = \sqrt{d^2_{aa"} - (d^x_a)^2} \quad [23]$$

es menor que un valor critico ε_2,

$$\varepsilon_2 \rangle d_{a"b"} \quad [24]$$

Si el numero de átomos de la molécula A es n_A, y el de la molécula B es n_B, el numero de operaciones del algoritmo escala en la forma $n_A^3\, n_B^3$. Sin embargo, a partir de una selección adecuada de ε_1 y ε_2, la mayoría de estas evaluaciones pueden eliminarse, por otro lado, al menos los lazos mas externos de este algoritmo son trivialmente paralelizables, y de esta forma aprovechar el hecho de que actualmente se disponen en muchos laboratorios de computadoras con mas de un procesador. En muchos casos no es conveniente incluir todos los átomos de A y de B en la optimización global. En general el alineamiento se reduce a explorar un esqueleto común de átomos compartido por todas las moléculas, por ejemplo, al comparar dos péptidos es conveniente la superposición de los átomos de la cadena principal, y no de las cadenas laterales de los aminoácidos.

Orientados los átomos {a,a',a"} y {b,b',b"}, se calcula Z_{AB} para esa orientación, y si el valor de similitud obtenido es mayor que una fracción f_1 (típicamente 80%) del valor de Z_{AB} mas grande obtenido con otras orientaciones previas, se procede a minimizar la similitud usando el método de gradientes conjugados (17).

En esta optimización final, se deja fija la molécula A, y se varia la orientación relativa de la molécula B usando un vector de dimensión 6 que contiene el conjunto de los tres vectores que definen la traslación del

Tabla 1
Algoritmo de alineamiento y optimización de la similitud molecular Z_{AB}.
Este algoritmo esta comentado en detalle en la sección 2.

Z_{AB}= 0
DO for a∈A
DO for b∈B
Trasladar los átomos (a) y (b) al origen de coordenadas.
DO for a∈A\a
DO for b∈B\b
IF $\varepsilon_1 > |d_{aa'} - d_{bb'}|$ THEN
Alinear (aa') y (bb') en el eje X.
DO for a"?A\(a^a')
DO for b"?B\(b^b")
IF $\varepsilon_2 > d_{a"b"}$
Alinear (a,a',a") y (b,b',b") en el plano XY, calcular Z_{AB}.*
IF $Z_{AB}^{*} > f_i Z_{AB}$
Optimizar Z_{AB} por un método de gradientes.*
END IF
Z_{AB}= max{Z_{AB},Z_{AB}^{*}}
END IF
END DO b"
END DO a"
END IF
END DO for b'
END DO for a'
END DO for b
END DO for a

centro de masas de la molécula B, $\{t_x, t_y, t_z\}$, y el conjunto de los tres ángulos de Euler que definen las rotaciones de la molécula B en los ejes X, Y y Z, $\{\psi, \theta, \phi\}$,

$$\Omega = (t_x, t_y, t_z, \psi, \theta, \phi) \quad [25]$$

Usando los vectores anteriores, la distancia entre dos átomos a y b de la molécula A y B, respectivamente, se define de la forma,

$$r_{ab}(\Omega) = \left| r_a - (R_x(\psi)R_y(\theta)R_z(\phi))r_b^o + T(t_x, t_y, t_z) \right| \quad [26]$$

donde la posición del átomo b, depende de los operadores de rotación (R_X,R_Y,R_Z), del operador de traslación *T* y de su posición inicial $r_b^0 = \{x_b^0, y_b^0, z_b^0\}$. En términos del vector de traslación y los ángulos de Euler, la posición del átomo b viene dada de la forma,

$$x_b = (\cos\theta\cos\phi)x_b^o + (\sin\psi\sin\theta - \cos\psi\sin\phi)y_b^a + (\cos\psi\sin\theta\cos\phi + \sin\psi\sin\phi)z_b^o + t_x \quad [27]$$

$$y_b = (\cos\theta\cos\phi)x_b^o + (\sin\psi\sin\theta + \cos\psi\sin\phi)y_b^a + (\cos\psi\sin\theta\sin\phi + \sin\psi\sin\phi)z_b^o + t_y \quad [28]$$

$$z_b = -\sin\theta x_b^o + \sin\psi\cos\theta y_b^o + \cos\psi\cos\theta z_b^o + t_z \quad [29]$$

En el caso particular de Ω= 0, es decir, cuando no se ha realizado ninguna traslación ni rotación de la molécula B, el vector gradiente de la integral de similitud molecular de la ecuación [13] entre el átomo i de la molécula A y el átomo j de la molécula B, tiene una forma relativamente simple. En el caso del operador de superposición, el gradiente con respecto al vector Ω toma la forma,

$$g_\Omega^o = \left.\frac{\partial z_{ij}}{\partial r - \{ab\}}\right|_{\Omega=0} = -\frac{2\zeta_i\zeta_j}{\zeta_i + \zeta_j} z_{ij} \left.\frac{\partial r_{ab}^2}{\partial\Omega}\right| \quad [30]$$

mientras que para el operador Coulombico se obtiene,

$$g_\Omega^C = \left.\frac{\partial z_{ij}}{\partial r - \{ab\}}\right|_{\Omega=0} = -\frac{8}{\sqrt{2\pi}}\left(\frac{\zeta_i\zeta_j}{\zeta_i + \zeta_j}\right)^{3/2} F_o'[T_o]\frac{\partial r_{ab}^2}{\partial\Omega} \quad [31]$$

La derivada parcial de r_{ab}^2, dada por la ecuación [26], con respecto al vector de traslación, evaluada en Ω= 0, tiene la forma,

$$\left.\frac{\partial r_{ab}^2}{\partial t_x}\right|_{\Omega=0} = -2(x_a - x_b) \quad [32]$$

$$\left.\frac{\partial r_{ab}^2}{\partial t_y}\right|_{\Omega=0} = -2(y_a - y_b) \quad [33]$$

$$\left.\frac{\partial r_{ab}^2}{\partial t_z}\right|_{\Omega=0} = -2(z_a - z_b) \quad [34]$$

En el caso del vector de rotación, se obtiene que para los tres ángulos de Euler la derivada parcial evaluada en Ω= 0 tiene la forma,

$$\left.\frac{\partial r_{ab}^2}{\partial\psi}\right|_{\Omega=0} = -2(z_a y_b - y_a z_b) \quad [35]$$

$$\left.\frac{\partial r_{ab}^2}{\partial\theta}\right|_{\Omega=0} = -2(x_a z_b - z_a x_b) \quad [36]$$

$$\left.\frac{\partial r_{ab}^2}{\partial\phi}\right|_{\Omega=0} = -2(y_a x_b - x_a y_b) \quad [37]$$

Los valores de las derivadas de z_{ij} son usados para calcular la derivada de la similitud molecular. El método de gradientes conjugados (17) fue implementado para optimizar Z_{AB}, actualizando en cada iteración el vector Ω.

Para finalizar esta sección, se mencionara algunas de las diferencias del algoritmo de la Tabla 1 y el implementado por Constants, Amat y Carbo-Dorca en la Tabla I de la referencia (14). En primer lugar se puede notar que en el presente algoritmo se traslada y rota tanto la molécula B como la molécula A (con la excepción de la optimización final usando el método de gradientes conjugados donde se deja fija la molécula A). Por el contrario en el algoritmo de la referencia (14) solo se traslada y rota la molécula B, manteniendo A fija. En segundo lugar, en el presente algoritmo son comparadas las distancias entre los átomos de A y B para decidir si se calcula o no la similitud, en el algoritmo de la referencia (14) se compara la integral de similitud z_{ij}, la cual depende de la distancia y el tipo de átomos. En tercer lugar, en el presente algoritmo cuando las moléculas alineadas son similares se optimiza usando el método de gradientes conjugados, mientras que en la referencia (14) la optimización final es realizada después de comparar todas las moléculas alineadas, además en la optimización final de la referencia (14) se emplea el método de Newton donde se calculan los gradientes y la matriz Hessiana.

Ejemplos

En esta sección, se describen dos ejemplos donde se usa el programa SIMILARITY para calcular la similitud molecular. Estos ejemplos no buscan comparar en detalle el programa SIMILARITY con otros programas, mas bien, el objetivo es mostrar que los resultados obtenidos por este programa son comparables a estrategias similares que han

aparecido en la literatura. Los análisis estadísticos presentados en esta sección han sido realizados usando el programa STATISTIC desarrollado recientemente como herramienta de análisis de 3D-QSAR al igual que el programa SIMILARITY. STATIC permite realizar 3 tipos de cálculos estadísticos: 1) regresión lineal múltiple (MR), 2) análisis de componentes principales (PCA) y 3) "Partial Least Square" (PLS) (18). La correlación es calculada usando las medidas clásicas de índice de correlación, r^2, y desviación standard, s^2, mientras que la robustez del modelo estadístico es evaluada usando validación cruzada eliminando 1 variable, a partir del error cuadrático en la predicción del modelo, PRESS, y el índice de predicción de validación cruzada, q^2. En general el número de variables latentes en el modelo estadístico se escoge de manera de maximizar q^2 (o minimizar PRESS).

1. Correlación entre la autosimilitud cuantica y la constante σ de la Ecuación de Hammett

Como primer ejemplo del programa SIMILARITY, se estudio la correlación entre las medidas de autosimilaridad, ecuación [4], y la constante σ de la ecuación de Hammett, que refleja el efecto electrónico de los sustituyentes en posición para y meta de ácidos benzoicos sustituidos. Para este estudio se uso un conjunto de 29 ácidos benzoicos sustituidos en posición para y meta (uno no sustituido, R=H, 14 sustituidos en posición para y 14 en posición meta). Los 29 sustituyentes usados se muestran en la primera columna de la Tabla 2. Para cada uno de los ácidos benzoicos sustituidos se optimizo su estructura molecular y se calcularon las cargas atómicas mediante el método semi-empírico AM1 (19) usando el programa MOPAC-7 (20). En todos los casos se uso el método de Mulliken para calcular las cargas. A partir de la geometría y las cargas obtenidas, se determino la autosimilaridad usando únicamente los átomos del grupo carboxílico de estos ácidos. Este calculo se realizo usando la aproximación ASA, incluyendo en el calculo de la densidad únicamente los átomos del grupo *COOH*. La Tabla 2 muestra los resultados obtenidos para las medidas de autosimilaridad, en la columna 2 se muestra el valor de la constante de Hammett para los sustituyentes estudiados (21), las columnas 3-5 corresponden a los valores de la autosimilitud usando el operador de superposición, Z_O, con los parámetros correspondientes a las bases de huzinaga, 3-21G y 6-311G respectivamente. Mientras que la columna 6-8 corresponden a los valores usando el operador Coulombico, Z_C, con las mismas bases.

En la Tabla 3 se muestra el coeficiente de correlación, r^2, y la desviación estándar, s, para la correlación entre las medidas de autisomilitud calculadas y la constante σ de la ecuación de Hammett. De la Tabla 3 se puede observar que: 1) en todos los casos la correlación usando el operador de Superposición es estadísticamente superior a la obtenida usando el operador Coulombico, y 2) la correlación obtenida usando el mismo operador, pero diferentes parámetros para la densidad electrónica, son estadísticamente equivalente, es decir que la correlación depende únicamente del operador usado. Un análisis más detallado de las correlaciones obtenidas muestra que con ambos operadores los sustituyentes peor predichos son los mismos. En el caso del operador de Superposición se tiene: m-NO_2 (valor experimental= 0,51; predicho= 0,77) y p-CN (valor experimental= 0.66; predicho= 0,45), es de resaltar que estos sustituyentes tienen los valores mas grandes de la constante de Hammett. Los resultados de este estudio son similares a los encontrados por Ponec, Amet y Carbo-Dorca [22-23] usando el operador de superposición que es el que mejor se correlaciona con la constante de Hammett. El hecho de que el operador de superposición, y no el Coulombico, sea el que muestra la mejor correlación con la constante de Hammett, puede justificarse debido a que las medidas de autosimilaridad son un reflejo de la densidad de cargas en el grupo carboxílico, y esta bien establecido en trabajos previos que

Tabla 2
Medidas de Autosimilitud en ácidos benzoicos sustituidos.

R	σ_R	Z_O Huzinaga	Z_O 3-21G	Z_O 6-311G	Z_C Huzinaga	Z_C 3-21G	Z_C 6-311G
H	0,00	203,6030	203,2340	206,4039	360,3097	362,0627	360,8944
p-F	0,06	203,5250	203,1563	206,3249	360,1301	361,8823	360,7146
m-F	0,34	203,4055	203,0371	206,2036	359,9882	361,7341	360,5662
p-Cl	0,23	203,4972	203,1285	206,2966	360,1186	361,8710	360,7030
m-Cl	0,37	203,4467	203,0782	206,2455	360,0367	361,7888	360,6209
p-Br	0,23	203,4512	203,0827	206,2500	360,0577	361,8099	360,6419
m-Br	0,39	203,4468	203,0783	206,2456	360,0272	361,7792	360,6114
p-I	0,18	203,4382	203,0697	206,2368	360,0476	361,7997	360,6317
m-I	0,35	203,4636	203,0950	206,2626	360,0564	361,8085	360,6406
p-CH_3	-0,17	203,6446	203,2756	206,4461	360,3693	362,1224	360,9541
m-CH_3	-0,07	203,6180	203,2490	206,4191	360,3406	362,0938	360,9254
p-CF_3	0,54	203,2997	202,9316	206,0964	359,8178	361,5692	360,4014
m-CF_3	0,43	203,3322	202,9640	206,1294	359,7993	361,5505	360,3831
p-Et	-0,15	203,6435	203,2745	206,4450	360,3680	362,1212	360,9529
m-Et	-0,07	203,6212	203,2522	206,4224	360,3476	362,1008	360,9324
p-CN	0,66	203,3621	202,9938	206,1596	359,9183	361,6700	360,5021
m-CN	0,56	203,3834	203,0151	206,1813	359,8973	361,6488	360,4813
p-NO_2	0,78	203,1517	202,7840	205,9464	359,5681	361,3186	360,1511
m-NO_2	0,51	203,1897	202,8219	205,9849	359,5384	361,2886	360,1217
p-NH_2	-0,66	203,8441	203,4744	206,6484	360,6040	362,3575	361,1897
m-NH_2	-0,16	203,5916	203,2227	206,3923	360,3339	362,0872	360,9186
p-OH	-0,37	203,6809	203,3117	206,4829	360,3586	362,1114	360,9436
m-OH	0,12	203,5156	203,1469	206,3153	360,1778	361,9305	360,7622
p-OMe	-0,27	203,7115	203,3422	206,5139	360,4110	362,1640	360,9962
m-OMe	0,12	203,5638	203,1950	206,3642	360,2784	362,0314	360,8629
p-SMe	0,00	203,6639	203,2948	206,4657	360,3697	362,1227	360,9546
m-SMe	0,15	203,5896	203,2207	206,3903	360,2809	362,0339	360,8655
p-COOH	0,45	203,3466	202,9784	206,1440	359,9150	361,6668	360,4988
m-COOH	0,37	203,4094	203,0409	206,2076	359,9412	361,6929	360,5253

Tabla 3
Resultados de la correlación entre la autosimilitud molecular cuantica y la constante sincerely de la ecuación de Hammett

Operador	Base	R^2	S
Superposición	Huzinaga	0,8843	0,1123
	3-21G	0,8844	0,1124
	6-311G	0,8844	0,1122
Coulombico	Huzinaga	0,8398	0,1311
	3-21G	0,8395	0,1328
	6-311G	0,8400	0,1321

la constante de Hammett se correlaciona con la carga en el grupo carboxílico calculada usando métodos semi-empiricos (24-26).

2. Modelaje de la afinidad de un conjunto de 31 esteroides

El conjunto de los 31 esteroides mostrados en la Figura 1, conocidos en la literatura como el conjunto de Cramer (14), ha sido usado como "benchmark" para evaluar prácticamente todos los métodos de 3D-QSAR actualmente disponibles en la literatura (14, 27, 28). En esta sección se presentara resultados de la correlación entre la similitud molecular cuántica calculada para este conjunto usando el programa SIMILARITY y la constante de afinidad de estos esteroides a dos proteínas globulares. El logaritmo de la constante de afinidad de este conjunto de esteroides a la CBG (corticosteroid binding globuline) y la TBG (testosterone binding globuline) es mostrada en la Tabla 4 (14). Como se puede observar de la Tabla 4, para los 21 primeros esteroides se dispone tanto de la afinidad a la CBG como a la TBG, este conjunto de esteroides será usado para ajustar el modelo estadístico y se denominara el conjunto de entrenamiento; para los 10 últimos esteroides solo se dispone de la afinidad a la CBG y son usados como un conjunto de prueba para medir la capacidad predictiva del modelo obtenido con los datos

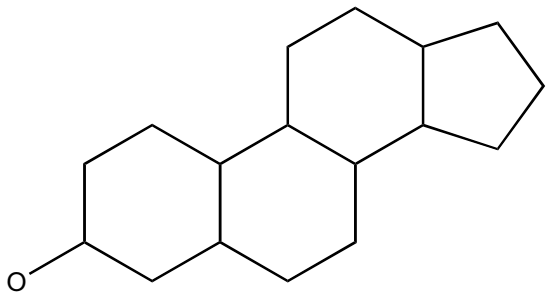

Figura 1. Conjunto de Cramer de 31 esteroides (14). Los esteroides 1-21 son usados como conjunto de entrenamiento para general el modelo estadístico. Los esteroides 22-31 son usados como conjunto de prueba para evaluar la capacidad predictiva del modelo.

de la CBG. Para cada uno de los esteroides se optimizo su estructura molecular, y se calcularon las cargas atómicas mediante el método semi-empírico AM1 (19) usando el programa MOPAC-7 (20). En todos los casos se uso el método de Mulliken para calcular las cargas. A partir de la geometría y las cargas obtenidas, se procedió a alinear todos los pares de esteroides diferentes superponiendo únicamente los 18 átomos de la Figura 2, los cuales son compartidos por las 31 molé-

Tabla 4
Constantes de afinidad, Log(Ka), de los esteroides usados en este trabajo.

N	Esteroide	CBG	TBG
1	Aldosterona	-6,279	-5,322
2	Androstanediol	-5,000	-9,114
3	Androstenediol	-5,000	-9,176
4	Androstenedion	-5,763	-7,462
5	Androsterona	-5,613	-7,146
6	Corticosterona	-7,881	-6,342
7	Cortisol	-7,881	-6,204
8	Cortisona	-6,892	-6,431
9	Dehidroepiandrosterona	-5,000	-7,819
10	Deoxicorticosterona	-7,653	-7,380
11	Deoxicortisol	-7,881	-7,204
12	Dehidrotestosterona	-5,919	-9,740
13	Estradiol	-5,000	-8,833
14	Estriol	-5,000	-6,633
15	Estrona	-5,000	-8,176
16	Etiocolanona	-5,225	-6,146
17	Pregnenolona	-5,225	-7,146
18	Hidroxipregnolona	-5,000	-6,362
19	Progesterona	-7,380	-6,944
20	Hidroxiprogesterona	-7,740	-6,996
21	Testosterona	-6,724	-9,204
22	Prednisolona	-7,512	–
23	Cortisolacetato	-7,553	–
24	4-pregnene-3,11,20-trione	-6,779	–
25	Epicortecosterona	-7,200	–
26	19-nortestosterona	-6,144	–
27	16,17–dihidroxiprogesterona	-6,247	–
28	17-metilprogesterona	-7,120	–
29	19-norprogesterona	-6,817	–
30	2-metilcortisol	-7,688	–
31	2-metil-9-fluorcortisol	-5,797	–

culas de la serie. Este calculo se realizo usando la aproximación ASA con el conjunto de parámetros de tipo *huzinaga*. La alineación se realizo por separado en el caso del operador de superposición y el operador Coulombico. A partir de la similitud molecular calculada en el proceso de alineamiento se calcularon los índices de similitud, C_{AB}. Las columnas de la matriz C_{AB} para el conjunto de entrenamiento (esteroides 1-21) fueron usadas para obtener un modelo estadístico que describa la constante de afinidad de los esteroides usando el método PLS (18). El modelo estadístico fue obtenido por separado para los datos de CBG y TBG, así como usando la matriz C_{AB} obtenida del operador de superposición y el operador Coulombico. La robustez del modelo estadístico es mostrado en la Tabla 5, en la cual se describe el tipo de descriptor usado (Superposición o Coulombico), la desviación standard del modelo de validación cruzada, s^2, el coeficiente de predicción, q^2, y el numero de variables latentes usadas en el modelo (los grados de libertad del modelo estadístico). Por comparación se incluyen los resultados originales de Cramer con el mismo conjunto de moléculas usando el método CoMFA/PLS con tres tipos de descriptores (electrostático, estérico, campo total = electrostático+estérico) (18). En el caso del CBG se obtuvo que el modelo obtenido con el operador de superposición es ligeramente superior al obtenido con el operador Coulombico y estadísticamente similar al modelo CoMFA con el campo de fuerza estérico, sin embargo este modelo es de inferior calidad que el obtenido usando CoMFA con el campo total de fuerza. En el caso de la afinidad a TBG el modelo obtenido con el operador Coulombico resulto ser mucho más robusto que el obtenido con el operador de superposición y de mejor calidad que el obtenido con CoMFA con el campo electrostático. Este análisis indicaría que el operador de superposición esta describiendo la información incluida en el campo molecular estérico del método CoMFA, mientras que el operador Coulombico describe la información en el descriptor asociado con el campo molecular electrostático. La conclusión preliminar descrita anteriormente, que se tendría que validar con otros ejemplos, permitiría establecer un vínculo entre los resultados del método de CoMFA (14) y los estudios de similitud molecular.

En la última columna de la Tabla 5 se muestra el análisis estadístico para la predicción de los 10 esteroides del conjunto de prueba usando el modelo obtenido para el caso de la afinidad a la CBG. En este conjunto hay un esteroide que es muy difícil de predecir ya que contiene un átomo de fluor, el 2-metil-9a-fluorocortisol [31], por esta razón la predicción se realizo en presencia y en ausencia de este punto particular. Como se observa de la Tabla 5 la desviación estándar cuando se incluye la molécula [31] aumenta en aproximadamente un 20% en todos los ejemplos, tanto con la presente metodología como usando CoMFA. Al igual que lo que se obtuvo para el conjunto de entrenamiento, la predicción del operador superposición fue superior a la obtenida con el operador Coulombico, y de una calidad comprable a la obtenida con el campo estérico de CoMFA.

Conclusiones

En este trabajo se presentó una descripción de la implementación y validación de un programa de similitud molecular cuántica. Se describió en detalle la metodología empleada para el cálculo de la similitud en todas sus etapas: construcción de la densidad electrónica, cálculo de las integrales de superposición y Coulombica, y el algoritmo de alineamiento. Este programa produce resultados de calidad similar a otras herramientas que actualmente están disponibles, y puede ser usado junto con herramientas estadísticas convencionales para producir modelos de QSAR robustos con un costo computacional relativamente pequeño.

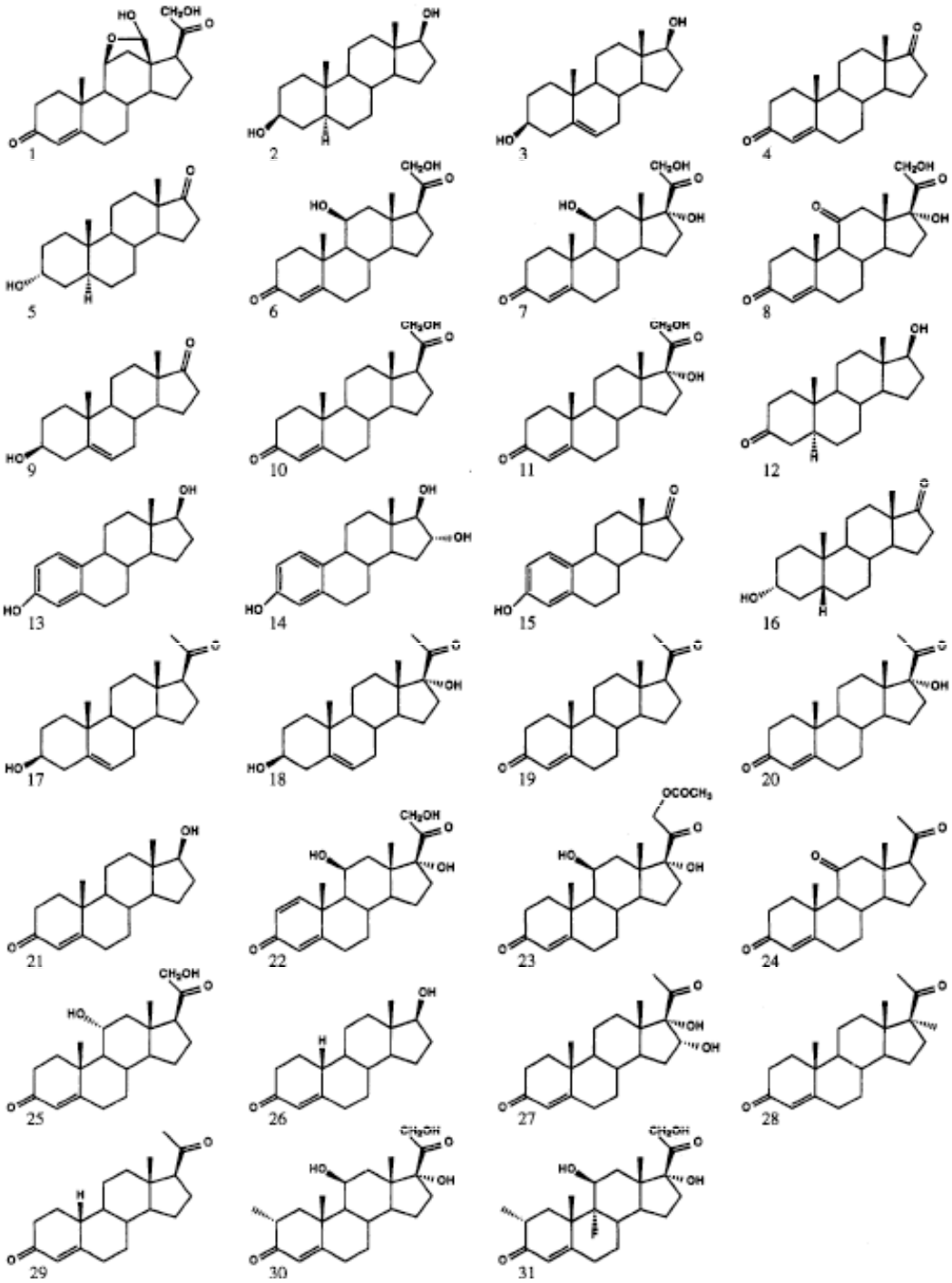

Figura 2. Esqueleto de átomos usados en el alineamiento.

Tabla 5
Resumen de los modelos estadísticos obtenidos para los esteroides de la Tabla 4.

Variables Independientes	Componentes usados	s^2 Entrenamiento	q^2 Entrenamiento	s^2 Validación[a]
		CBG		
Operador de Superposición	4	0,821	0,633	0,404 (0,710)
Operador Coulombico	1	0,820	0,533	0,514 (0,646)
CoMFA Electrostatico	1	0,718	0,644	0,352 (0,619)
CoMFA Estarico	2	0,604	0,761	0,421 (0,760)
CoMFA Total	3	0,678	0,716	0,567 (0,835)
		TBG		
Operador de Superposición	3	1,146	0,244	–
Operador Coulombico	3	0,665	0,743	–
CoMFA Electrostatico	3	0,826	0,603	–
CoMFA Esterico	4	0,972	0,483	–
CoMFA Total	3	1,004	0,414	–

a. valores entre paréntesis corresponden a la desviación estándar incluyendo el esteroide 31.

Agradecimiento

Este programa ha sido posible desarrollarlo gracias al apoyo del FONACIT (subvención S1-2001001186).

Referencias Bibliograficas

1. CARBO-DORCA R., ROBERT D., AMAT LL., GIRONES X., BESALU E. ***Molecular Quantum Similarity in QSAR and Drug Design***. Lectures Notes in Chemistry 73, Springer-Verlag, Berlin, 2000.
2. BESALU E., GIRONES X., AMAT LL., CARBO-DORCA R. ***Acc Chem Res*** 35: 289-295, 2002.
3. MEZEY P.G. ***Shape in Chemistry: An introduction to molecular shape and topology***. VCH, New York (USA), 1993.
4. COOPER D.L., ALLAN N.L. ***J Comput-Aided Mol Design*** 3: 253-259, 1989.
5. COOPER D.L., ALLAN N.L. ***J Am Chem Soc*** 114: 4773-4776, 1992.
6. COOPER D.L., ALLAN N.L. ***J Chem Inf Comp Sci*** 32: 587-590, 1992.
7. ALLAN N.L., COOPER D.L. ***Top Curr Chem*** 173: 85-111, 1995.
8. CIOSLOWSKI J., FLEISCHMANN E.D. ***J Am Chem Soc*** 113: 64-67, 1991.
9. CIOSLOWSKI J. ***Theor Chim Acta*** 81: 319-327, 1992.
10. AMAT LL., BESALU E., CARBO-DORCA R., PONEC R. ***J Chem Inf Comput Sci*** 41: 978-991, 2001.
11. CARBO-DORCA R. ***Int J Quantum Chem*** 79: 163-177, 2000.

12. GIRONES X., CARBO-DORCA R. ***J Chem Inf Comput Sci*** 42: 317-325, 2002.

13. http://iqc.udg.es/cat/similarity/ASA/

14. CRAMER R.D., PATERSON D.E., BUNCE J.D. ***J Am Chem Soc*** 110: 5959-5967, 1988.

15. CONSTANTS P., AMAT LL., CARBO-DORCA R. ***J Comput Chem*** 18: 826-846, 1997.

16. MCMAHON A.J., KING P.M. ***J Comput Chem*** 18: 151-158, 1997.

17. PRESS W.H., TEUKOLSKY S.A., VETTERLING W.T., FLANNERY B.P. ***Numerical Recipes in FORTRAN***. Cambridge U.P., Cambridge, 1999.

18. GELADI P., KOWALSKI B.R. ***Ann Chim Acta*** 185: 1-17, 1986.

19. DEWAR M.J.S., ZOEBISCH E.G., HEALY E.F., STEWART J.J.P. ***J Am Chem Soc*** 107: 3902, 1985.

20. STEWART J.J.P. ***MOPAC-7***, 1993.

21. HANSCH C., LEO A., TAFT R.W. ***Chem Rev*** 91: 165-195, 1991.

22. PONEC R., AMAT LL., CARBO-DORCA R. ***J Phys Org Chem*** 12: 447, 1999.

23. PONEC R., AMAT LL., CARBO-DORCA R. ***J Comp-Aid Mol Design*** 13: 259, 1999.

24. SOTOMATSU T. MURATA Y., FUJITA T. ***J Comp Chem*** 10: 94, 1988.

25. KIM K.H., MARTIN Y.C. ***J Org Chem*** 56: 2723, 1991.

26. SULLIVAN J.J., JONES A.D., TANJI K.K. ***J Chem Inf Comp Sci*** 40: 1113, 2000.

27. WAGNER M., SADOWSKI J., GASTAIGER J. ***J Am Chem Soc*** 117: 7769, 1995.

28. ROBERT D., AMAT LL., CARBO-DORCA R. ***J Chem Inf Comp Sci*** 39: 333, 1999.