Comparación de modelos novedosos de proximidad en Quimioinformática

Oscar Miguel Rivera Borroto[1, 3*], Yoandy Hernández Díaz[1], José Manuel García de la Vega[2], Ricardo del Corazón Grau Ábalo[1], Yovani Marrero Ponce[3]

[1]Laboratorio de Bioinformática, Centro de Estudios de Informática, Facultad de Matemática, Física y Computación, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, 54830 Villa Clara, Cuba. [2]Departamento de Química Física Aplicada, Facultad de Ciencias, Universidad Autónoma de Madrid (UAM), 28049 Madrid, Spain. [3]Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatics Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba.

Comparison of novel proximity models in Chemoinformatics

Comparació de models nous de proximitat en Quimioinformàtica

Recibido: 12 de enero de 2012; revisado: 29 de septeimbre de 2012; aceptado: 1 de octubre de 2012

RESUMEN

El presente trabajo comprende la implementación computacional en el ambiente Java de 21 modelos de proximidad para usarlos en experimentos simulados de búsqueda de similitud; nueve de estos modelos son novedosos en Quimioinformática pues proceden del área de la Psicología, los otros 12 son medidas ya establecidas de la literatura especializada. Posteriormente, las medidas de similitud fueron comparadas y validadas en la "recuperación temprana" usando nueve conjuntos de datos de la Química Medicinal, representados por descriptores numéricos seleccionados por Aprendizaje Automático y un algoritmo de búsqueda eficiente. Los resultados muestran que en tendencia promedio los nuevos modelos se comportan superiormente a los de referencia y que más de la mitad de los mismos se sitúan entre los diez modelos más potentes.

Palabras clave: búsqueda de similitud, conjunto de datos de la Química Medicinal, modelo de proximidad, selección de descriptor por Aprendizaje Automático.

SUMMARY

This work comprises the computational implementation in the Java environment of 21 proximity models to be used in simulated experiments of similarity searching, nine out of which are novel in Chemoinformatics since they come from the psychology field, and other 12 are measures already established in the specialized literature. Afterwards, the similarity measures were compared and assessed at the "early retrieval" using nine data sets from medicinal chemistry, represented by machine learning-selected real descriptors, and one efficient matching algorithm. Results show that in average trends the new models perform superiorly with respect to the reference ones, and more than half of them are among the top-10 models.

Keywords: machine-learning descriptor selection, medicinal chemistry data set, proximity model, similarity searching.

RESUM

Aquest treball comprèn la implementació computacional en l'entorn ambient Java de 21 models de proximitat per usar-los en experiments simulats de recerca de similitud; nou d'aquests models són nous en Quimioinformàtica doncs procedeixen de l'àrea de la Psicologia, els altres 12 són mesures ja establertes de la literatura especialitzada. Posteriorment, les mesures de similitud van ser comparades i validades en la "recuperació primerenca" usant nou conjunts de dades de la Química Medicinal, representats per descriptors numèrics seleccionats per Aprenentatge Automàtic i un algoritme de recerca eficient. Els resultats mostren que en tendència mitjana els nous models es comporten millor que els de referència i que més de la meitat se situen entre els deu models més potents.

Paraules clau: recerca de similitud, conjunt de dades de la Química Medicinal, model de proximitat, selecció de descriptor per Aprenentatge Automàtic.

*Autor para la correspondencia: oscarrb@uclv.edu.cu; Tel: 53 42 281515

INTRODUCCIÓN

La búsqueda de similitud es una prestación importante en los sistemas modernos de gestión de la información química para acceder a la rica información contenida en los enormes repositorios químicos modernos. Básicamente, dadas una representación molecular, una medida de similitud y un algoritmo de búsqueda, el resultado de la técnica devuelve una lista ordenada de moléculas del conjunto de datos en orden decreciente de similitud con respecto a la molécula consultada especificada por el usuario [1].

Tradicionalmente, los estudios de similitud molecular se han concentrado en el uso de "huellas dactilares" (*fingerprints*, en inglés) que no son más que cadenas binarias que codifican la presencia (o ausencia) de fragmentos moleculares [2]. Este enfoque está cimentado en la lógica de base "eficiencia primero, luego efectividad". Sin embargo, los enormes recursos computacionales actuales permite al investigador adoptar la línea de pensamiento contraria de "efectividad primero, luego eficiencia" que consiste en aprovechar toda la información estadística contenida en los descriptores no binarios, teniendo en cuenta su relación con las *escalas de medición* absoluta, razón, aditiva, intervalo, ordinal y nominal [3].

El objetivo principal de este trabajo es comparar nueve modelos de proximidad novedosos en el área de la Quimioinformática con otros 12 ya establecidos en la literatura, usando conjuntos de datos farmacológicos, técnicas de aprendizaje automatizado para la selección de los descriptores moleculares y un algoritmo de búsqueda eficiente.

Tabla 1. *Medidas de proximidad no binarias como referente de comparación*

Medida	Fórmula[a]	Tipo[b]	No.[c]
Manhattan Media	$MM_{XY} = \frac{\sum_{j=1}^{n} \lvert x_j - y_j \rvert}{n}$	D	**1**
Euclidiana Media	$EM_{XY} = \frac{\sqrt{\sum_{j=1}^{n} \lvert x_j - y_j \rvert^2}}{n}$	D	**2**
Euclidiana Cuadrada Media	$ECM_{XY} = \frac{\sum_{j=1}^{n} \lvert x_j - y_j \rvert^2}{n}$	D	**3**
Bray/Curtis	$BC_{XY} = \frac{\sum_{j=1}^{n} \lvert x_j - y_j \rvert}{\sum_{j=1}^{n} (\lvert x_j \rvert + \lvert y_j \rvert)}$	D	**4**
Tan	$T_{XY} = \frac{\sum_{j=1}^{n} x_j y_j}{\sum_{j=1}^{n} x_j^2 + \sum_{j=1}^{n} y_j^2 - \sum_{j=1}^{n} x_j y_j}$	A	**5**
Dice	$D_{XY} = \frac{2\sum_{j=1}^{n} x_j y_j}{\sum_{j=1}^{n} x_j^2 + \sum_{j=1}^{n} y_j^2}$	A	**6**
Fossum	$F_{XY} = \frac{n\left(\sum_{j=1}^{n} x_j y_j - \frac{1}{2}\right)^2}{\sum_{j=1}^{n} x_j^2 \sum_{j=1}^{n} y_j^2}$	A	**7**
Sokal/Sneath(1)	$SS1_{XY} = \frac{\sum_{j=1}^{n} x_j y_j}{2\sum_{j=1}^{n} x_j^2 + 2\sum_{j=1}^{n} y_j^2 - 3\sum_{j=1}^{n} x_j y_j}$	A	**8**
Kulczynski(1)	$Kul1_{XY} = \frac{\sum_{j=1}^{n} x_j y_j}{\sum_{j=1}^{n} x_j^2 + \sum_{j=1}^{n} y_j^2 - 2\sum_{j=1}^{n} x_j y_j}$	A	**9**
Cosine/Ochiai	$Cos_{XY} = \frac{\sum_{j=1}^{n} x_j y_j}{\sqrt{\sum_{j=1}^{n} x_j^2 \sum_{j=1}^{n} y_j^2}}$	A	**10**
Simpson	$Sim_{XY} = \frac{\sum_{j=1}^{n} min(x_j, y_j)}{min\left(\sum_{j=1}^{n} x_j, \sum_{j=1}^{n} y_j\right)}$	A	**11**
Pearson	$r_{XY} = \frac{\sum_{j=1}^{n} (x_j - \overline{X})(y_j - \overline{Y})}{\sqrt{\sum_{j=1}^{n} (x_j - \overline{X})^2 \sum_{j=1}^{n} (y_j - \overline{Y})^2}}$	C	**12**

[a]$x_j(y_j)$ representa el valor del descriptor de la molécula X (Y) en el atributo j; [b]Clasificación de las medidas de proximidad acorde a su naturaleza de definición. D, coeficientes de distancia; A, coeficientes de asociación; C, coeficientes de correlación. [c]Identificador usado a lo largo del trabajo.

MATERIALES Y MÉTODOS

CONJUNTOS DE VALIDACIÓN

Se seleccionaron nueve conjuntos de datos de la Química Medicinal. Los primeros ocho repositorios con la variable externa binarizada fueron usados en estudios QSAR previos [4], estos conjuntos son: inhibidores de la enzima convertidora de angiotensina (ACE), inhibidores de la acetilcolinesterasa (AchE), ligandos para el receptor de la benzodiacepina (BZR), Inhibidores de la ciclooxigenasa (COX-2), inhibidores de la hidrofolato reductasa (DHFR), inhibidores de la glicógeno fosforilasa b (GPB), inhibidores de la termolisina (THER), e inhibidores de la trombina (THR). El noveno caso de estudio consiste en la base de datos abierta del programa de cribado antiviral frente al SIDA del Instituto Nacional del Cáncer de los Estados Unidos de América -NCI's AIDS *Antiviral Screen*, en inglés, (NAAS)- [5]. Para el conjunto NAAS, los compuestos "activos confirmados" (CA) y "moderadamente activos" (CM) se han agrupado en la categoría "**1**" (activos), y los "inactivos confirmados" (CI) se han agrupado en la categoría "**0**" (inactivos).

DESCRIPTORES MOLECULARES Y PROCEDIMIENTO INFORMÁTICO

Los conjuntos de datos fueron editados con la utilidad *JChem for Excel* [6], y reoptimizados con el software CORINA [7], generador de estructuras 3D, con el objetivo de "estandarizar" las bases de datos. Los ficheros de salida fueron cargados en el software DRAGON [8], para el cálculo de descriptores moleculares y luego en el software de minería de datos Weka [9], para un tratamiento que incluyó prefiltrado, rellevado a escala, y selección de rasgos (ver más detalles en ref. [1]).

MODELOS DE PROXIMIDAD

Las medidas tomadas como referencia de comparación en el presente estudio son las reportadas en el trabajo de Al Khalifa et al. [10]. El conjunto resultante consiste en 12 medidas de proximidad cuyas fórmulas y clasificación se brindan en la **Tabla 1**.

Las otras medidas de similitud molecular propuestas en nuestro trabajo están basadas en la teoría Zegers y ten Berge [11]. Estos autores propusieron una fórmula general para los coeficientes de asociación bivariada correspondientes a las escalas métricas. Más adelante, Zegers [12] reporta que la elección de un coeficiente de asociación entre dos variables depende del tipo de escala de las variables, definida por la clase de transformaciones admisibles. Stine [13], cambió el enfoque de "asociación" entre dos variables al de "acuerdo relacional" entre observadores.

Las medidas reportadas en el trabajo de Zergers y ten Berge [11] son las siguientes:

Identidad

No corregido: es igual al coeficiente de Dice (ver **Ec. 6, Tabla 1**)

Corregido

$$e_{XY}^{c} = \frac{2s_{XY}}{s_X^2 + s_Y^2 + \left(\overline{X} - \overline{Y}\right)^2} \qquad \textbf{(13)}$$

Aditividad

No corregido (es el caso especial del "punto de anclaje" de Winer [14])

$$a_{XY} = \frac{2s_{XY}}{s_X^2 + s_Y^2} \qquad \textbf{(14)}$$

donde,

$$s_{XY} = \frac{\sum_{j=1}^{n}(x_j - \overline{X})(y_j - \overline{Y})}{n} \;,\; \overline{X} = \frac{\sum_{j=1}^{n} x_j}{n} \;,\; s_X^2 = \frac{\sum_{j=1}^{n}(x_j - \overline{X})^2}{n}$$

Corregido (no cambia la expresión)

Proporcionalidad

No corregido: es igual al coeficiente Cosine/Ochiai (ver **Ec. 10, Tabla 1**)

Corregido

$$p_{XY}^{c} = \frac{s_{XY}}{T_X + T_Y - \overline{XY}} \qquad \textbf{(15)}$$

donde, $T_X = +\sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n}} T_X = +\sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n}}$

Linealidad

No corregido: es el coeficiente de correlación de Pearson (r_{XY}) (ver **Ec. 12, Tabla 1**)

Corregido (no cambia la expresión)

Las medidas restantes son las reportadas en el trabajo de Stine [13]:

Log-razón o Log-proporcionalidad

No corregido

$$Lp_{XY} = \frac{2\sum_{j=1}^{n} x_j^{1/L_X} y_j^{1/L_Y}}{\sum_{j=1}^{n} x_j^{2/L_X} + \sum_{j=1}^{n} y_j^{2/L_Y}} \qquad \textbf{(16)}$$

donde, $L_X = \sqrt{\frac{\sum_{j=1}^{n}\left[ln(x_j)\right]^2}{n}}$

Corregido

$$Lp_{XY}^{c} = \frac{2\left[\sum_{j=1}^{n} x_j^{1/L_X} y_j^{1/L_Y} - (1/n)\sum_{j=1}^{n} x_j^{1/L_X}\sum_{j=1}^{n} y_j^{1/L_Y}\right]}{\left[\sum_{j=1}^{n} x_j^{2/L_X} + \sum_{j=1}^{n} y_j^{2/L_Y} - (2/n)\sum_{j=1}^{n} x_j^{1/L_X}\sum_{j=1}^{n} y_j^{1/L_Y}\right]} \qquad \textbf{(17)}$$

Log-intervalo o Log-linealidad

No corregido

$$Lr_{XY} = \frac{2\sum_{j=1}^{n} x_j^{1/N_X} y_j^{1/N_Y}}{\left(G_Y^{1/N_Y} / G_X^{1/N_X}\right)\sum_{j=1}^{n} x_j^{2/N_X} + \left(G_X^{1/N_X} / G_Y^{1/N_Y}\right)\sum_{j=1}^{n} y_j^{2/N_Y}} \qquad \textbf{(18)}$$

donde, $G_X = \sqrt[n]{\prod_{j=1}^{n} x_j}$, $N_X = \sqrt{\frac{\sum_{j=1}^{n}\left[ln\left(x_j / G_X\right)\right]^2}{n}}$

Corregido

$$Lr_{XY}^{c} = \frac{2\left[\sum_{j=1}^{n} x_j^{1/N_X} y_j^{1/N_Y} - (1/n)\sum_{j=1}^{n} x_j^{1/N_X}\sum_{j=1}^{n} y_j^{1/N_Y}\right]}{\left(G_Y^{1/N_Y} / G_X^{1/N_X}\right)\sum_{j=1}^{n} x_j^{2/N_X} + \left(G_X^{1/N_X} / G_Y^{1/N_Y}\right)\sum_{j=1}^{n} y_j^{2/N_Y} - (2/n)\sum_{j=1}^{n} x_j^{1/N_X}\sum_{j=1}^{n} y_j^{1/N_Y}} \qquad \textbf{(19)}$$

Ordinal

No corregido

$$e_{XY}^{R} = \frac{2\sum_{j=1}^{n} R(x_j)R(y_j)}{\sum_{j=1}^{n} R(x_j)^2 + \sum_{j=1}^{n} R(y_j)^2} \quad \textit{(20)}$$

Corregido (es el mismo que la ρ de Spearman)

$$e_{XY}^{Rc} = 1 - \frac{6\sum_{j=1}^{n}[R(x_j) - R(y_j)]^2}{n(n^2-1)} \quad \textit{(21)}$$

Donde, $R(x_j)$ es el rango de la molécula X en el atributo j una vez ranqueada esa molécula con las restantes del conjunto de datos.

DISEÑO EXPERIMENTAL

Los modelos de proximidad fueron evaluados a través de la técnica de validación cruzada de 10 pliegues (10-CV) estándar usando solamente subconjunto de activos, o sea, en cada ciclo de la técnica, cada pliegue de activos "dejado fuera" fue utilizado como multiconsulta para ranquear los restantes nueve pliegues de activos junto al conjunto de inactivos. El clasificador empleado para comparar las medidas de similitud fue MAX-SIM [15]. Para evaluar la calidad de la recuperación temprana, se usó la métrica AUC [CROC] "área bajo la curva (de las) características concentradas del operador receptor", con un factor de magnificación $\alpha = 20$ [16-17], cuyos valores medios y desviaciones estándares procedentes del 10-CV fueron usadas para la comparación estadística final de las medidas a través de la prueba *t* de Student con la corrección de Welch [18].

DETALLES DE LA PROGRAMACIÓN

Es de interés en nuestro trabajo el acceder al comportamiento global de las medidas de di(similitud) para lo cual fue necesario ordenar los conjuntos de datos completamente, por lo que se decidió implementar el algoritmo de fuerza bruta paralelizado (utilizando hilos en Java), con una complejidad temporal *O(N*M*f),* siendo *N*M* el resultado de aplicar el clasificador MAX-SIM de cada elemento del conjunto de datos a la multiconsulta y *f* una constante de la fórmula correspondiente a la medida seleccionada.

RESULTADOS Y DISCUSIÓN

COMPARACIÓN DE LOS MODELOS DE PROXIMIDAD

La efectividad del clasificador MAX-SIM usando las medidas de proximidad estudiadas en el "enriquecimiento temprano" de activos fue expresada a través de los valores medios y desviaciones estándares de las AUC[CROC] obtenidas del experimento 10-CV (ver **Tablas 2-4).**

Con el objetivo de detectar diferencias globales entre estos modelos se aplicó un contraste de Friedman, el cual arrojó diferencias extremadamente significativas. A partir de la **Tabla 5** se puede calcular que el rango promedio para los modelos que no son de acuerdo relacional es 9.67 y que el 55.56% de los modelos propuestos en el trabajo están incluidos entre los 10 modelos más potentes ("top-10") inspeccionados.

Luego, con el objetivo de detectar y agrupar modelos con similitudes poblacionales estadísticamente significativas se aplicó una prueba de Wilcoxon por pares. El resultado de esta prueba, i.e., la matriz binaria de diferencias signi-

Tabla 2. *Exactitud de los modelos de proximidad en el "reconocimiento temprano" a través del AUC [CROC]: Primer grupo de modelos*

Datos	MM*	EM	ECM	BC	Tan	D	F
ACE	**(0.1926, 0.1702)****	(0.1865, 0.1471)	(0.1865, 0.1471)	(0.2782, 0.1833)	(0.3766, 0.1992)	(0.3766, 0.1992)	(0.3888, 0.1851)
AchE	(0.2443, 0.2051)	(0.2868, 0.1591)	(0.2868, 0.1591)	(0.3217, 0.2018)	(0.3814, 0.2336)	(0.3814, 0.2336)	(0.3082, 0.2417)
BZR	(0.1636, 0.0941)	(0.2173, 0.1011)	(0.2173, 0.1011)	(0.2850, 0.0754)	(0.3186, 0.1063)	(0.3186, 0.1063)	(0.3184, 0.1253)
COX2	(0.2348, 0.0663)	(0.235, 0.0599)	(0.2350, 0.0599)	(0.4121, 0.0737)	(0.4242, 0.0898)	(0.4242, 0.0898)	(0.4383, 0.1075)
DHFR	(0.4711, 0.0895)	(0.4816, 0.0711)	(0.4816, 0.0711)	(0.4865, 0.0878)	(0.4843, 0.0704)	(0.4843, 0.0704)	(0.2580, 0.0513)
GBP	(0.0831, 0.0872)	(0.0939, 0.1010)	(0.0939, 0.1010)	(0.0872, 0.0876)	(0.1359, 0.1219)	(0.1359, 0.1219)	(0.0791, 0.0783)
THERM	(0.0646, 0.1081)	(0.0387, 0.1058)	(0.0387, 0.1058)	(0.1338, 0.0991)	(0.1269, 0.0990)	(0.1269, 0.0990)	(0.0434, 0.1023)
THR	(0.1502, 0.1320)	(0.1554, 0.160)	(0.1554, 0.1600)	(0.1551, 0.1257)	(0.1404, 0.1215)	(0.1404, 0.1215)	(0.1431, 0.1446)
NAAS	(0.9297, 0.0045)	(0.9325, 0.0046)	(0.9325, 0.0046)	(0.9385, 0.0037)	(0.9404, 0.0034)	(0.9404, 0.0034)	(0.9369, 0.0030)

**Las siglas representan el nombre de los modelos de proximidad. **El formato presentado en la tabla es de la forma: ("media", "desviación estándar").*

Tabla 3. *Exactitud de los modelos de proximidad en el "reconocimiento temprano" a través del AUC [CROC]: Segundo grupo de modelos*

Datos	SS1*	Kul1	Cos	Sim	r	e^c	a
ACE	**(0.3766, 0.1992)****	(0.3766, 0.1992)	(0.3739, 0.1920)	(0.2875, 0.0437)	(0.3592, 0.2152)	(0.3252, 0.1788)	(0.3288, 0.1922)
AchE	(0.3814, 0.2336)	(0.3814, 0.2336)	(0.3689, 0.2349)	(0.0117, 0.0259)	(0.3734, 0.2435)	(0.3852, 0.2445)	(0.3890, 0.2523)
BZR	(0.3186, 0.1063)	(0.3186, 0.1063)	(0.3402, 0.1125)	(0.0107, 0.0307)	(0.3679, 0.1302)	(0.3517, 0.1098)	(0.3439, 0.1274)
COX2	(0.4242, 0.0898)	(0.4242, 0.0898)	(0.4306, 0.0879)	(0.0123, 0.0321)	(0.4249, 0.0905)	(0.4209, 0.0912)	(0.4284, 0.0927)
DHFR	(0.4843, 0.0704)	(0.4843, 0.0704)	(0.4998, 0.0727)	(0.0071, 0.0138)	(0.4917, 0.0813)	(0.4871, 0.0755)	(0.4831, 0.0784)
GBP	(0.1359, 0.1219)	(0.1359, 0.1219)	(0.1422, 0.1197)	(0.0554, 0.0711)	(0.1081, 0.1148)	(0.1223, 0.1214)	(0.1087, 0.1124)
THERM	(0.1269, 0.0990)	(0.1269, 0.0990)	(0.1426, 0.1122)	(0.1585, 0.1210)	(0.1075, 0.1026)	(0.1076, 0.1038)	(0.1108, 0.1082)
THR	(0.1404, 0.1215)	(0.1404, 0.1215)	(0.1516, 0.1296)	(0.0015, 0.0023)	(0.1748, 0.1260)	(0.1448, 0.1125)	(0.1735, 0.1402)
NAAS	(0.9404, 0.0034)	(0.9404, 0.0034)	(0.9393, 0.0032)	(0.9436, 0.0027)	(0.9387, 0.0037)	(0.9407, 0.0039)	(0.9402, 0.0038)

**Las siglas representan el nombre de los modelos de proximidad. **El formato presentado en la tabla es de la forma: ("media", "desviación estándar").*

Tabla 4. *Exactitud de los modelos de proximidad en el "reconocimiento temprano" a través del AUC [CROC]: Tercer grupo de modelos*

Datos	p^{c*}	Lp	Lp^c	Lr	Lr^c	e^R	e^{Rc}
ACE	**(0.5671, 0.2093)****	(0.2735, 0.1898)	(0.2990, 0.1941)	(0.0243, 0.0271)	(0.0872, 0.0555)	(0.3842, 0.1994)	(0.3842, 0.1994)
AchE	(0.2977, 0.1715)	(0.1099, 0.1369)	(0.2214, 0.1622)	(0.0924, 0.1046)	(0.0301, 0.0500)	(0.3633, 0.2052)	(0.3633, 0.2052)
BZR	(0.4372, 0.1308)	(0.2189, 0.0970)	(0.2339, 0.0961)	(0.0148, 0.0155)	(0.0428, 0.0497)	(0.3586, 0.1208)	(0.3586, 0.1208)
COX2	(0.2676, 0.0889)	(0.2307, 0.0478)	(0.2530, 0.0694)	(0.0535, 0.0279)	(0.0643, 0.0322)	(0.4690, 0.0535)	(0.4690, 0.0535)
DHFR	(0.0799, 0.0301)	(0.3692, 0.0881)	(0.2947, 0.0805)	(0.0559, 0.0356)	(0.0233, 0.0208)	(0.5340, 0.0876)	(0.5340, 0.0876)
GBP	(0.2279, 0.1009)	(0.1484, 0.1151)	(0.1139, 0.1098)	(0.1301, 0.1746)	(0.1320, 0.1809)	(0.1956, 0.1270)	(0.1956, 0.1270)
THERM	(0.3745, 0.1771)	(0.0464, 0.1025)	(0.0710, 0.0990)	(0.0562, 0.1369)	(0.0403, 0.0609)	(0.1435, 0.1178)	(0.1435, 0.1178)
THR	(0.2277, 0.2144)	(0.0955, 0.1072)	(0.1352, 0.1535)	(0.0343, 0.0376)	(0.0852, 0.1173)	(0.1601, 0.1216)	(0.1601, 0.1216)
NAAS	(0.9136, 0.0045)	(0.9417, 0.0036)	(0.9250, 0.0049)	(0.9320, 0.004)	(0.8877, 0.0036)	(0.9422, 0.0044)	(0.9422, 0.0044)

**Las siglas representan el nombre de los modelos de proximidad. ¨El formato presentado en la tabla es de la forma: ("media", "desviación estándar").*

Tabla 5. *Potencia relativa de los modelos de similitud en la recuperación temprana de activos según la prueba de Friedman*

Orden[a]	Modelo	Rango Medio	Orden	Modelo	Rango Medio
1*t	e^R	18.39	12	*BC*	10.67
2*t	e^{Rc}	18.39	13	*F*	9.78
3*	*Cos*	15.22	14**	*Lp*	8.11
4**	p^c	14.33	15	EM^t	6.83
5*	*r*	14.11	16	ECM^t	6.83
6**	*a*	14.00	17**	Lp^c	6.78
7**	e^c	13.67	18	*Sim*	6.11
8t	*Tan*	13.61	19	*MM*	6.00
9*t	*D*	13.61	20**	*Lr*	3.89
10t	*SS1*	13.61	21**	Lr^c	3.44
11t	*Kul1*	13.61			

*[a]Cuanto más alto es el rango medio, más potente es el modelo; *medidas de similitud planteadas por Al Khalifa et al. [10], ¨modelos de acuerdo relacional aportados por el presente trabajo, t modelos atados por el "Rango Medio".*

ficativas, se usó como entrada para el algoritmo de agrupamiento de Ward (ver más detalles en ref. [19, 20]). En la **Fig. 1** se observa la formación de dos grandes grupos, según la prueba de Mojena [21]. El primer grupo está formado por

$$G_1 = \{MM, EM, ECM, F, Sim, Lp, Lp^c, Lr, Lr^c\}$$

y el otro por

$$G_2 = \{BC, Tan, D, SS1, Kul1, Cos, r, e^c, a, p^c, Lr, Lr^c\}$$

Todo esto sugiere la separación de los modelos en un grupo de 12 mejores modelos y otro de 9 peores modelos.

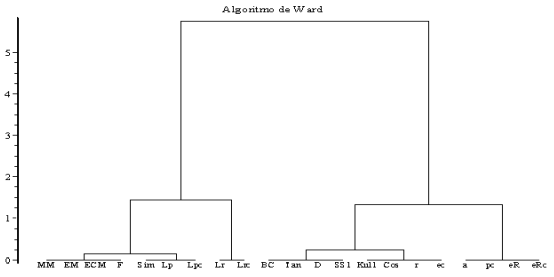

Fig. 1. *Dendrograma obtenido del algoritmo de Ward para el agrupamiento de modelos homogéneos según la prueba de Wilcoxon.*

Posteriormente se aplicó la prueba de comparación por pares *t*-Student con la corrección de Welch empleando los datos de valores medios y desviaciones estándares de las **Tablas 2-4** obtenidos en el repositorio NAAS para desambiguar los mejores modelos en un conjunto representativo de los problemas reales en el área de la Quimioinformática, donde la matriz binaria de diferencias significativas fue sometida a un análisis de agrupamiento jerárquico de Ward (ver **Fig. 2**).

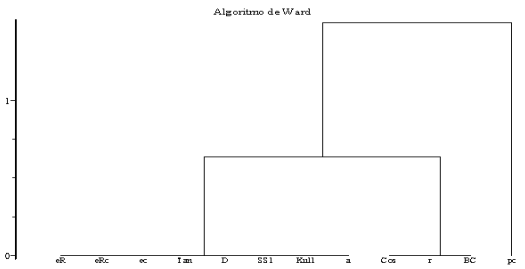

Fig. 2. *Dendrograma obtenido del algoritmo de Ward para el agrupamiento de modelos homogéneos según la prueba t-Student con la corrección de Welch.*

La **Fig. 2** muestra la presencia de tres grupos homogéneos

$$G_1 = \{e^R, e^{Rc}, e^c, Tan, D, SS1, Kul1, a\}, G_2 = \{Cos, r, BC\},$$

$$G_3 = \{p^c\}$$

que han sido dispuestos convenientemente en orden decreciente de efectividad, de modo que los modelos del primer grupo son más potentes que los del segundo grupo, y estos a su vez, son más potentes que p^c.

Por último, para propósitos de comprobación visual de estos resultados, en el **Gráfico 1** se muestran las curvas de

calidad CROC para el grupo de modelos de proximidad anteriores.

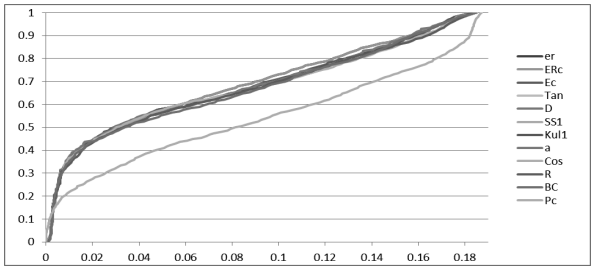

Gráfico 1. *Secciones de curvas CROC para los mejores modelos de proximidad mostrando "el reconocimiento temprano" de los activos en el conjunto de datos NAAS.*

CONCLUSIONES

En este trabajo se implementaron herramientas para el cribado virtual de conjuntos quimioinformáticos que consisten en 21 medidas de similitud para datos numéricos acopladas a un algoritmo de ordenación. Los modelos de proximidad basados en el acuerdo relacional, novedosos en este trabajo, se comportan relativamente superiores a otros modelos no definidos a partir de esta teoría en el reconocimiento temprano de compuestos líderes. Más de la mitad de los modelos propuestos como novedosos en el trabajo están incluidos entre los 10 modelos más potentes, los cuales resultaron ser: e^R, e^{Rc}, p^c, *a* y e^c.

REFERENCIAS

1. Rivera Borroto, O. M.; Hernández Díaz, Y.; García de la Vega, J. M.; Grau Ábalo, R. C.; Marrero Ponce, Y. *Afinidad* **2011**, *68*.
2. Geppert, H.; Vogt, M.; Bajorath, J. R. *J. Chem. Inf. Model.* **2010**, *50*, 205.
3. Siegel, S.; Castellan, N. J. *Nonparametric statistics for the behavioral sciences*; McGraw-Hill: New York, USA, 1988.
4. Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. *J. Chem. Inf. Model.* **2007**, *47*, 219.
5. Voigt, J. H.; Bienfait, B.; Wang, S.; Nicklaus, M. C. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702.
6. ChemAxon *JChem for Excel*, 5.3.8 (166); 2010. URL: http://www.chemaxon.com (visitado el 28 de septiembre de 2012).
7. Sadowski, J.; Gasteiger, J.; Klebe, G. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000. El software "CORINA", generador de estructuras 3D, esta disponible en el sitio de la compañía alemana "Molecular Networks GmbH". URL: http://www.molecular-networks.com (visitado el 28 de septiembre de 2012).
8. Talete srl *DRAGON for Windows* 5.5; Milano, Italy, 2007. URL: http://www.talete.mi.it (visitado el 28 de septiembre de 2012).
9. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. *SIGKDD Explor. Newsl.* **2009**, 11, 10. El software "Weka" v. 3-6-4 esta disponible gratuitamente en el sitio del Grupo de Aprendizaje Automático de la Universidad de Waikako. URL: http://www.cs.waikato.ac.nz/ml/weka/ (visitado el 28 de septiembre de 2012).
10. Al Khalifa, A.; Haranczyk, M.; Holliday, J. *J. Chem. Inf. Model.* **2009**, *49*, 1193.
11. Zegers, F. E.; ten Berge, J. M. F. *Psychometrika* **1985**, *50*, 17.
12. Zegers, F. E. *Psychometrika* **1986**, *51*, 559.
13. Stine, W. W. *Psychol. Bull.* **1989**, *106*, 341.
14. Winer, B. J. *Statistical principles in experimental design*; 2nd ed.; McGraw-Hill: New York, 1971.
15. Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. *J. Med. Chem.* **2005**, *48*, 7049.
16. Swamidass, S. J.; Azencott, C.-A.; Daily, K.; Baldi, P. *Bioinformatics* **2010**, *26*, 1348.
17. Truchon, J.; Bayly, C. I. *J. Chem. Inf. Model.* **2007**, *47*, 488.
18. Sawilowsky, S. S. *J. Mod. Appl. Stat.* **2002**, 1, 461.
19. Rivera-Borroto, O. M.; Marrero-Ponce, Y.; García-de la Vega, J. M.; Grau-Ábalo, R. d. C. *J. Chem. Inf. Model.* **2011**, 51, 3036.
20. Rivera-Borroto, O. M.; Rabassa-Gutiérrez, M.; Grau-Ábalo, R. d. C.; Marrero-Ponce, Y.; García-de la Vega, J. M. *Can. J. Physiol. Pharmacol.* **2012**, 90, 425.
21. Mojena, R. *Comput. J.* **1977**, *20*, 359.