UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS

FACULTAD DE MATEMÁTICA, FÍSICA Y COMPUTACIÓN

NUEVAS MEDIDAS DE SIMILITUD PARA LA BÚSQUEDA DEL VECINO MÁS CERCANO EN REPOSITORIOS QUIMIOINFORMÁTICOS

Tesis presentada en opción al grado científico de Máster en Ciencia de la Computación

Autor: Lic. Yoandy Hernández Díaz

Tutor: Inv. Agr. Dr. Oscar Miguel Rivera Borroto

Santa Clara, Cuba
2015

El que suscribe, Yoandy Hernández Díaz, hago constar que el presente trabajo para optar por el Título de Máster en Ciencia de la Computación ha sido realizado en la Facultad de Matemática, Física y Computación de la Universidad Central "Marta Abreu" de Las Villas como parte de la culminación de los estudios de la Maestría en Ciencia de la Computación, autorizando a que el mismo sea utilizado por la institución, para los fines que estime conveniente, tanto de forma parcial como total y que además no podrá ser presentado en eventos ni publicado sin la autorización de la Universidad.

_____

Firma del autor

Los abajo firmantes, certificamos que el presente trabajo ha sido realizado según acuerdos de la dirección de nuestro centro y el mismo cumple con los requisitos que debe tener un trabajo de esta envergadura referido a la temática señalada.

_____ _____

Firma del tutor Firma del jefe del Laboratorio

_____

Fecha

PENSAMIENTO

Piensa que en ti está el futuro

y encara la tarea con orgullo y sin miedo.

Aprende de quienes puedan enseñarte.

Las experiencias de quienes nos precedieron

de nuestros "poetas muertos",

te ayudan a caminar por la vida.

La sociedad de hoy somos nosotros:

Los "poetas vivos".

No permitas que la vida te pase a ti sin que la vivas...

Walt Whitman

AGRADECIMIENTOS

A mis padres, por darme en todo momento su apoyo incondicional.

A mi tutor, Oscar. A él le debo todo el conocimiento que he adquirido en esta etapa, sin él este trabajo no existiría.

A profesores que pertenecen al equipo de trabajo del laboratorio de Bioinformática, en especial a la Dra. Gladys Casas Cardoso por el voto de confianza.

A todos los compañeros de trabajo que laboran en la Unidad Empresarial Básica Aplicaciones de Redes Sancti Spíritus, en especial a María Ela, Raúl, Nayi y Reinier que me apoyaron en la etapa final de la realización de esta tesis.

En fin, a todos lo que creyeron en mí y a los que no creyeron también, por obligar a esforzarme para demostrarles lo contrario.

DEDICATORIA

A toda mi familia, en especial a mis padres.

SÍNTESIS

Se introducen medidas matemáticas de similitud y se integran a algoritmos de cribado virtual de repositorios quimioinformáticos representados por descriptores moleculares informativos y relevantes. Para ello se implementó en el ambiente de programación Java una estrategia computacional para llegar a los resultados finales. Se fundamenta la disimilitud de Ruzicka de la Ecología para búsqueda de similitud, logrando un desempeño similar a algunas medidas reportadas. Paralelamente, se prueban nueve medidas de la Teoría Estadística de las Mediciones para búsqueda de similitud en Quimioinformática. Dichas medidas superan de forma general a las reportadas, incluyendo al coeficiente de elección (Tanimoto), y proporcionan además información de la relación funcional entre los vectores de representación. Por último, se considera la extensión de las medidas bivariadas de acuerdo relacional al caso de múltiples objetos biomoleculares y se demuestra que las medidas multivariadas pueden expresarse como la media ponderada de sus contrapartes bivariadas. Como resultado, se proponen tres medidas de acuerdo relacional multivariadas que son nuevas en Estadística aplicada en Quimioinformática.

TABLA DE CONTENIDOS

INTRODUCCIÓN

INTRODUCCIÓN

El concepto de similitud juega un rol prominente en Quimioinformática y Bioinformática (1-2). La aplicación prolífica de la similitud en estos campos del conocimiento está cimentada en el *principio de similitud-propiedad* de Johnson y Maggiora (3), el cual plantea que "moléculas estructuralmente similares se espera que exhiban propiedades (actividades) similares". Este paradigma parece ser una adaptación directa del proceso fundamental del pensamiento, *razonamiento por analogía* (4). La comprobación práctica de este principio ha sido apoyada por un buen número de resultados experimentales (5-7); sin embargo, otros hallazgos sugieren que *eventualmente* bio(macro)moléculas estructuralmente similares exhiben comportamientos disimilares, así como bio(macro)moléculas estructuralmente disimilares exhiben comportamientos similares (8-10). Para sistematizar este cuerpo de evidencias algunos autores han propuesto, en el contexto del diseño de fármacos, un cuadro (*matriz de confusión*) de cuatro *hipótesis bayesianas*, o sea, *i-)biomoléculas estructuralmente similares es muy plausible que tengan funciones bioquímicas similares, ii-)biomoléculas estructuralmente similares es plausible que tengan funciones bioquímicas disimilares, iii-)biomoléculas estructuralmente disimilares es plausible que tengan funciones bioquímicas similares, iv-)biomoléculas estructuralmente disimilares es muy plausible que tengan funciones bioquímicas disimilares* (11-12). La validez general de las hipótesis *ii-)* y *iii-)* debe ser aceptada con reservas ya que no se han presentado evidencias suficientemente claras en la literatura, en el sentido que aún no se ha conducido un estudio exhaustivo sobre la selección de rasgos relevantes bio(macro)moleculares en contextos bioquímicos diversos. La causa para la dificultad en establecer un grupo de experimentos controlados pudiera residir en la carencia de mediciones "húmedas" o experimentales fiables, a la promiscuidad observada de algunos ligandos, y a la multiplicidad de funciones de las macromoléculas (genes, proteínas, etc) (13-14). Por otro lado, el aceptar estas dos hipótesis como realidades infalibles presupondría un reto filosófico interesante pues algunos pensadores han postulado que el núcleo cognitivo humano

opera analógicamente (15). La cuestión no solamente radica en que el concepto de analogía es más general que el de similitud (16), sino también que en Quimio(Bio)informática lo que los investigadores usualmente llaman "relaciones de similitud" probablemente sean en realidad *relaciones de similitud estructurada o analógicas*. La consideración anterior se basa en que la comparación entre objetos químicos frecuentemente comprende el emparejamiento de los elementos correspondientes (fragmentos moleculares) y las relaciones entre estos (los fragmentos se encuentran relacionados por enlaces bioquímicos), y esta información aparece codificada casi siempre en los descriptores que luego se usan para los cálculos de similitud. Cualquiera que fuere la situación, si es que estas hipótesis están poco fundamentadas o que en el futuro habremos de aceptarlas como parte de nuestra realidad, el hecho es que las hipótesis *i-)* y *iv-)* conforman la lógica de base para técnicas como la búsqueda de similitud, muestreo de compuestos intraclústeres y modelación de QSAR (17-19).

En Quimioinformática y Bioinformática las medidas de (di)similitud bio(macro)moleculares se han tratado tradicionalmente como relaciones *binarias* y generalizaciones explícitas al caso de varios objetos o *k-arias* han recibido poca atención en estas áreas; la mayoría de las veces el uso de estas últimas se ha restringido a las funciones objetivo para la selección de compuestos basada en disimilitud (25), o para la comparación de secuencias y estructuras de proteínas y genomas (26). Por otra parte, salvo algunas excepciones (27), la mayoría de las medidas de similitud usadas en estas áreas carecen de una distribución estadística asociada, por tanto, la decisión acerca de si un valor de similitud en particular es significativo debe hacerse sobre bases irregulares. Por último, las medidas de similitud usadas en estas áreas cuantifican el grado de semejanza entre las bio(macro)moléculas pero no brindan información sobre la relación funcional explícita entre los vectores de representación correspondientes.

A partir del análisis de la literatura se nota que las cadenas binarias o fingerprints (p.e., las huellas MACCs, las huellas dactilares Daylight, las huellas BCI, etc) han sido adoptadas como el medio

de facto para la representación molecular (28), y la razón para ello radica principalmente en la eficiencia con la que estas pueden ser generadas, comparadas y almacenadas (29-30). Alternativamente, en contadas publicaciones, otros autores han introducido diferentes tipos de descripción numérica que capturan mayor información química de las entidades moleculares (31-32). Aunque, desde la perspectiva computacional, el usar descriptores numéricos es una estrategia menos eficiente que usar huellas dactilares, desde el punto de vista estadístico el transformar descriptores de escala continua (discreta) en descriptores de escala nominal (binarios) conlleva a una pérdida de información estadística que afecta la potencia, capacidad de resolución de las ataduras de proximidad y versatilidad en general de los métodos basados en similitud (33). También se constata que la selección de rasgos es una etapa casi ignorada en los estudios de comparación de técnicas basadas en similitud, usualmente consideradas como "*técnicas* (obligatoriamente) *no supervisadas*" a pesar de que la mayoría de estos se han conducido en escenarios *supervisados*. Sin embargo, la importancia de la selección automática de rasgos para este tipo de técnicas se ha resaltado en otros campos del conocimiento (34-38). En Quimioinformática, la selección de rasgos se ha guiado mayoritariamente por la experiencia acumulada de estudios anteriores o por la estrategia de "prueba y error" y, en la extensión de nuestros conocimientos, poquísimos trabajos han usado técnicas automáticas en esta etapa (39-40), que sean consistentes con el principio de vecindad (o similitud) (41-42).

Como consecuencia de las razones expuestas anteriormente surge el **problema científico:**

¿Cómo introducir medidas de similitud biomoleculares novedosas en Quimioinformática, superiores cuantitativamente y cualitativamente a las ya reportadas, que usen representaciones biomoleculares informativas relevantes al contexto bioquímico estudiado y permitan mejorar el desempeño de los principales algoritmos de búsqueda de similitud en la clasificación de datos quimioinformáticos?

Como vía para solucionar este problema científico se formula la **hipótesis de investigación** siguiente:

- El uso de medidas de similitud de otras áreas del conocimiento en Quimioinformática usando representaciones biomoleculares numéricas coherentes con el principio de similitud mejorará las medidas de similitud reportadas en el problema de la búsqueda del vecino más cercano.

Como **objetivo general** se plantea introducir y validar medidas novedosas de similitud en Quimioinformática usando una metodología apropiada que integre un algoritmo de búsqueda eficiente, conjunto de datos farmacológicos estándares, descriptores moleculares (DMs) seleccionados de acuerdo al *principio de similitud* y una métrica de validación apropiada para el *problema de la detección temprana*, superiores a las medidas ya reportadas en cuanto a la interpretación de las relaciones de semejanza biomoleculares y al desempeño de los principales algoritmos de búsqueda de similitud.

Para dar cumplimiento a este objetivo general se han trazado tres **objetivos específicos**:

- Implementar técnicas de cribado virtual en el lenguaje libre *Java* para datos quimio-/bio-informáticos basadas en la búsqueda de similitud, disponiendo de varias medidas de similitud, modelos de fusión, algoritmos de búsqueda y métricas de validación automática.
- Emplear medidas binarias de similitud de la Ecología y compararlas con medidas ya reportadas en el desempeño de algoritmos de búsqueda de similitud de datos químico-médicos representados por DMs numéricos relevantes.
- Comparar medidas binarias de similitud de la Teoría Estadística de las Mediciones con las ya reportadas en cuanto al desempeño de la búsqueda de similitud de datos quimioinformáticos representados por DMs numéricos relevantes, y lograr interpretaciones tipológicas de las relaciones de semejanza biomoleculares.

- Realizar un estudio teórico de las medidas de acuerdo relacional que comprenda el generalizar las medidas bivariadas al caso de múltiples objetos y analizar la dependencia funcional entre medidas multivariadas y bivariadas.

La *novedad científica* de este trabajo se fundamenta en la propuesta de medidas de similitud novedosas en Quimioinformática, ya que mejoran a las medidas clásicas en cuanto a un mejor desempeño de los principales algoritmos de búsqueda similitud y algunas de ellas incorporan información distribucional y funcional entre los vectores de representación, y, pueden ser generalizadas para la modelación de las relaciones de semejanza de múltiples objetos biomoleculares.

La presente tesis se estructura en tres capítulos. El capítulo 1 está dedicado a aspectos filosóficos y matemáticos relacionados con los conceptos de similitud, y a la descripción de las principales técnicas de búsqueda de similitud a través de sus componentes esenciales usadas en este trabajo. El segundo capítulo se dedica a la exposición secuencial de las etapas de la investigación, detallando los materiales y métodos empleados en cada caso. En el tercer capítulo se presentan los resultados principales del trabajo, tanto computacionales como teóricos, que ilustran la relevancia de las medidas de similitud propuestas por primera vez en Quimioinformática. A ello le siguen las conclusiones, recomendaciones y referencias bibliográficas.

1. MARCO TEÓRICO

1. MARCO TEÓRICO

Este capítulo de la tesis está dedicado primeramente al tratamiento epistemológico del concepto de similitud. También, se brindan los modelos más extendidos para el estudio de la similitud; especial énfasis se hace en el modelo geométrico donde se trata el concepto dual de disimilitud. A continuación se muestran los modelos matemáticos clásicos, tanto para el caso usual donde la disimilitud se considera como una relación entre dos objetos matemáticos, como para el caso tratado más recientemente en la literatura, donde este concepto se extiende como una relación entre múltiples objetos. Más adelante, se abordan los métodos y técnicas características de la Quimioinformática; de interés particular resultan las técnicas de VS basadas en similitud donde se describen detalladamente una de sus componentes esenciales: los algoritmos de búsqueda de similitud.

1.1. Similitud

Cuando puede establecerse solamente una correspondencia entre los elementos de dos objetos hablamos entonces de *similitud* o *similaridad* entre los mismos. En la rama de la Quimio(Bio)informática, la similitud debe entenderse como similitud de estructuras moleculares, que en principio garantice al menos con basta probabilidad igualdad de propiedades (funciones). La cuestión es cómo medir la similitud y este es el objetivo esencial del presente trabajo.

1.1.1. Enfoques generales sobre similitud.

Las evaluaciones humanas de similitud son fundamentales para la cognición porque las similitudes son reveladoras para el mundo en que vivimos. Este mundo es un lugar suficientemente ordenado de modo que los objetos y eventos tienden a comportarse similarmente; este hecho, observado en múltiples fenómenos no es una coincidencia fortuita, sino que se debe al principio básico fundamental de que *objetos que son similares también tenderán a comportarse similarmente* en la mayoría de los aspectos (52).

1.1.1.1. Modelos más difundidos de similitud.

En la literatura científica ha aparecido un número importante de tratados que brindan enfoques teóricos de similitud y describen como la misma puede ser medida empíricamente (53). Estos modelos han tenido un profundo impacto en estadística, reconocimiento automático de patrones, minería de datos y marketing. Los modelos de similitud más importantes son: a) el geométrico, b) el basado en rasgos, c) basado en alineamiento y d) transformacional (54).

1.1.1.1.1. Modelo geométrico.

El *modelo geométrico* funciona bajo la premisa de que dos cosas son similares cuando están cerca una de la otra en un espacio matemático. Estos enfoques son ejemplificados por la técnica de modelación estadística de escalamiento multidimensional (MDS, de sus siglas en inglés, *Multidimensional Scaling*). Los modelos MDS representan las relaciones de similitud entre entidades en término de un modelo geométrico que consiste en un conjunto de puntos embebidos en un espacio (matemático). La entrada de las rutinas MDS pueden ser juicios de similitud, confusiones entre entidades, patrones de coocurrencia en muestras grandes de texto, u otra medida de proximidad por pares. La salida de una rutina MDS es un modelo geométrico de un conjunto de objetos, con cada objeto representado en un espacio *n*-dimensional (55). La similitud entre un par de objetos se evalúa entonces como relacionada inversamente a la distancia entre dos objetos-punto en dicho espacio. En la técnica MDS, la distancia entre los puntos *X* e *Y* se calcula típicamente mediante la distancia de Minkowsky generalizada (o distancia l_p, $p \geq 1$) como:

$$d_{XY} = \left[\sum_{j=1}^{n} \left|x_j - y_j\right|^p\right]^{1/p} \quad \textbf{(1.1)}$$

donde *n* es la dimensión del espacio y x_j e y_j son los valores de las variables *X* e *Y* en el rasgo *j*, respectivamente.

1.1.1.1.2. Modelos de características.

En los años 1970, se observó que las evaluaciones subjetivas de similitud no siempre satisfacen los supuestos de los modelos geométricos de similitud:

i. $D(X,Y) \geq D(X,X) = 0$ (minimalidad)

ii. $D(X,Y) = D(Y,X)$ (simetría)

iii. $D(X,Y) \leq D(X,Z) + D(Y,Z)$ (desigualdad triangular)

En este caso, $D(X,Y)$ debe ser interpretado en el sentido más amplio de disimilitud entre los ítems X e Y.

En la práctica, de hecho, se han obtenido empíricamente evidencias que violan los tres supuestos. El modelo más radical fue presentado en 1977 por Amos Tversky (56), quien propuso una modelación de la similitud en términos del emparejamiento o desparejamiento de características o rasgos, en vez de distancias entre dimensiones psicológicas. En su modelo, las entidades se representan como colecciones de rasgos y la similitud se calcula por:

$$S(X,Y) = \theta f(X \cap Y) - \alpha f(X/Y) - \beta f(Y/X) \quad \textbf{(1.2)}$$

donde, $S(X,Y)$ es la similitud del objeto X al objeto Y, y es expresada como una combinación lineal de la medida de los rasgos comunes y distintivos. El termino $(X \cap Y)$ representa los rasgos que los objetos X e Y tienen en común. La expresión (X/Y) representa los rasgos que tiene X y no tiene Y. Similarmente, (Y/X) representa los rasgos que tiene Y y no tiene X. Por otra parte, f representa una función de emparejamiento (p.e., el cardinal de conjuntos). Finalmente, θ, α y β son los pesos de los componentes comunes (θ) y distintivos (α y β); nótese que el peso asignado a los rasgos distintivos no es el mismo y entonces el modelo permite describir asimetrías observadas en experimentos que involucran estímulos psicológicos.

1.1.1.1.3. Modelo basado en alineamiento.

Una comunalidad entre las representaciones *geométricas* y de *características* es que ambas usan representaciones relativamente no estructuradas. Sin embargo, entidades como los objetos naturales con partes, escenarios del mundo real, palabras, oraciones, historias, teorías científicas, y hasta caras humanas no son simplemente una "bolsa de atributos", y es que las relaciones entre las partes de las entidades es algo muy importante para lograr una modelación más realista de estas. Inspirados en los modelos de emparejamiento de estructuras de Dedre Gentner (57), en los modelos basados en alineamiento, el emparejamiento de rasgos influye la similitud más acentuadamente si pertenecen a partes que están situadas en correspondencia. Las partes tienden a estar situadas en correspondencia si estas tienen muchos rasgos en común y son consistentes con otras correspondencias emergentes.

1.1.1.1.4. Modelos transformacionales.

Este enfoque sobre similitud sostiene que la similitud entre dos objetos está directamente relacionada con el número de *transformaciones* requeridas para convertir un objeto a otro. Las operaciones de alineamiento rotan, escalan, trasladan y deforman topográficamente las descripciones de los objetos. De acuerdo a Hahn y coautores (53), la similitud entre dos entidades está basada en cuan compleja es la secuencia de transformaciones que transforma una entidad en la otra; desde este punto de vista, mientras más simple es la secuencia de transformaciones entre dos objetos, más similares tienden a ser estos.

1.1.2. Tratamiento matemático de la similitud.

Se abordan ahora posibles enfoques diferentes de tratamiento matemático de la similitud.

1.1.2.1. Relaciones de similitud entre dos objetos o binarias.

En Matemática, el concepto de similitud es analizado usualmente a través de su concepto opuesto o dual de *disimilitud*. De hecho, las medidas de similitud y disimilitud pueden ser interconvertidas a través de transformaciones matemáticas relativamente sencillas que conservan las propiedades topológicas del par (58). Uno de los modelos matemáticos más difundidos en la literatura exige que para que una medida de disimilitud sea "bien comportada", esta debe satisfacer dos conjuntos de supuestos, esto es, dimensionales y métricos. Los supuestos dimensionales son necesarios y suficientes para la sustractividad intradimensional y la aditividad interdimensional, mientras que las propiedades métricas se refieren a los siguientes axiomas:

Sea $\mathbb{R}^+$ el conjunto de números reales no negativos. Una métrica es un par (E, d), donde E es un conjunto no vacío y d una función de disimilitud $d: ExE \rightarrow \mathbb{R}^+$, que satisface para todo $x, y, z \in E$:

i.) $d(x,y) = 0 \Leftrightarrow x = y$ (minimalidad)

ii.) $d(x,y) = d(y,x)$ (simetría)

iii.) $d(x,y) \leq d(x,z) + d(y,z)$ (desigualdad triangular)

Tversky y Gatti (59) argumentan que a los axiomas anteriores debe agregarse un cuarto axioma definido como:

iv.) $d(x,y) = d(x,z) + d(y,z)$ (aditividad segmentaria, cuando z yace en el segmento entre *x* a *y*)

Esta última propiedad es ignorada usualmente, pero es muy importante porque en la ausencia de segmentos aditivos, la desigualdad triangular puede satisfacerse trivialmente adicionando una constante suficientemente grande a todas las distancias entre los distintos puntos.

Cuadras (60) ha propuesto un esquema más extendido que abarca otros modelos matemáticos para las relaciones binarias de distancia, de este modo sean las siguientes propiedades o axiomas:

P1. $d_{ij} \geq 0$

P2. $d_{ii} = 0$

P3. $d_{ij} = d_{ji}$

P4. $d_{ij} \leq d_{ik} + d_{jk}$

P5. $d_{ij} = 0$ ssi $i = j$

P6. $d_{ij} \leq \max\{d_{ik}, d_{jk}\}$ (desigualdad ultramétrica)

P7. $d_{ij} + d_{kl} \leq \max\{d_{ik} + d_{jl}, d_{il} + d_{jk}\}$ (desigualdad aditiva)

P8. d_{ij} es euclidea; esto significa que existen dos puntos $\boldsymbol{X_i} = (x_{i1}, x_{i2}, x_{i3}, \dots, x_{in})$ y $\boldsymbol{X_j} = (x_{j1}, x_{j2}, x_{j3}, \dots, x_{jn})$ de $\mathbb{R}^n$ tales que:

$$d_{ij}^2 = (\boldsymbol{X_i} - \boldsymbol{X_j})^t (\boldsymbol{X_i} - \boldsymbol{X_j})$$

Es decir, si d_{ij} es la distancia euclídea entre los puntos $\boldsymbol{X_i}$ y $\boldsymbol{X_j}$, entonces (E, d) puede representarse en el espacio euclídeo $(\mathbb{R}^n, d)$.

P9. d_{ij} es riemaniana; lo cual significa que (E, d) puede ser representado mediante una variedad de Rieman (M, d_M).

P10. d_{ij} es una divergencia; supongamos que hemos definido una medida de probabilidad μ sobre *E*, entonces esta propiedad significa que *d* es una expresión funcional sobre μ.

A partir de estas propiedades, podemos clasificar las distancias según las mismas en:

Disimilaridad: *P1*, *P2*, *P3*

Distancia métrica: *P1*, *P2*, *P3*, *P4*, *P5*

Distancia ultramétrica: *P1*, *P2*, *P3*, *P6*: (E, d) puede ser representado a través de un dendrograma

Distancia euclídea: *P1*, *P2*, *P3*, *P4*, *P8*

Distancia aditiva: *P1*, *P2*, *P3*, *P7*

Actualmente, existe una gran variedad de modelos matemáticos que "demandan" propiedades características, dando como resultado una gran variedad de *métricas*. Una fuente excelente que hace énfasis en esta materia importante se brinda en la referencia (61).

1.1.2.2. Generalización al caso de varios objetos o *k*-arias.

La generalización de la noción de métrica como relación entre *dos objetos* o *binaria* al caso de *multimétrica* como la relación entre *varios objetos* o *k-arias* ha sido estudiada en gran medida por Warrens (62). En este punto, es útil hacer una presentación de tales definiciones pues precisamente varias medidas de similitud propuestas en este trabajo están inspiradas en esta idea. Para ello utilizaremos la notación de este autor.

Sea $x_{1,k} = (x_1, x_2, x_3, \dots, x_k)$ denote la *k*-upla y sea $x_{1,k}^{-i} = (x_1, x_2, x_{i-1}, x_{i+1} \dots, x_k)$ denote la (*k*-1)-upla donde el menos en el supraíndice de $x_{1,k}^{-i}$ se usa para indicar que el objeto x_i ha sido retirado de la *k*-upla. A partir de aquí se define una multimétrica como una medida de disimilitud que satisface:

i.) $d_k(x_i, x_i, x_i, \dots, x_i) = 0 \; \forall i = \overline{1,k}$ (Minimalidad generalizada)

ii.) Una disimilitud $d_k: E^k \to \mathbb{R}^+$es totalmente simétrica para todos los $x_1, x_2, x_3, \dots, x_k \in E$ y cada permutación π de $\{x_1, x_2, x_3, \dots, x_k\}$ si:

$d_k\left(x_{\pi(1)}, \dots, x_{\pi(k)}\right) = d_k(x_1, \dots, x_k)$ (Simetría generalizada)

iii.) $d_k\left(x_{1,k}\right) \leq \sum_{i=1}^{k} d_k\left(x_{1,k+1}^{-i}\right)$ (Desigualdad triangular generalizada para métricas débiles)

iv.) $(k-1)d_k\left(x_{1,k}\right) \leq \sum_{i=1}^{k} d_k\left(x_{1,k+1}^{-i}\right)$ (Desigualdad poliédrica o desigualdad triangular generalizada para métricas fuertes)

v.) $d_k\left(x_1, x_{1,k-1}\right) = d_k\left(x_{1,2}, x_{2,k-1},\right) = \cdots = d_k\left(x_{1,k-1}, x_{k-1}\right)$ (Requerimiento de que si cualesquiera dos objetos son iguales d_k debe permanecer invariante)

vi.) $d_{k-1}\left(x_{1,k-1}\right) = \frac{1}{p} d_k\left(x_1, x_{1,k-1}\right)$ (Expresa que d_k y d_{k-1} son iguales hasta un factor de multiplicación *p* cuando dos objetos son idénticos)

vii.) $d_k\left(x_1, x_{1,k-1}\right) \leq d_k\left(x_1, x_{2,k}\right)$ (Expresa que d_k sin objetos idénticos siempre es mayor o igual que d_k con objetos idénticos)

1.2. Métodos computacionales en Quimio(Bio)informática.

Debido a la necesidad de explotar las cantidades masivas de datos generados por las tecnologías de alto rendimiento, los métodos computacionales se han ido implementando de manera creciente en el proceso de descubrimiento de fármacos y de manera general en la química. Para unificar la combinación de los métodos computacionales y la química, F. K. Brown acuñó en 1998 el término "Quimioinformática" definiéndola como: *"la combinación de aquellos recursos de información para transformar datos en información y la información en conocimiento con el propósito de tomar mejores y más rápidas decisiones en el área de la identificación y optimización de compuestos líderes"* (63). La Quimioinformática implica el uso de las tecnologías informáticas para procesar los datos químicos. Esta definición general engloba múltiples aspectos, en particular, la representación, almacenamiento, recolección y análisis de la información química en un sistema informático (64). Otro de los retos que enfrenta este nuevo campo es establecer relaciones claras entre los patrones estructurales y las actividades o propiedades. Uno de los primeros estudios quimioinformáticos involucra las representaciones de estructuras químicas, tales como descriptores estructurales.

1.2.1. Cribado virtual.

El cribado virtual (VS, del inglés Virtual Screening) consiste en el análisis computacional de bases de datos de compuestos, dirigido a identificar y seleccionar un número limitado de candidatos que posean la actividad biológica deseada sobre un blanco terapéutico específico (73). Pueden evaluarse *in silico* cantidades arbitrarias de moléculas reales o virtuales y se pueden ahorrar como promedio 140 millones de USD y 0.9 años por cada fármaco (74).

En la práctica, el VS requiere del conocimiento de la estructura del blanco terapéutico, usualmente obtenido por métodos cristalográficos, o de la actividad, medida experimentalmente, en un conjunto de compuestos. Si la estructura de la diana farmacológica es conocida, el enfoque más

común para el VS son los estudios de acoplamiento o "docking", que consisten en la derivación de una puntuación o "score" de la actividad a partir del posicionamiento óptimo del ligando en el sitio activo del blanco (76). Este enfoque de VS suele brindar los resultados más confiables y al mismo tiempo resulta atractivo por el gran número (alrededor de 500) de dianas farmacológicas disponibles (ver **Fig. 1.1**).

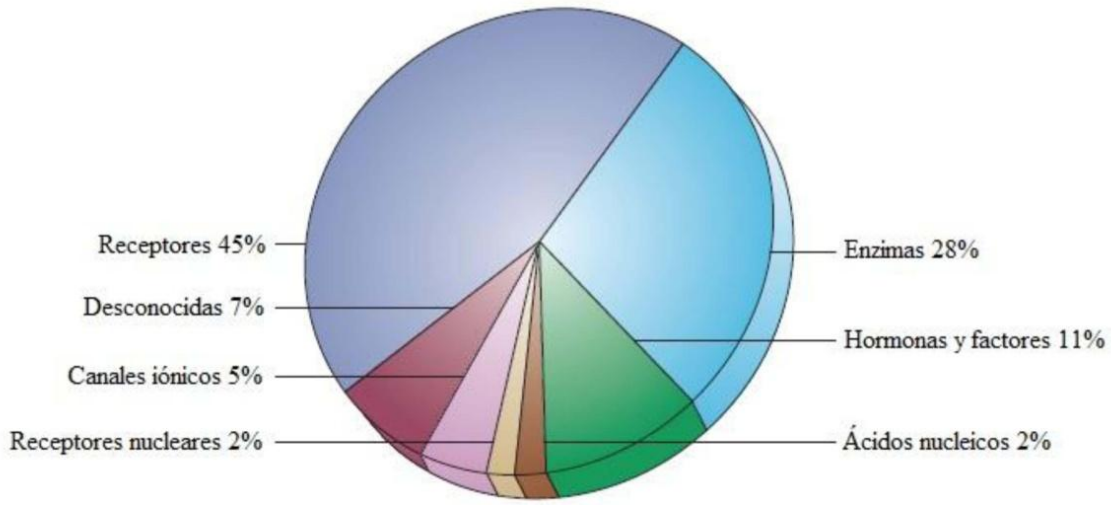


Figura 1.1. Dianas farmacológicas distribuidas en siete clases bioquímicas principales, donde las enzimas y receptores representan la parte mayoritaria.

Si se desconoce la estructura del blanco, los métodos de VS pueden derivarse de un conjunto de compuestos con actividad conocida obtenidos de ensayos experimentales previos. Estos métodos se conocen como enfoques de VS basados en ligandos y en la práctica suelen complementarse con el enfoque anterior basado en la estructura del blanco (77).

1.2.2. Componentes básicos de una técnica de VS basada en similitud.

El VS basado en similitud molecular comprende cuatro componentes esenciales: la medida de (di)similitud, que cuantifica el grado y tipo de semejanza entre dos compuestos químicos; el conjunto de datos estructurales químicos, que cubre cierta región del espacio químico a explorar; la representación matemática de los compuestos químicos, a través de DMs y la medida de la efectividad y eficiencia de la técnica usada para el problema quimioinformático en cuestión (78).

1.2.2.1. Medidas de similitud.

El concepto de similitud es fundamental para varios aspectos del razonamiento y análisis químicos, y cae bajo la rúbrica general del análisis de la **similitud molecular**. La determinación de la similitud caracteriza el grado de concordancia, de asociación, semejanza, alineamiento, porcentaje de identidad, similitud o más generalmente *proximidad* entre pares de moléculas dadas por sus "patrones moleculares", que se componen de conjuntos de rasgos (79). Consideremos dos objetos químicos arbitrarios A y B descritos mediante vectores *X* e *Y*, respectivamente, de *n* atributos, de modo que $X = (x_1, x_2, x_3, \dots, x_n)$ e $Y = (y_1, y_2, y_3, \dots, y_n)$. En la **Tabla 1.1** se muestran algunas medidas de (di)similitud con aplicaciones en Quimioinformática (80).

Tabla 1.1. Algunas de las medidas de distancia reportadas en la literatura Quimioinformática.

Nombre	Fórmula[a]	Rango
Manhattan/ City-Block	$d_{XY} = \sum_{j=1}^{n} \lvert x_j - y_j \rvert$	$[0, \infty)$
Euclídea	$d_{XY} = \sqrt{\sum_{j=1}^{n} \lvert x_j - y_j \rvert^2}$	$[0, \infty)$
Chebyshev/Lagrange	$d_{XY} = max\{\lvert x_j - y_j \rvert\}$	$[0, \infty)$
Bhattacharyya	$d_{XY} = \sqrt{\sum_{j=1}^{n} \left(\sqrt{x_j} - \sqrt{y_j}\right)^2}$	$[0, \infty)$
Mahalanobis	$d_{XY} = \sqrt{(X-Y)^t \boldsymbol{S}^{-1} (X-Y)}$	$[0, \infty)$
Correlación	$d_{XY} = 1 - \frac{\sum_{j=1}^{n}(x_j - \bar{X})(y_j - \bar{Y})}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{X}) \sum_{j=1}^{n}(y_j - \bar{Y})}}$	[0,2]
Separación Angular	$d_{XY} = 1 - \frac{\sum_{j=1}^{n} x_j y_j}{\sqrt{\sum_{j=1}^{n} x_j^2 \sum_{j=1}^{n} y_j^2}}$	[0,2]
Camberra	$d_{XY} = \sum_{j=1}^{n} \frac{\lvert x_j - y_j \rvert}{\lvert x_j \rvert + \lvert y_j \rvert}$	$[0, n]$
Soergel	$d_{XY} = \frac{\sum_{j=1}^{n} \lvert x_j - y_j \rvert}{\sum_{j=1}^{n} max\{x_j, y_j\}}$	[0,2]
Lance-Williams/Bray-Curtis I	$d_{XY} = \frac{\sum_{j=1}^{n} \lvert x_j - y_j \rvert}{\sum_{j=1}^{n} (\lvert x_j \rvert + \lvert y_j \rvert)}$	[0,1]
"Wave-Edges"	$d_{XY} = \sum_{j=1}^{n} \left(1 - \frac{min\{x_j, y_j\}}{max\{x_j, y_j\}}\right)$	$[0, 2n]$

[a] $x_j(y_j)$ es el valor del vector *X* (*Y*) correspondiente al atributo *j*. $\boldsymbol{S}$ representa la matriz de varianza-covarianza.

Cuando los valores de los atributos se limitan a 0 y 1, las expresiones utilizadas por varias similitudes y medidas de distancia a menudo pueden simplificarse considerablemente. Si los objetos A y B que se caracterizan por vectores *X* e *Y* que contienen *n* valores binarios (tales como huellas dactilares) se pueden definir las cantidades *a*, *b*, *c*, *d* o elementos de la *matriz de confusión* como:

$a = \sum_{j=1}^{n} x_j$, es el número de bits activos en A (**1.3**)

$b = \sum_{j=1}^{n} y_j$, es el número de bits activos en B (**1.4**)

$c = \sum_{j=1}^{n} x_j y_j$, es el número de bits activos en A y B (**1.5**)

$d = \sum_{j=1}^{n} (1 - x_j - y_j + x_j y_j)$, es el número de bits inactivos en A y B (**1.6**)

Por tanto, $n = a + b - c + d$ (**1.7**)

Las cantidades anteriores también se pueden expresar en notación de teoría de conjuntos dando lugar a las formulaciones análogas basadas en este tipo de representación (81).

Como ejemplo ilustrativo de la simplificación en las fórmulas tomemos el coeficiente de Tanimoto:

$$T_{XY} = \frac{\sum_{j=1}^{n} x_j y_j}{\sum_{j=1}^{n} x_j^2 + \sum_{j=1}^{n} y_j^2 - \sum_{j=1}^{n} x_j y_j} \quad (\mathbf{1.8})$$

Para el caso binario la fórmula correspondiente vendría dada por:

$$T_{XY} = c/[a + b - c] \quad (\mathbf{1.9})$$

Este coeficiente aplicado a las huellas dactilares 2D (cadenas binarias) constituye actualmente la medida de elección de los sistemas de software comerciales para la gestión de la información química (26). También forma parte de sistemas de acceso público importantes como el PubChem (82).

1.2.2.2. Conjuntos de datos quimioinformáticos.

La medición del rendimiento de los índices de similitud, DMs e incluso enfoques de validación, es estrictamente dependiente de las bases de datos de prueba, de la configuración del espacio químico

y de la problemática tratada. Este problema se pudiera arreglar evaluando los métodos nuevos en bases de datos populares como el conjunto de datos (CD) de esteroides, el CD del NCI (*National Cancer Institute*), las bases de datos WDI (*World Drug Index*) y MDDR (*MACCS Drug Data Report*), o estableciendo metodologías concisas. Desafortunadamente, la comunidad científica internacional no ha adoptado ningún CD estándar para la comparación de medidas de similitud y DMs, probablemente por la imposibilidad de encontrar un grupo único de moléculas que reagrupe todas las necesidades de cribado de la Quimioinformática moderna (8). Por este motivo se ha sugerido que, para validar un método nuevo, los investigadores deben presentar al menos 10 conjuntos con actividades diversas con más de un estándar de comparación (90).

1.2.2.3. Espacio químico y representación molecular.

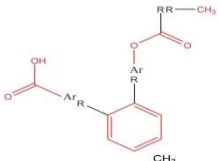
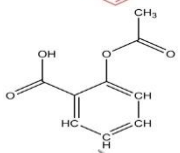
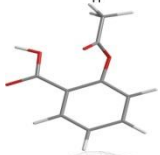
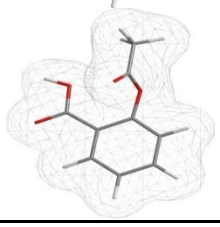
Cercanamente aliado con la noción de similitud molecular es el de *espacio químico*. Los espacios químicos proveen un medio para conceptualizar y visualizar la similitud molecular. El concepto de espacio químico se deriva de la noción de espacio usado en Matemáticas y consiste en un conjunto de moléculas y proximidades asociadas entre las moléculas, lo cual le da al espacio cierta "estructura" (91).

El espacio químico se puede describir usando una codificación *basada en coordenadas* o una codificación *libre de coordenadas* de las estructuras químicas. En la codificación individual de moléculas (espacio basado en coordenadas), cada molécula se describe mediante un vector de fragmentos o subestructuras, traducido posteriormente en un vector de DMs y, por tanto, tiene una posición absoluta en un espacio multidimensional. Por otra parte, en la codificación por pares de moléculas (espacio libre de coordenadas) solo se calculan las distancias entre dos moléculas usando una medida de similitud explícita o implícita. La posición absoluta de las moléculas en este espacio se puede calcular solamente si se miden todas las distancias por pares y se conoce la dimensionalidad del espacio (92-93).

De acuerdo a la naturaleza en su definición y a la complejidad de los rasgos moleculares estructurales que se codifican, los DMs se pueden clasificar de forma general según las dimensiones que abarcan: DMs-0D (Descriptores Constitucionales), DMs-1D (Descriptores Unidimensionales), DMs-2D (Descriptores Bidimensionales o Invariantes de Grafos), DMs-3D (Descriptores Tridimensionales), DMs-4D (Descriptores Tetradimensionales) y DMs-*n*D (Descriptores n-dimensionales, estando *n* desde 4-7).

Los DMs-0D son descriptores que se obtienen directamente de la fórmula molecular y son independientes de cualquier conocimiento sobre la estructura molecular. Los DMs-1D están basados en la representación unidimensional de la molécula o representación que consiste en una lista de fragmentos estructurales de la misma. Los DMs-2D se basan en la representación bidimensional o *topológica* de la molécula, o sea, que consideran la conectividad de los átomos (vértices) en la molécula (pseudografo) en términos de la presencia y naturaleza de los enlaces químicos (aristas). Los DMs-3D son derivados de la representación tridimensional de la molécula y se basan no solo en la naturaleza y conectividad de los átomos, sino también en la configuración espacial de la misma. Finalmente los DMs-nD son descriptores basados no solo en la configuración espacial de la molécula, sino también en los campos escalares de interacción que se originan como consecuencia de la distribución electrónica en dicha entidad química, cuando se consideran los átomos del receptor, las moléculas de agua en las vecindades del ligando y receptor, etc (94). La presentación que se muestra en la **Tabla 1.2** procura solo ejemplificar y no pretende ser representativa, los lectores interesados en ampliar este tema pueden consultar la referencia (80).

Tabla 1.2. Relación entre la dimensionalidad de los descriptores y la complejidad de la representación que describen.

Dimensión	Representación Típica	Descriptores Típicos
0D	C9H8O4	Peso molecular, Conteo de átomos
1D		Conteo de grupos funcionales
2D		Descriptores topológicos
3D		Descriptores geométricos
4D		Energía de interacción

1.2.2.3.1. Selección de rasgos.

A medida que la dimensionalidad de los datos se incrementa, muchos tipos de análisis de datos y problemas de clasificación se vuelven computacionalmente difíciles. En la literatura, este fenómeno se refiere como *la maldición de la dimensionalidad* (95). Para propósitos de búsqueda de similitud, el aspecto más relevante de la maldición de la dimensionalidad concierne a la medida de distancia o similitud. Para ciertas distribuciones de datos, la diferencia relativa entre las distancias de los puntos más cercanos y lejanos a un punto, independientemente seleccionado, tiende a cero a medida que la dimensionalidad aumenta (96). Por otra parte, un número grande de descriptores en la representación pueden contener rasgos irrelevantes o débilmente relevantes, que se conoce afectan negativamente la exactitud de los algoritmos de predicción (97), el caso extremo de este fenómeno se ilustra en *el teorema del patito feo* de Watanabe; básicamente, si se considera

el universo de rasgos de los objetos y no se tiene algún sesgo cognitivo acerca de cuáles de ellos son mejores, no importa cuales dos objetos se compare, todo resultará igualmente similar (disimilar) (98). Una estrategia para solucionar esta dificultad es seleccionar un conjunto de descriptores en particular para los cuales se demostró que funcionan bien en un problema determinado. Otra estrategia es calcular primero un gran número de descriptores y luego eliminar aquellos descriptores del conjunto que muestran un coeficiente de correlación por encima de cierto valor umbral. Un enfoque diferente es seleccionar de forma automática (computacional) la combinación óptima de descriptores para el problema en cuestión (99).

Numerosos métodos automáticos se han propuesto en Quimioinformática para la selección de rasgos y/o reducción de la dimensionalidad, por ejemplo, la técnica paso a paso de los procesos de integración hacia adelante o eliminación hacia atrás y el análisis de componentes principales (100); también ha sido propuesto el uso de los *k*-vecinos más cercanos (101), entre muchos otros. Una fuente excelente que aborda el tema de la selección de rasgos en el contexto del Aprendizaje Automático lo constituye la revisión de Guyon y Elisseeff (102). Un buen número de estas técnicas aparecen implementadas en el software de Aprendizaje Automático y Minería de Datos disponible gratuitamente Weka (103).

1.2.2.4. Métricas de desempeño.

Existe un debate en curso en la literatura sobre "puntajes de mérito" adecuados (o indicadores de desempeño) para evaluar los ensayos de VS retrospectivos. Una métrica popular es el *factor de enriquecimiento* porque es intuitivo y sencillo de interpretar; sin embargo, este depende de un valor de corte arbitrario, por lo general el 1 o 5% de la base de datos a cribar. Nicholls (104) aboga firmemente por el uso de medidas estándares como el área bajo la curva de las características del operador receptor AUC[ROC]. Sin embargo, Truchon y Bayly (105) detectaron que la curva ROC no tiene en cuenta explícitamente el llamado *problema de la detección temprana*, esto es, la propiedad de un método para recuperar compuestos activos "tempranamente" al principio de la

lista de clasificación. Específicamente, este fenómeno se ejemplifica en tres situaciones donde el algoritmo de búsqueda: 1-) ranquea la mitad de los candidatos positivos al principio de la lista y la mitad al final, 2-) distribuye los candidatos positivos uniformemente por toda la lista, 3-) ranquea todos los candidatos positivos exactamente en la mitad de la lista; para todos los casos anteriores AUC[ROC] = 0.5. En este sentido, estos autores desarrollaron un mejoramiento de la curva ROC a través de la métrica Discriminación de la Curva ROC Mejorada por la Distribución de Boltzmann (BEDROC, del inglés Boltzmann-Enhanced Discrimination of ROC), que utiliza una ponderación exponencial para asignar mayor peso a la detección temprana. Del mismo modo, se ha sugerido el escalado semilogarítmico de la curva ROC, pROC (106).

1.2.3. Búsqueda de Similitud como una de las Técnicas principales de VS basadas en (di)similitud.

La búsqueda de similitud es una de las técnicas de VS más simples, en la cual una estructura bioactiva conocida se usa como consulta frente a una base de datos para identificar las moléculas vecinas más cercanas, que al mismo tiempo son las más probables que exhiban la bioactividad de interés (131).

1.2.3.1. Algoritmos de emparejamiento o "matching".

El concepto de emparejamiento "matching" exacto o parcial, y los algoritmos de búsqueda de emparejamiento, son ampliamente utilizados en sistemas de información química asistidos por ordenadores con el propósito de buscar una subestructura idéntica. Una facilidad menos común es la provisión para la búsqueda del mejor par, o vecino más cercano, en la cual se recuperan las estructuras más similares a una estructura de consulta, donde la similitud se define sobre la base de algún coeficiente de similitud o de distancia, que refleja el número de fragmentos comunes de la consulta y de una molécula de la base de datos (132). El problema general de encontrar los mejores pares se define por Friedman y coautores (133) como: "... dado un conjunto de *m*

instancias (cada uno de los cuales es descrito por *n* atributos con valores reales) y una medida de (di)similitud, encontrar las *k* instancias más cercanas a la instancia de consulta con los atributos especificados…".

Un algoritmo eficiente del vecino más cercano será uno que evite el cálculo de la mayoría de las distancias, calculando solamente aquellas que rodean la instancia o estructura de consulta (134-138). Para problemas quimioinformáticos, Baldi y coautores (139) plantean un algoritmo diferente a los tradicionales, que consiste en almacenar para cada molécula A de la base de datos no solamente su vector correspondiente $\vec{A}$, sino también almacenar información adicional contenida en un pequeño vector $\vec{a}$, de tamaño *n* (esto es, si $\vec{A}$ tiene tamaño $N = 2^p$entonces el tamaño de $\vec{a} es\ n = p$). El vector $\vec{a}$ se obtiene aplicando el operador lógico de disyunción exclusiva XOR al vector $\vec{A}$. Esta información adicional permite explorar menos del 50% de la base de datos y acelera la búsqueda significativamente. Más recientemente, Cao y coautores (140) han reportado un algoritmo de búsqueda basado en técnicas de imbibición e indexado multidimensional que mejora en 20-400 veces a los métodos secuenciales en cuanto al tiempo de búsqueda de los 100 primeros vecinos más cercanos (el algoritmo de Baldi y coautores (139) los mejora en 5.5 veces) en conjuntos de datos de 260 000-19 millones de compuestos, mientras que mantiene exactitudes comparables. Además, este algoritmo es aplicable a un amplio espectro de medidas de similitud y puede ser escalable a conjuntos de datos de hasta cientos de millones de objetos químicos.

1.2.3.2 Fusión de datos

Como Sheridan y Kearsley (90) han señalado, es muy poco probable que un solo mecanismo de búsqueda pueda comportarse consistentemente superior a los demás en todos los problemas (nuevamente se "recuerda" el teorema "*No free lunch*"). Por esta razón, tiene sentido aplicar técnicas de búsqueda complementarias y combinar los resultados individuales en un *resultado*

consenso para extender el dominio de problemas con resultados satisfactorios, este enfoque se ha dado a conocer en los últimos años como *fusión de datos* (141).

Básicamente, existen tres técnicas de fusión de datos y una de estas es la *fusión de similitud*, que implica la búsqueda con una estructura de referencia y varias medidas de similitud. Otra variante es la *fusión de grupo*, que consiste en buscar múltiples estructuras de referencia con una sola medida de similitud y se ha mostrado que es más eficaz que la fusión de similitud. El tercer enfoque es la *turbo similitud*, en analogía a los motores turbos que reutilizan los gases de escape y le imprimen una potencia mayor al vehículo. Esta técnica utiliza una estructura de referencia y una medida de similitud, sin embargo, es más efectiva que la *búsqueda simple* porque utiliza los primeros vecinos más cercanos recuperados como estructuras de referencias, ya que estos es probable que también sean bioactivos y al mismo tiempo introducen otros rasgos estructurales que aumentan el éxito de la búsqueda al encontrar otros quimiotipos en el espacio químico (141).

Adicionalmente a las técnicas de fusión de datos anteriores, algunos investigadores han trabajado la ponderación de rasgos binarios orientados por clases de actividad sobre la base de compuestos de referencia múltiples y aplicados para enfatizar algunas posiciones de *bits* específicas durante la búsqueda de similitud. En esencia, estas técnicas también pudieran considerarse como una cuarta estrategia de fusión de datos, más específicamente *fusión de representación*, *y*, actualmente constituyen un área de investigación activa en Quimioinformática (28).

1.3. Consideraciones finales del capítulo.

Actualmente, el proceso de descubrimiento y desarrollo de fármacos es complejo y costoso, por lo cual las técnicas de VS surgen como una alternativa prometedora para palear estas dificultades en las primeras fases de este proceso. Mención especial merecen las técnicas de cribado basadas en similitud por su complejidad matemática y computacional relativamente baja, ya que solamente "demandan" de una medida de proximidad, un algoritmo de búsqueda de similitud y algunos otros elementos comunes con las otras técnicas como conjuntos de datos debidamente curados y

representados por descriptores numéricos informativos seleccionados de acuerdo al principio de vecindad o similitud. En el caso de los objetos químicos, generalmente hablamos de una evaluación de similitud estructurada ya que la comparación comprende tanto los fragmentos de las biomoléculas como las relaciones entre los mismos dadas por los enlaces químicos. Por tanto, es de esperar que bio(macro)moleculares similares (disimilares) tengan funciones bioquímicas similares (disimilares) y cualquier desviación de esta regularidad tenga más que ver con una mala conducción en la etapa de la selección de rasgos.

2. MÉTODOS COMPUTACIONALES

2. MÉTODOS COMPUTACIONALES

En este capítulo se discute la planificación general de la investigación, basada en la introducción de algunas medidas novedosas de similitud y búsqueda de similitud para su validación en Quimioinformática, lo cual es centro de la presente tesis.

2.1 Planificación general de la investigación.

La presente investigación fue planificada en tres fases principales: i-) Estudio preliminar sobre la introducción de algunas medidas novedosas de (di)similitud de otras áreas del conocimiento en técnicas de búsqueda de similitud; ii-) Introducción de medidas de similitud novedosas basadas en el *acuerdo relacional* y su validación en la búsqueda de similitud; iii-) Análisis teórico de la generalización de las medidas de similitud por pares del acuerdo relacional al caso de varios objetos químicos y derivación de nuevas medidas *k*-arias para su uso en la Quimioinformática.

Para materializar toda la investigación fue necesario usar productos de software para las técnicas de VS, conjuntos de datos, programas de cálculo de DMs y algunas otras técnicas de selección de rasgos y validación, que en algunos casos se reutilizaron en las distintas fases. A continuación, se ofrece con un grado mayor de detalle de cómo fueron concebidas cada una de las fases:

2.1.1. Estudio preliminar sobre la introducción de algunas medidas novedosas de (di)similitud de otras áreas del conocimiento en técnicas de búsqueda de similitud.

Se concreta en este epígrafe cómo se usan los conceptos esbozados en el Marco Teórico para la búsqueda de similitud.

2.1.1.1. Búsqueda de similitud.

El propósito de este estudio fue introducir dos medidas de similitud novedosas procedentes del área de la Ecología [manual de usuarios de programa SYN-TAX (142)]. Las bases de datos estructurales usadas fueron las de Sutherland y coautores (143) (ver **Tabla 2.1**). Como

características de los datos originales, estos fueron obtenidos aplicando criterios cuantitativos de diversidad molecular, derivando finalmente en repositorios fiables y diversos, aunque relativamente pequeños y balanceados.

Tabla 2.1. Descripción de los conjuntos de datos procedentes de la Química Médica

Datos[a]	Diana farmacológica	Número de compuestos	Variable farmacodinámica	Rango de valores
ACE	Inhibidores de la enzima convertidora de angiotensina	114	pIC50	2.1-9.9
AchE	Inhibidores de la acetilcolinesterasa	111	pIC50	4.3-9.5
BZR	Ligandos para el receptor de la benzodiacepina	163	pIC50	5.5-8.9
COX-2	Inhibidores de la ciclooxigenasa	322	pIC50	4.0-9.0
DHFR	Inhibidores de la hidrofolato reductasa	397	pIC50	3.3-9.8
GPB	Inhibidores de la glicógeno fosforilasa b	66	pKi	1.3-6.8
THER	Inhibidores de la termolisina	76	pKi	0.5-10.2
THR	Inhibidores de la trombina	88	pKi	4.4-8.5

[a]Los conjuntos de datos se presentan en el orden de la fuente original (143). Los mismos se pueden acceder libremente en http://www.cheminformatics.org/datasets/index.shtml; [b]*pIC50* = - *logIC50*, donde *IC50* representa la mitad de la concentración inhibitoria máxima, se usa como una medida de la potencia del fármaco, pK_i = - $logK_i$, donde K_i representa la constante de inhibición del fármaco, también se usa como una medida de la potencia del fármaco.

Los conjuntos de datos estructurales fueron editados con el software JChemfor Excel(144). Cada uno de los mismos fue "reoptimizado" con el software generador de estructuras 3D CORINA (145). Los parámetros más relevantes fijados en este proceso para la curación de las estructuras químicas fueron: *wh*, escribir átomos de hidrógeno adicionados; *rs*, eliminar pequeños fragmentos; *neu*, neutralizar cargas formales. Los ficheros resultantes fueron cargados en el software para cálculo de DMs, DRAGON (146); todas las familias de descriptores fueron calculadas (3224 índices en total) y entonces los rasgos binarios fueron eliminados. Los ficheros resultantes fueron cargados en el software para minería de datos Weka (103), donde posteriormente fueron sujetos a un tratamiento que incluye los procesos de prefiltrado, rellevado a escala y selección de rasgos. En esta etapa, los atributos nominales fueron eliminados con el filtro "RemoveType", los atributos que no varían mucho o en lo absoluto fueron eliminados con el filtro "RemoveUseless", manteniendo los parámetros por defecto y los atributos numéricos resultantes así como la clase o

variable respuesta fueron estandarizados con el filtro "Standardize". Luego se llevó a cabo la selección de rasgos con el filtro "AtributeSelection"; en esta etapa "CfsSubsetEval" fue fijado con el resto de los parámetros por defecto, el mismo selecciona subconjuntos de rasgos altamente correlacionados con la clase mientras que guardan una baja correlación entre ellos (147). Según estudios previos, solo los descriptores linealmente dependientes con la clase satisfacen el *principio de vecindad* (o similitud) (41-42).

La medida de efectividad del recobrado fue el *factor de enriquecimiento* definido como el porcentaje de activos por delante de un valor de corte prefijado que en este caso consistió en los primeros 10 candidatos de la lista de recuperación. Las pruebas de comparación múltiple por lote y pares fueron las clásicas de Friedman y Wilcoxon, respectivamente (33).

2.1.2. Introducción de medidas de similitud novedosas basadas en el Acuerdo Relacional y su validación en la búsqueda de similitud.

El propósito de esta fase de investigación fue introducir una nueva clase de medidas de similitud, sistematizadas en la familia de los coeficientes de *acuerdo relacional* e integrarlos a la técnica de búsqueda de similitud. Una primera etapa consistió en un estudio comparativo preliminar de estas medidas con un grupo de otras medidas reportadas en la literatura quimioinformática empleando para ello los conjuntos de datos de Sutherland y coautores (143) para la validación; pero también se usó como confirmación el conjunto de datos estructurales anti-VIH del NCI, que refleja más cercanamente las características reales de un conjunto de cribado masivo experimental. Este conjunto de referencia internacional proviene del cribado experimental antiviral frente al VIH (virus de inmunodeficiencia en humanos) de mayo del 2004, el cual contiene los resultados para 43850 compuestos probados *in vitro* con el objetivo de proteger las células CEM humanas (derivadas de los linfocitos T, tipo de glóbulos blancos) del VIH (causante del SIDA). Los datos de cribado disponibles se encuentran divididos en tres clases: activos confirmados (CA),

moderadamente activos (CM) y los inactivos confirmados (CI). En nuestro estudio, las primeras dos clases se combinaron en una sola clase "activa", que representa el 4.13% de los datos estructurales (83).

La curación, optimización y el cálculo de los descriptores se llevó a cabo exactamente como en la fase *i-)* (ver **Sección 2.1.1.1**). En la etapa de acondicionamiento de los datos, en el software Weka no se eliminaron los rasgos de escasa variabilidad pues estos pudieran resultar muy relevantes para modelar las clases de los compuestos del NCI que es altamente desbalanceada. Para la selección de rasgos, en el trabajo con el filtro "AtributeSelection" se mantuvo el evaluador "CfsSubsetEval" usando la heurística de búsqueda "BestFirst" pero la dirección de búsqueda "backward" como fue recomendado por Guyon y Elisseff (102) para obtener los descriptores más cercanos al óptimo de efectividad y en un tiempo razonable.

En una segunda etapa de validación exhaustiva, las medidas de acuerdo relacional se compararon con una gama más amplia de medidas de proximidad reportadas que en la primera etapa. Además, se usaron conjuntos de validación más adecuados, los datos MUV propuestos por Rohrer y coautores (151). Estos conjuntos de datos fueron construidos usando estrategias de diseño experimental monitoreadas por funciones de análisis de los *vecinos más cercanos refinados* en 17 repositorios estructurales, que contienen información de cribado experimental extraída del proyecto PubChem (ver **Tabla 2.2**) (82). Cada conjunto de datos diseñado consiste en 30 compuestos activos lo más disímiles posibles y un subconjunto de fondo de 15000 señuelos de activos o "decoys". Este procedimiento fue propuesto para evitar un desempeño inflado de los clasificadores en el VS causado por el *enriquecimiento artificial*, el *sesgo de análogos*, y el *desbalance de clases inapropiado*, los cuales han sido factores encontrados usualmente en los conjuntos de referencia internacional.

Tabla 2.2. Descripción bioquímica de conjuntos de datos de la base de datos de PubChem.

Diana farmacológica	Modo de interacción	Clase objeto	Ensayo primario (AID)[a]	Ensayo confirmatorio (AID)[b]	Tipo de ensayo
S1P1 rec.	agonistas	GPCR	449	466	Gen reportero
PKA	inhibidores	quinasa	524	548	enzima
SF1	inhibidores	receptor nuclear	525	600	Gen reportero
Rho-Kinase2	inhibidores	quinasa	604	644	enzima
HIV RT-RNase	inhibidores	ARNasa	565	652	enzima
Eph rec. A4	inhibidores	receptor de tirosín quinasa	689	689	enzima
SF1	agonistas	receptor nuclear	522	692	gen reportero
HSP 90	inhibidores	Chaperona	429	712	enzima
ER-α-Coact. Bind.	inhibidores	PPI	629	713	enzima
ER-β-Coact. Bind.	inhibidores	PPI	633	733	enzima
ER-a-Coact. Bind.	potenciadores	PPI	639	737	enzima
FAK	inhibidores	quinasa	727	810	enzima
Cathepsin G Catepsina G	inhibidores	proteasa	581	832	enzima
FXIa	inhibidores	proteasa	798	846	enzima
FXIIa	inhibidores	proteasa	800	852	enzima
D1 rec.	moduladores alostéricos	GPCR	641	858	Gen reportero
M1 rec.	Inhibidores alostéricos	GPCR	628	859	Gen reportero

[a]Identificador de ensayo PubChem del cribado experimental primario donde se han probado los activos y señuelos. [b]Identificador PubChem para el ensayo de confirmación de las moléculas activas encontradas en el cribado primario.

Similarmente a la primera etapa, el conjunto de datos MUV se usó para confirmar la efectividad de las medidas de proximidad propuestas. Al igual que en la etapa anterior, se mantuvieron los mismos parámetros para la curación, "optimización" y representación (cálculo de descriptores, acondicionamiento de datos, selección de rasgos).

2.1.2.1. Curva ROC concentrada.

Para superar los inconvenientes observados en el uso de la curva ROC en la *detección temprana*, Swamidass y coautores (152) propusieron la curva ROC Concentrada (CROC, del inglés Concentrated ROC) que consiste en magnificar uno de los ejes de la curva ROC [*X* representa la

razón de falsos positivos (*fpr*), *Y* representa la razón de verdaderos positivos (*tpr*)] a través de una transformación de magnificación suave ya sea exponencial, de potencia o logarítmica. Mediante resultados gráficos y empleando pruebas estadísticas robustas los autores concluyen que las variantes CROC son más potentes que los métodos de umbrales de corte fijo, que las variantes de Curva de Acumulación Concentrada (CAC, del inglés Concentrated Acumulation Curve), pROC y ROC.

La curva CROC se obtiene aplicando la transformación de magnificación exponencial del eje *X* (*fpr*) de la curva ROC, esta transformada viene dada por:

$$h(x) = \frac{1-e^{-\delta x}}{1-e^{-\delta}} \quad \textbf{(2.1)}$$

donde, δ es el factor de magnificación, que para caso recomendado toma el valor δ = 20, el cual se corresponde aproximadamente a un 8% de enriquecimiento temprano (123).

Una vez establecida la función de magnificación $h(x)$, el área bajo la curva CROC se puede calcular como el promedio de los valores de *fpr* correspondientes a los rangos de las instancias positivas como:

$$AUC[CROC] = \frac{\sum_{i=1}^{n}[1-h(fpr_i)]}{n} \quad \textbf{(2.2)}$$

donde, fpr_i es la razón de falsos positivos al nivel (rango) de cada instancia positiva *i* del total *n*.

Por último, los valores del área bajo CROC se pueden comparar con el valor correspondiente al clasificador aleatorio, que se calcula por la fórmula:

$$AUC[CROC]_{aleat} = \frac{1}{\delta} - \frac{e^{-\delta}}{1-e^{-\delta}} \quad \textbf{(2.3)}$$

Entonces, para δ = 20 el valor correspondiente para el clasificador aleatorio es de $AUC[CROC]_{aleat}$ = 0.2809

2.1.2.2. Pruebas estadísticas de comparación múltiple.

Tanto los valores del factor de enriquecimiento promedio como los valores de AUC[CROC] promedio se usaron para llevar a cabo comparaciones múltiples entre series de modelos de

proximidad según su desempeño en la búsqueda de similitud usando el estadístico de Iman y Davenport (153). Ellos demostraron que la χ^2 de Friedman es indeseablemente conservativa y obtuvieron un mejor estadístico:

$$F_F = \frac{(N_{ds}-1)\chi^2}{N_{ds}(N_{pm}-1)-\chi^2} \quad \textbf{(2.4)}$$

que se distribuye de acuerdo a la *F* de Snedecor con $N_{pm} - 1$ y $(N_{pm} - 1)(N_{ds} - 1)$ grados de libertad; donde, N_{pm} es el número de modelos de proximidad a comparar y N_{ds} es el número de conjunto de datos.

En caso de que se detecten diferencias significativas, la prueba *post hoc* de Holm puede usarse para comparar la mejor medida ranqueada por la prueba de Friedman (medida de control) con las restantes. Esta prueba representa un mejoramiento a la prueba de Wilcoxon porque permite controlar el error tipo I de la *familia de hipótesis*, que surge de la comparación múltiple por pares.

Básicamente, este método funciona en forma de etapas que se presentan a continuación:

1. Ordenar crecientemente los modelos de proximidad con respecto al mejor usando el puntaje estandarizado:

   $$Z_s = \left[R_j - \left(R_j\right)_{min}\right] / \sqrt{N_{pm}(N_{pm}+1)/6N_{ds}} \quad \textbf{(2.5)}$$

   donde, R_j es el valor del ranqueo promedio asignado por la prueba de Friedman al modelo de proximidad j y $\left(R_j\right)_{min}$ es el menor de tales valores, asignado al modelo de proximidad mejor ranqueado.

2. Calcular la probabilidad asociada a Z_s, asumiendo que la misma se distribuye por una función normal estandarizada.

3. Comparar el valor de probabilidad observada, p_j, con la significación ajustada α/j, donde $j = \overline{1, N_{pm}}$ y $\alpha = 0.05$, y $p_{N_{pm}}$ es la probabilidad asociada a la mayor diferencia con respecto al mejor modelo (que se mantiene fijo).

4. Tan pronto como cierta hipótesis nula o de no diferencia deba ser rechazada, todas las hipótesis anteriores (y los modelos de proximidad correspondientes incluyendo al mejor modelo) son retenidas y las demás descartadas pues serían las peores.

Una prueba más potente que la prueba de Holm es la prueba de Li de dos etapas (156), que modifica los pasos 3 y 4 de la prueba de Holm mediante el algoritmo:

3. Rechazar todas las H_i si $p_1 \leq \alpha = 0.05$. De otro modo, aceptar la hipótesis asociada a p_1 e ir al paso 4.
4. Rechazar cualquier hipótesis remanente H_i con $p_i \leq [(1 - p_1)/(1 - \alpha)]\alpha$.

Como resultado, cualquiera de estas pruebas devuelve una lista de medidas de proximidad en orden decreciente de desempeño, con diferencias estadísticamente no significativas con respecto a la mejor medida (157-158).

2.1.2.3. Implementación computacional del diseño estadístico experimental

La búsqueda de similitud se materializó con el software FUSE, el cual también constituye un producto de software que es otro resultado de la investigación del Laboratorio de Bioinformática de la UCLV (ver **Fig. 2.1**). Básicamente, el software FUSE 1.0 se compone de un conjunto de herramientas de VS para aplicaciones quimioinformáticas, con el fin de realizar búsqueda de similitud en conjuntos de datos quimioinformáticos. Incluye, además, múltiples criterios de búsqueda como la fusión de datos, que está estrechamente relacionado con el nombre del software FUSE o "fusionar". La idea general de esta aplicación consiste en recuperar las moléculas del conjunto de datos que son más similares a una estructura de consulta seleccionada internamente, utilizando una definición cuantitativa de la similitud estructural intermolecular. La versión actual ejecuta un experimento simulado de autoentrenamiento para seleccionar la medida de (di)similitud con mejor desempeño, que posteriormente puede ser usada para ejecutar búsquedas en otros repositorios con una finalidad común.

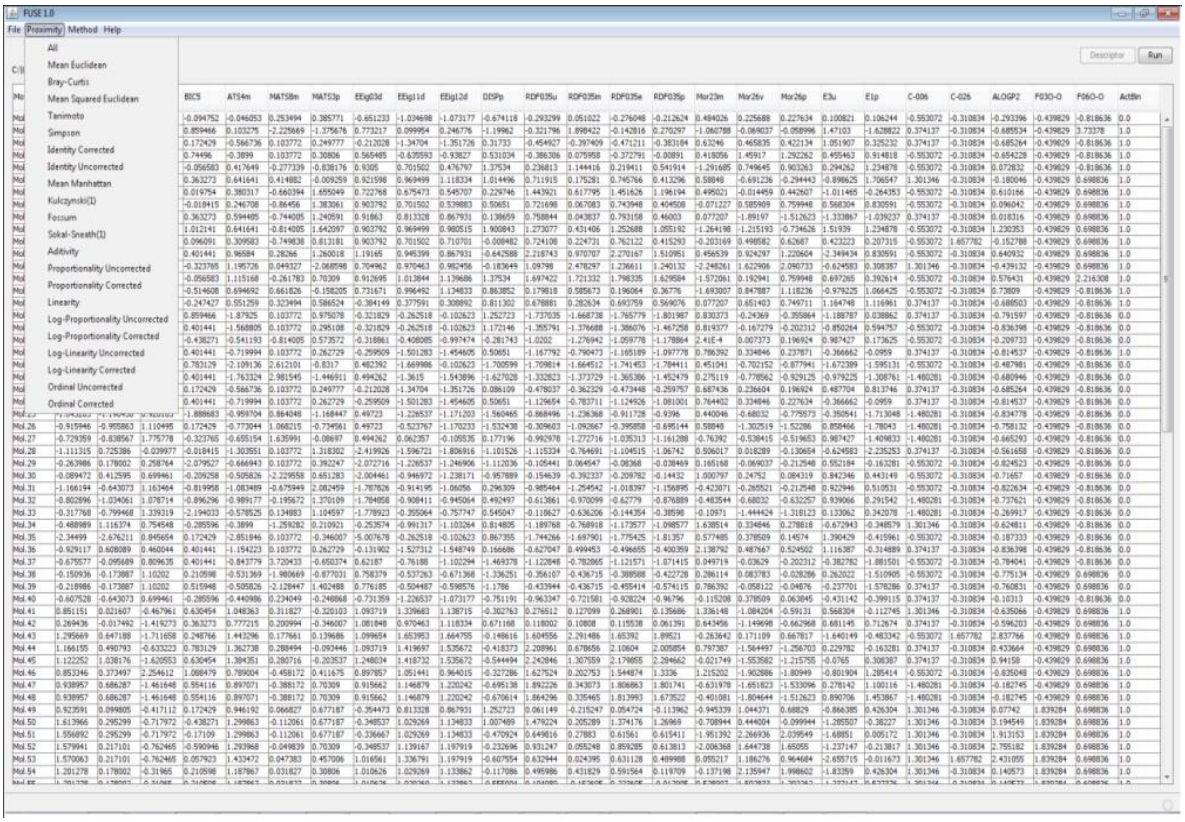


Figura 2.1. Interfaz gráfica del software FUSE

2.1.3. Análisis teórico de la generalización de las medidas de similitud por pares del acuerdo relacional al caso de varios objetos químicos y derivación de nuevas medidas *k*-arias para su uso en la Quimioinformática.

El propósito de esta etapa es el análisis teórico de la generalización de las medidas de similitud bivariadas al caso de varios objetos químicos, la relación funcional entre los coeficientes multivariados y sus contrapartes por pares y la derivación de nuevas medidas multivariadas para su uso en el campo de la Quimioinformática.

2.2. Consideraciones finales del capítulo.

Se enumeraron y describieron brevemente las tres fases que componen nuestra investigación. De forma general, se realizó un estudio preliminar sobre la introducción de nuevas medidas de similitud basadas en el concepto de acuerdo relacional procedentes del área de la Quimioinformática y la Ecología junto con otras medidas de semejanza molecular reportadas en la

literatura. Se seleccionaron veintiséis conjuntos de datos internacionales de la Química Medicinal, uno de los cuales resulta de gran interés para el laboratorio de Bioinformática. Las bases de datos fueron representadas por descriptores seleccionados mediante una técnica de Aprendizaje Automático que es consistente con el *principio de similitud*. Se seleccionó el AUC[CROC] como medida de exactitud por ser la más apropiada para el problema del "reconocimiento temprano". Se propuso un diseño experimental estadístico junto a una estrategia computacional definida para llegar a los resultados finales.

3. RESULTADOS Y DISCUSIÓN

3. RESULTADOS Y DISCUSIÓN

A lo largo del presente capítulo se brinda la posibilidad al lector de corroborar que efectivamente durante el presente trabajo se introdujeron medidas de similitud que tienen aplicaciones novedosas en Quimioinformática y que logran mejorar el desempeño de los principales algoritmos de clasificación basados en similitud, así como permiten un entendimiento superior de las relaciones de semejanza biomolecular con respecto a las medidas reportadas anteriormente.

En el **Epígrafe 3.1** se introducen dos medidas procedentes del área de la Ecología para búsqueda de similitud de repositorios de la Química Médica. En el **Epígrafe 3.2** se introducen medidas novedosas de acuerdo relacional de naturaleza estadística, usadas en el área de la Psicología, que se integran a algoritmos de búsqueda de similitud; dichas medidas se validan, primeramente, de forma preliminar o parcial y luego de forma exhaustiva.

En el **Epígrafe 3.3** se hace un análisis teórico en cuanto a la generalización de las medidas de acuerdo relacional bivariadas al caso de múltiples objetos biomoleculares y la relación funcional que existe entre ambos tipos de medidas para dos y múltiples objetos.

3.1. Estudio preliminar sobre la introducción de algunas medidas novedosas de (di)similitud de otras áreas del conocimiento en técnicas de búsqueda de similitud.

3.1.1. Análisis de la dimensionalidad de los datos.

3.1.1.1 Conjuntos de datos de la Química Medicinal

Los descriptores linealmente relevantes a cada actividad farmacológica (conjunto de datos de la Química Médica, Sutherland y coautores (143)) seleccionados por el evaluador del Weka "CfsSubsetEval" se muestran en (66). El porcentaje de reducción de dimensionalidad en la etapa de eliminación de los descriptores binarios (selección del Weka) correspondientes a cada conjunto de datos fue ACE 55.18% (98.27%), AchE 56.27% (97.94%), BZR 54.28% (98.98%), COX-2 52.70% (98.62%), DHRF 53.07 (98.94%), GBP 55.89% (99.30%), THERM 56.51% (98.93%), y

THR 56.36% (98.93%), siendo la media geométrica de estos valores 55.01% (98.74%). Estos resultados indican que aproximadamente el 55% de los DMs del software DRAGON (3224 descriptores en total) son binarios y, por tanto, fueron descartados en este estudio. Los pasos subsecuentes de limpieza y selección de los descriptores restantes usando el Weka produjeron una eliminación del 98.74% de rasgos no relevantes. Este valor significativo sugiere un alto grado de especificidad en las relaciones entre las actividades farmacológicas y los rasgos biomoleculares.

3.1.1.2 Conjunto de datos estructurales anti-VIH del NCI y MUV

En base a los resultados obtenidos anteriormente se decidió no calcular el total de 3224 descriptores del DRAGON sino solamente calcular las familias 2-D autocorrelacionadas, descriptores RDF (3-D), descriptores 3-D-Morse y descriptores WHIM (3-D) pues resultaron ser las familias predominantes en la modelización de los conjuntos químico medicinales anteriores; por tanto se espera observar (aunque con algún grado de incertidumbre) este mismo comportamiento en el juego de datos anti-VIH del NCI y MUV, que también pertenece a la química medicinal o terapéutica. Este tipo de proceder es consistente con la teoría del *meta aprendizaje* (Biggs, 1985). La cantidad inicial de descriptores considerada fue de 702 y luego del proceso de selección de rasgos con el evaluador *CfsSubsetEval* del Weka se obtuvieron 22 rasgos linealmente relevantes, para un porcentaje de reducción de dimensionalidad de los datos en un 96.87%. Los descriptores seleccionados se muestran en la **Tabla 3.1**. La interpretación para el problema que nos ocupa es equivalente al descrito en la subsección anterior, i.e. estos valores significativos sugieren un grado alto de especificidad en las relaciones rasgos moleculares-actividad farmacológica.

Tabla 3.1 Descriptores seleccionados con el evaluador *CfsSubsetEval* del Weka para el conjunto NAAS y MUV

MATS8v	MATS1p	GATS2v	GATS6p	RDF125m	Mor02m	Mor06m	Mor11m
Mor13m	Mor32m	Mor30v	Mor13e	Mor16e	Mor06p	Mor26p	
E1u	L1e	L1s	H1m	H4m	H8m	R1m	

3.1.2 Desempeño de las medidas de similitud.

El estudio de comparación del desempeño de las medidas de disimilitud fue conducido a través de experimentos de búsqueda de similitud simulada usando el algoritmo de Friedman (133). Para cada repositorio, el 5% de los compuestos clasificados como fármacos fueron muestreados aleatoriamente sin reposición para actuar como las estructuras de referencia o consulta.

Las medidas de disimilitud seleccionadas para este estudio de comparación comprendieron dos ya conocidas, Euclídea y Correlación (complemento del coeficiente de Pearson), y otras dos usadas normalmente en el área de la Ecología, la medida Bray-Curtis II y Ruzicka, que en el momento de este estudio eran novedosas en Quimioinformática. Las fórmulas correspondientes se muestran a continuación:

1. Euclídea

$$d_1(X,Y) = [\sum_{i=1}^{n}(x_i - y_i)^2]^{1/2} \quad \textbf{(3.1)}$$

2. Correlación

$$d_2(X,Y) = 1 - \frac{\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})}{\left[\sum_{i=1}^{n}(x_i-\bar{X})^2 \sum_{i=1}^{n}(y_i-\bar{Y})^2\right]^{1/2}} \quad \textbf{(3.2)}$$

3. Bray-Curtis II

$$d_3(X,Y) = 1 - 2\frac{\sum_{i=1}^{n} min\{x_i,y_i\}}{\sum_{i=1}^{n}(x_i+y_i)} \quad \textbf{(3.3)}$$

4. Ruzicka

$$d_4(X,Y) = 1 - 2\frac{\sum_{i=1}^{n} min\{x_i,y_i\}}{\sum_{i=1}^{n} max\{x_i,y_i\}} \quad \textbf{(3.4)}$$

donde, *X* e *Y* representan los vectores moleculares comparados y x_i e y_i las componentes de los vectores correspondientes o DMs.

Los resultados obtenidos a partir de los experimentos de búsqueda de similitud se muestran en la **Tabla 3.2**. Como se observa, las probabilidades de la prueba de Friedman arrojan diferencias significativas entre las medianas de los factores de enriquecimiento correspondientes a cada medida.

Tabla 3.2. Resultados generales del factor de enriquecimiento para cada medida y comparación entre estas.

Datos	**[1]Euclídea[a]**	**[2]Correlación[a]**	**[3]Bray-Curtis II[a]**	**[4]Ruzicka[a]**
AchE	64.00	70.00	52.00	66.00
ACE	82.00	92.00	30.00	82.00
BZR	73.75	82.50	45.00	63.75
COX-2	69.38	72.50	41.86	68.75
DHFR	82.63	86.84	33.68	80.53
GPB	86.66	96.66	36.66	70.00
THERM	76.66	63.33	23.33	100.00
THR	45.00	55.00	45.00	55.00
Friedman[b]		**0.003****		
Ranq. Prom.[c]	2.75	3.38	1.13	2.75
Wilcoxon[d]	1-2	0.674	**2-3**	**0.018***
	1-3	**0.018***	2-4	0.866
	1-4	0.933	**3-4**	**0.012***

[a]Los números en supraíndice son los identificadores usados para las pruebas múltiples por pares, [b]Significación asintótica para la prueba de comparación múltiple por lotes χ^2 de Friedman. [c]Ranqueo medio asignado por la prueba de Friedman, mientras mayor es este valor, mejor es el desempeño de la medida correspondiente. [d]Significación asintótica para la prueba *post hoc* de Wilcoxon de comparación múltiple por pares, de dos colas. *Pruebas estadísticas significativas ($p < 0.05$); **Pruebas estadísticas altamente significativas ($p < 0.01$).

Con el objetivo de comprobar cuáles medidas causaron esta diferencia, también se ejecutó la prueba de Wilcoxon para cada par de medidas. La observación y análisis de los valores de significación y los ranqueos promedios correspondientes permitieron organizar las medidas por su nivel de desempeño en dos grupos homogéneos, o sea Euclídea = Correlación = Ruzicka>Bray-Curtis II; lo cual indica que las primeras tres medidas son equivalentes en la recuperación efectiva de los conjuntos de datos estudiados, y su desempeño es superior a la medida Bray-Curtis II. También se observa que estas primeras tres medidas mostraron valores de efectividad entre 63-100%, indicando así un desempeño no aleatorio (dado por un valor del 50%), mientras que la

medida Bray-Curtis II arrojó valores de efectividad entre 23-52% indicando un desempeño casual o pobre. Por último, los resultados sugieren que la disimilitud de Ruzicka es comparablemente útil con respecto a las ya tradicionales Euclídea y Correlación en el escenario de la Química Médica.

3.2. Introducción de medidas de similitud novedosas basadas en el Acuerdo Relacional y su validación en la búsqueda de similitud.

3.2.1. Primera etapa: Introducción de medidas de similitud basadas en el Acuerdo Relacional y su validación preliminar en la búsqueda de similitud de repositorios de la Química Médica.

3.2.1.1. Modelos de proximidad.

Las medidas tomadas como referencia de comparación en el presente estudio son las reportadas en el trabajo de Al Khalifa y coautores (32). El conjunto resultante consiste en 12 medidas de proximidad cuyas fórmulas y clasificación se brindan en la **Tabla 3.3**.

Tabla 3.3. Medidas de proximidad como referente de comparación.

Medida	Fórmula[a]	Tipo[b]	Ec[c]
Manhattan Media	$MM_{XY} = \sum_{j=1}^{n} \lvert x_j - y_j \rvert / n$	D	(**3.5**)
Euclídea Media	$EM_{XY} = \sqrt{\sum_{j=1}^{n} \lvert x_j - y_j \rvert^2} \Big/ n$	D	(**3.6**)
Euclídea Cuadrática Media	$ECM_{XY} = \sum_{j=1}^{n} \lvert x_j - y_j \rvert^2 / n$	D	(**3.7**)
Bray-Curtis I	$BC_{XY} = \frac{\sum_{j=1}^{n} \lvert x_j - y_j \rvert}{\sum_{j=1}^{n} (\lvert x_j \rvert + \lvert y_j \rvert)}$	D	(**3.8**)
Tanimoto	$T_{XY} = \frac{\sum_{j=1}^{n} x_j y_j}{\sum_{j=1}^{n} x_j^2 + \sum_{j=1}^{n} y_j^2 - \sum_{j=1}^{n} x_j y_j}$	A	(**3.9**)
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	(**3.10**)
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	(**3.11**)
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	(**3.12**)

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	(**3.13**)
Coseno/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	(**3.14**)
Simpson	$Sim_{XY} = \frac{\sum_{j=1}^{n} min(x_j, y_j)}{min(\sum_{j=1}^{n} x_j, \sum_{j=1}^{n} y_j)}$	A	(**3.15**)
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	(**3.16**)

[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]Número de ecuación usado a lo largo del trabajo.

Las otras medidas de similitud molecular propuestas en nuestro trabajo están basadas en la teoría de Zegers y ten Berge (176), y Zegers (177). Estos autores propusieron una fórmula general para los coeficientes de asociación bivariada correspondientes a las escalas métricas. 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*.

Un coeficiente de asociación entre dos variables tiene que ser invariante a las transformaciones admisibles de las mismas y sensible a transformaciones no admisibles de éstas. Con el objetivo de sistematizar esta idea, estos autores derivaron una fórmula general para los coeficientes de asociación basada en la *versión uniforme* de las variables (ver **Tabla 3.4**)

Tabla 3.4. Relación entre las escalas métricas, transformaciones admisibles y transformaciones uniformadoras de las variables.

Escala Métrica	Transformación Admisible[b]	Versión Uniforme (*U*)
absoluta	$\varphi' = \varphi$	X
diferencia	$\varphi' = \varphi + \alpha$	$X - \bar{X}$
razón	$\varphi' = \beta\varphi$	X/T_X
intervalo	$\varphi' = \beta\varphi + \alpha$	$(X - \bar{X})/S_X$
log-razón	$\varphi' = \varphi^{\gamma}$	X^{1/L_X}
log-intervalo	$\varphi' = \beta\varphi^{\gamma}$	$(X/G_X)^{1/N_X}$
ordinal	$\varphi' = f(\varphi)$	$Rank(X)$

[a]Las escalas 1[ra]-4[ta] fueron estudiadas por Zegers y ten-Berge (176) y las escalas 5[ta]-7[ma] fueron adicionadas por Stine (178). φ y φ' son las funciones original y transformada (homomorfismos), respectivamente; α y γ son números reales, βes un número real positivo, y f es una función estrictamente monótona. [c]$X = \{x_1, x_2, x_3, \ldots, x_n\}$ representa una muestra aleatoria de la variable X; $\bar{X}$ y S_X son la media aritmética y desviación estándar de X, respectivamente; $T_X = +\sqrt{(\sum_{j=1}^{n} x_j^2)/n}$, $L_X = +\sqrt{\{\sum_{j=1}^{n}[ln(|x_j|)]^2\}/n}$, $G_X = +\sqrt[n]{\prod_{j=1}^{n}|x_j|}$, $N_X = +\sqrt{\{\sum_{j=1}^{n}[ln(|x_j|/G_X)]^2\}/n}$; *Rank*, representa la transformación RT-2 de Conover e Iman (186).

Considerando que un coeficiente de asociación entre dos variables puede ser definido en la extensión en la cual sus versiones uniformes son idénticas y asumiendo que la diferencia cuadrática media de las versiones uniformes es un indicador común de tal identidad, o sea:

$$f(U_i, U_j) = 1 - \frac{c\sum_{t=1}^{n}(u_{it}-u_{jt})^2}{n} \quad (\mathbf{3.17})$$

donde, *c* es una constante positiva determinada inequívocamente requiriendo que $f(-U_i, U_j) = -f(U_i, U_j)$; Zegers y ten-Berge (176) arribaron a su fórmula general de asociación en las versiones uniformes de las variables *i* y *j*:

$$g_{ij} = \frac{2\sum_{t=1}^{n} u_{it}u_{jt}}{\sum_{t=1}^{n} u_{it}^2+\sum_{t=1}^{n} u_{jt}^2} \quad (\mathbf{3.18})$$

Luego, usando la transformación para corrección por aleatoriedad:

$$g_{ij}^{'} = \frac{g_{ij}-g_{ij}^{c}}{1-g_{ij}^{c}} \quad (\mathbf{3.19})$$

donde, $g_{ij}^{c} = 2n^{-1}\sum_{t=1}^{n} u_{it}\sum_{j=1}^{n} u_{jt}$ es la esperanza de g_{ij}

Zegers (177) propuso la versión de (**3.18**) corregida por aleatoriedad como:

$$g_{ij}^{'} = \frac{2\left(\sum_{t=1}^{n} u_{it}u_{jt}-n^{-1}\sum_{t=1}^{n} u_{it}\sum_{t=1}^{n} u_{jt}\right)}{\sum_{t=1}^{n} u_{it}^2+\sum_{t=1}^{n} u_{jt}^2-2n^{-1}\sum_{t=1}^{n} u_{it}\sum_{t=1}^{n} u_{jt}} \quad (\mathbf{3.20})$$

Stine (178) hizo una mejora a la teoría reinterpretando la asociación entre dos variables como *acuerdo relacional* (RA, de sus siglas en inglés, Relational Agreement) *entre observadores* (mediciones), el cual se refiere al grado en el cual una tasación es una transformación admisible de otra tasación. Su idea es que un coeficiente de acuerdo relacional debe atenuarse por *desacuerdos con significado*, mas debe ser independiente de *desacuerdos sin significado*. Stine (187) también presentó la organización entre las escalas métricas acorde a su Teoría de la Significatividad como se ilustra en la **Fig. 3.1**.

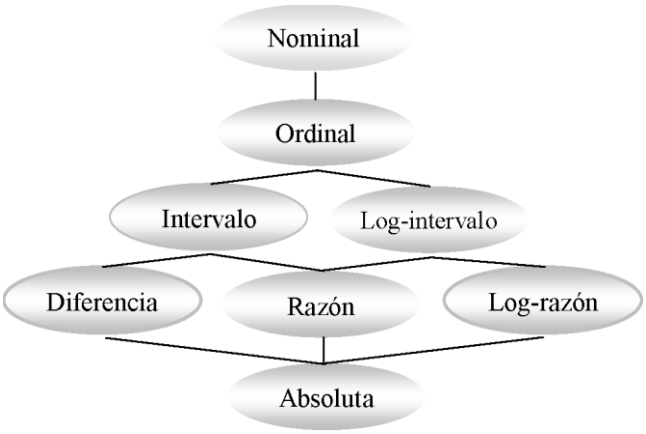


Figura 3.1. Relación entre las escalas métricas. Las aristas dirigidas hacia arriba ilustran las relaciones subconjunto-superconjunto. Las escalas que están en el mismo camino son isomorfas unas con otras.

Finalmente, insertando las transformaciones uniformadoras apropiadas (ver **Tabla 3.4**) en (**3.18**) y (**3.20**), se obtienen los coeficientes no corregidos y corregidos por aleatoriedad de identidad (e), aditividad (*a*), proporcionalidad (*p*), linealidad (*r*), log-proporcionalidad (*w*), log-linealidad (*l*), y monotonicidad (*m*) (ver **Ec. 3.21-3.29** y **Tabla 3.5**).

Las medidas de similitud reportadas en los trabajos de Zegers y ten Berge (176), y Zegers (177) son las siguientes:

Identidad

No corregido: es igual al coeficiente de Dice (ver **Ec. 3.10, Tabla 3.3**)

Corregido

$$e_{XY}^{c} = \frac{2s_{XY}}{s_X^2 + s_Y^2 + \left(\overline{X} - \overline{Y}\right)^2} \quad \textbf{(3.21)}$$

Aditividad

No corregido (es el caso especial del "punto de anclaje" de Winer (179))

$$a_{XY} = \frac{2s_{XY}}{s_X^2 + s_Y^2} \quad \textbf{(3.22)}$$

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 Coseno/Ochiai (ver **Ec. 3.14, Tabla 3.3**)

Corregido

$$p_{XY}^{c} = \frac{s_{XY}}{T_X T_Y - \overline{XY}} \quad \textbf{(3.23)}$$

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

Linealidad

No corregido: es el coeficiente de correlación de Pearson (r_{XY}) (ver **Ec. 3.16, Tabla 3.3**)

Corregido (no cambia la expresión)

Las medidas restantes son las reportadas en el trabajo de Stine (178):

Log-razón o Log-proporcionalidad

No corregido

$$Lp_{XY} = \frac{2\sum_{j=1}^{n} x_j^{1/L_X} y_j^{1/L_Y}}{\sum_{j=1}^{n} x_j^{2/L_X} + \sum_{j=1}^{n} y_j^{2/L_Y}} \quad \textbf{(3.24)}$$

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

Corregido

$$Lp_{XY}^{c} = \frac{2\left[\sum_{j=1}^{n} x_j^{1/L_X} y_j^{1/L_Y} - (1/n)\sum_{j=1}^{n} x_j^{1/L_X} \sum_{j=1}^{n} y_j^{1/L_Y}\right]}{\left[\sum_{j=1}^{n} x_j^{2/L_X} + \sum_{j=1}^{n} y_j^{2/L_Y} - (2/n)\sum_{j=1}^{n} x_j^{1/L_X} \sum_{j=1}^{n} y_j^{1/L_Y}\right]} \quad \textbf{(3.25)}$$

Log-intervalo o Log-linealidad

No corregido

$$Lr_{XY} = \frac{2\sum_{j=1}^{n} x_j^{1/N_X} y_j^{1/N_Y}}{\left(G_Y^{1/N_Y} \Big/ G_X^{1/N_X}\right)\sum_{j=1}^{n} x_j^{2/N_X} + \left(G_X^{1/N_X} \Big/ G_Y^{1/N_Y}\right)\sum_{j=1}^{n} y_j^{2/N_Y}} \quad \textbf{(3.26)}$$

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

$$Lr_{XY}^{c} = \frac{2\left[\sum_{j=1}^{n} x_j^{1/N_X} y_j^{1/N_Y} - (1/n)\sum_{j=1}^{n} x_j^{1/N_X}\sum_{j=1}^{n} y_j^{1/N_Y}\right]}{\left(G_Y^{1/N_Y} \Big/ G_X^{1/N_X}\right)\sum_{j=1}^{n} x_j^{2/N_X} + \left(G_X^{1/N_X} \Big/ G_Y^{1/N_Y}\right)\sum_{j=1}^{n} y_j^{2/N_Y} - (2/n)\sum_{j=1}^{n} x_j^{1/N_X}\sum_{j=1}^{n} y_j^{1/N_Y}} \quad \textbf{(3.27)}$$

Ordinal

No corregido

$$e_{XY}^{R} = \frac{2\sum_{j=1}^{n} R(x_j)R(y_j)}{\sum_{j=1}^{n} R(x_j)^2 + \sum_{j=1}^{n} R(y_j)^2} \quad \textbf{(3.28)}$$

Corregido (es el mismo que la ρ de Spearman)

$$e_{XY}^{Rc} = 1 - \frac{6\sum_{j=1}^{n}\left[R(x_j) - R(y_j)\right]^2}{n(n^2-1)} \quad \textbf{(3.29)}$$

donde, $R(x_j)$ es el ranqueo del atributo x_j relativo a los n descriptores de la molécula X.

Tabla 3.5. Coeficientes bivariados de acuerdo relacional y algunas de sus propiedades.

Escala	Nombre[a]	Fórmula[b]	Fórmula equivalente	Prueba de Inferencia	Rango	ID[c]
absoluta	identidad no corregido	(**3.10**)	- Coeficiente de Dice (32)	-	[-1, +1]	p_{31}
	Identidad corregido	(**3.21**)	- Coeficiente de igualdad de Jobson (191) - Coeficiente de correlación de concordancia de Lin (192)	Estadístico U (193)	[-1, +1]	p_{32}
diferencia	aditividad no corregido y corregido	(**3.22**)	- Coeficiente de McDonald (194)	Estadístico *u* (195-197)	[-1, +1]	p_{33}
razón	proporcionalidad no corregido	(**3.14**)	- Coseno/Ochiai (32) - Coeficiente de Burt (198) - Coeficiente de congruencia de Tucker (199) - Medida *H* de Sjöberg (200)	-	[-1, +1]	p_{34}
	proporcionalidad corregido	(**3.23**)	-	NA	[-1, +1]	p_{35}
intervalo	linealidad no corregido y corregido	(**3.16**)	- CPM de Pearson (32)	Estadístico *t*-Student	[-1, +1]	p_{36}
log- razón	log-proporcionalidad no corregido	(**3.24**)	-	-	[-1, +1]	p_{37}
	log-proporcionalidad corregido	(**3.25**)	-	NA	[-1, +1]	p_{38}
log-intervalo	log-linealidad no corregido	(**3.26**)	-	-	[-1, +1]	p_{39}
	log-linealidad corregido	(**3.27**)	-	NA	[-1, +1]	p_{310}
ordinal	monotonicidad no corregido	(**3.28**)	-	-	[+0.5,+1]	p_{311}
	monotonicidad corregido	(**3.29**)	- CCR de Spearman (ρ)	- Estadístico *t*-Student - Prueba de Kendall (201)	[-1, +1]	p_{312}

[b]Número de ecuación usado en las demostraciones. [c]Identificador para la comparación con los otros modelos de proximidad bivariados.

3.2.1.2. 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 (180). 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 $\alpha = 20$ (105,152), 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 (181).

3.2.1.3. 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 3.6-3.8**).

Tabla 3.6. 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)

Tabla 3.7. 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)

Tabla 3.8. 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").

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 3.9** 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.

Tabla 3.9. Potencia relativa de los modelos de similitud en la recuperación temprana de activos según la prueba de Friedman.

Posición[a]	Modelo	Ranqueo Medio	Posición	Modelo	Ranqueo 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
8[t]	*Tan*	13.61	19	*MM*	6.00
9[*t]	*D*	13.61	20[**]	*Lr*	3.89
10[t]	*SS1*	13.61	21[**]	Lr^c	3.44
11[t]	*Kul1*	13.61			

[a]Cuanto más alto es el ranqueo medio, más potente es el modelo; [*]medidas de similitud planteadas por Al Khalifa y coautores (32), [**]modelos de acuerdo relacional aportados por el presente trabajo, [t] modelos atados por el "Ranqueo Medio".

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, esto es, la matriz binaria de diferencias significativas, se usó como entrada para el algoritmo de agrupamiento de Ward (ver más detalles en ref. (182-183)). En la **Fig.3.2** se observa la formación de dos grandes grupos, según la prueba de Mojena (184). El primer grupo está formado por $G_1 = \{MM, EM, ECM, F, Sim, Lp, Lp^c, Lr, Lr^c\}$ y el otro por $G_2 = \{BC, Tan, D, SS1, Kul1, Cos, r, e^c, a, p^c, e^R, e^{Rc}\}$. Todo esto sugiere la separación de los modelos en un grupo de 12 mejores modelos y otro de 9 peores modelos.

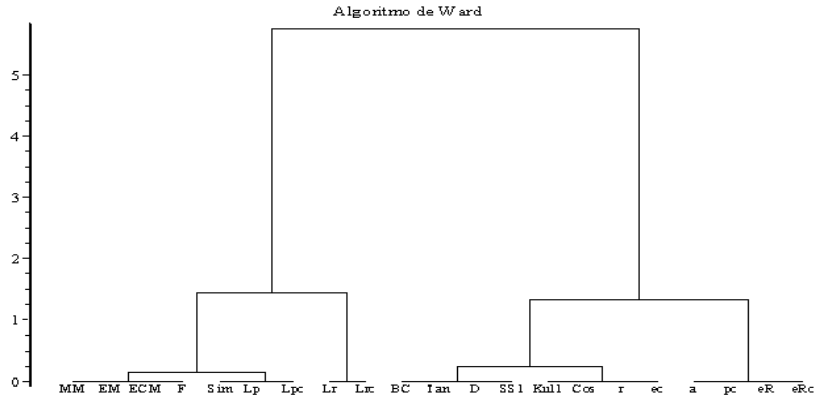


Figura 3.2. Dendrograma obtenido del algoritmo de Ward para el agrupamiento de modelos homogéneos según la prueba de Wilcoxon.

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 3.6-3.8** obtenidos en el repositorio NCI (43 850 compuestos probados *in vitro* contra el VIH en células CEM humanas) 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. 3.3**).

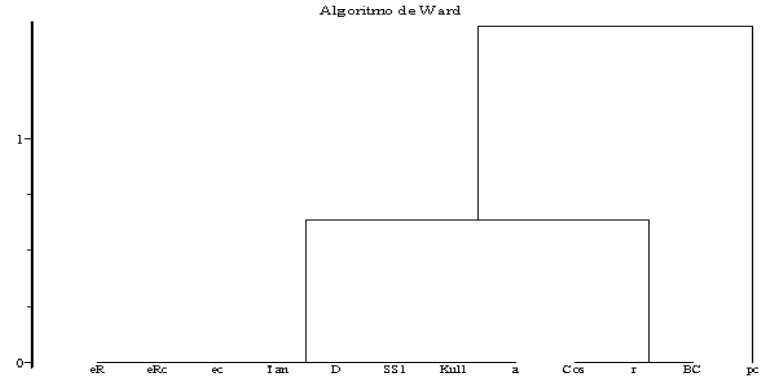


Figura 3.3. Dendrograma obtenido del algoritmo de Ward para el agrupamiento de modelos homogéneos según la prueba *t*-Student con la corrección de Welch.

La **Fig. 3.3** 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 3.1** se muestran las curvas de calidad CROC para el grupo de modelos de proximidad anteriores.

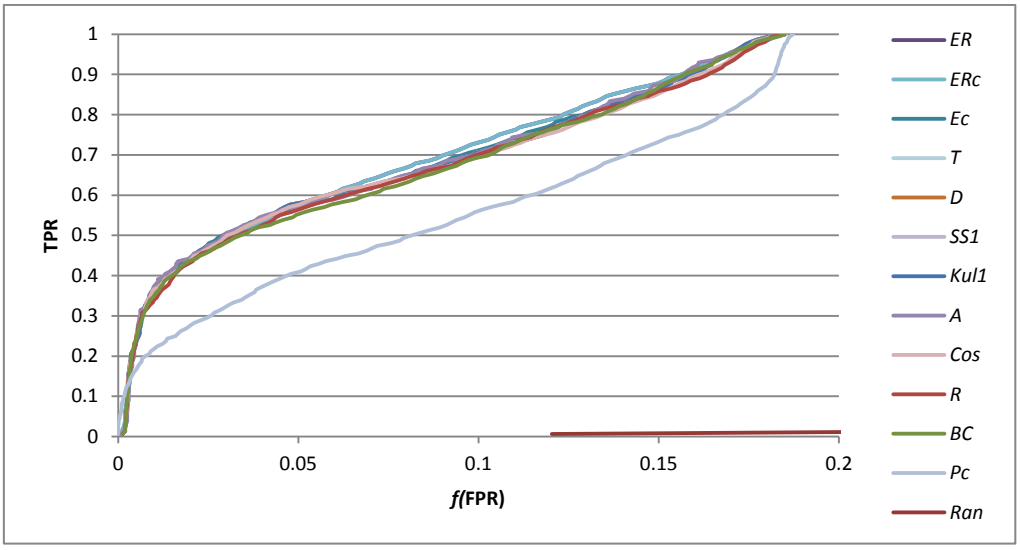


Gráfico 3.1. Secciones de curvas CROC para los mejores modelos de proximidad mostrando el reconocimiento temprano de los activos en el conjunto de datos anti-VIH del NCI.

3.2.2. Segunda etapa: Sistematización de medidas de similitud basadas en el acuerdo relacional y su validación exhaustiva en la búsqueda de similitud de repositorios quimioinformáticos.

Con el objetivo de mostrar la utilidad de las medidas RA en Quimioinformática, estas fueron comparadas con otras medidas de proximidad (PMs: medidas de similitud, disimilitud, distancia) reportadas en el trabajo de Al Kalifa y coautores (32) (p_{11}-p_{112}) y Varin y coautores (188) (p_{21}-p_{29}), para más detalles acerca de las fórmulas correspondientes ver **Ec. 3.5-3.16** y **Tabla 3.10**, correspondientemente.

De forma similar a un trabajo anterior (189), cada PM fue acoplado a un algoritmo de auto entrenamiento que comprende una técnica 10-CV usando el conjunto de compuestos activos. En cada ciclo, nueve de 10 pliegues fueron usados como el grupo de entrenamiento multi consulta mientras que el pliegue restante dejado fuera se combinó con el conjunto inactivo, actuando así

como el grupo de control. El grupo de predicción fue ordenado con el multi clasificador MAX-SIM (190).

Tabla 3.10. Medidas de proximidad de la literatura quimioinformática.

Nombre	Fórmula[a]	Tipo[b]	ID[c]
Manhattan Promedio	(**3.5**)	D	p_{11}
EuclídeaPromedio	(**3.6**)	D	p_{12}
EuclídeaCuadráticaPromedio	(**3.7**)	D	p_{13}
Bray-Curtis I	(**3.8**)	D	p_{14}
Tanimoto	(**3.9**)	A	p_{15}
Dice	(**3.10**)	A	p_{16}
Fossum	(**3.11**)	A	p_{17}
Sokal-Sneath (1)	(**3.12**)	A	p_{18}
Kulczynski (1)	(**3.13**)	A	p_{19}
Coseno/Ochiai	(**3.14**)	A	p_{110}
Simpson	(**3.15**)	A	p_{111}
Pearson	(**3.16**)	C	p_{112}
-	$M_1 = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left\|x_j - y_j\right\|^2}$	D	p_{21}
-	$M_2 = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\frac{\left\|x_j - y_j\right\|^2}{R_j^2}}$	D	p_{22}
MahalanobisPromedio	$M_3 = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\frac{\left\|x_j - y_j\right\|^2}{\sigma_j^2}}$	D	p_{23}
-	$M_5 = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\frac{\left\|x_j - y_j\right\|}{R_j}}$	D	p_{24}
-	$M_6 = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\frac{\left\|x_j - y_j\right\|}{\sigma_j}}$	D	p_{25}
Clark Promedio/ Coeficiente de Divergencia	$M_7 = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\frac{\left(x_j - y_j\right)^2}{\left(\left\|x_j\right| + \left|y_j\right|\right)^2}}$	D	p_{26}
CamberraPromedio	$M_8 = \frac{1}{n}\sum_{j=1}^{n}\frac{\left\|x_j - y_j\right\|}{\left\|x_j\right| + \left|y_j\right|}$	D	p_{27}
SoergelPromedio	$M_9 = \frac{1}{n}\sum_{j=1}^{n}\frac{\left\|x_j - y_j\right\|}{max\{x_j, y_j\}}$	D	p_{28}
Wave-Edges Promedio	$M_{10} = \frac{1}{n}\sum_{j=1}^{n}\left(1 - \frac{min\{x_j, y_j\}}{max\{x_j, y_j\}}\right)$	D	p_{29}

[a]"R" se refiere a "rango" y no a "ranqueo" y es definido como $R = max\{x_j\} - min\{x_j\}$; σ_j es la desviación estándar de la variable j. [b]Clasificación en Asociación (A), Distancia (D), y Correlación (C). [c]Identificador del PM usado para las comparaciones.

La lista ordenada fue escaneada en busca de cada compuesto etiquetado como activo y entonces, los compuestos activos e inactivos que precedieron cada compuesto activo detectado, fueron usados para calcular la razón de positivos verdaderos (*tpr*) y la razón de positivos falsos (*fpr*), respectivamente. Los pares ordenados (*fpr*, *tpr*) y el factor de magnificación δ = 20 sirvió para graficar la curva CROC y calcular la métrica de desempeño AUC[CROC] (152).

3.2.2.1. Desempeño relativo de las medidas de proximidad.

La serie 1 de PMs mostró un desempeño en el rango 82-100% (ver **Tablas 3.11 y 3.12**).

Tabla 3.11. Nombres e identificadores de los modelos de proximidad utilizados en el "reconocimiento temprano" a través del AUC [CROC] calculado en los conjuntos de datos MUV.

Al Kalifa et al. (2009)	**ID**	**Varin et al.(2009)**	**ID**	**Medidas RA**	**ID**
Media Manhattan	p_{11}	-	p_{21}	Identidad corregida	p_{32}
Media Euclídea	p_{12}	-	p_{22}	aditividad no corregida, y corregida	p_{33}
Euclídea Cuadrática Media	p_{13}	Promedio Mahalanobis	p_{23}	proporcionalidad corregida	p_{35}
Bray-Curtis I o Lance-Williams	p_{14}	-	p_{24}	log-proporcionalidad no corregida	p_{37}
Tanimoto	p_{15}	-	p_{25}	log-proporcionalidad corregida	p_{38}
Dice	p_{16}	Clark Promedio/ Coeficiente de Divergencia	p_{26}	log-linealidad no corregida	p_{39}
Fossum	p_{17}	Camberra Promedio	p_{27}	log-linealidad corregida	p_{310}
Sokal-Sneath (1)	p_{18}	Soergel Promedio	p_{28}	monotonicidad no corregida	p_{311}
Kulczynski (1)	p_{19}	Wave-Edges Promedio	p_{29}	monotonicidad corregida	p_{312}
Cosine/Ochiai	p_{110}				
Simpson	p_{111}				
Pearson	p_{112}				

Tabla 3.12. Exactitud de los modelos de proximidad en el "reconocimiento temprano" a través del AUC [CROC] calculado en los conjuntos de datos MUV: Serie 1 de los PMs

AID	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	p_{16}	p_{17}	p_{18}	p_{19}	p_{110}	p_{111}	p_{112}
466	0.8862	0.8845	0.8845	0.9062	0.9032	0.9032	0.8907	0.9032	0.9032	0.8958	0.8921	0.8954
548	0.9457	0.9454	0.9454	0.9672	0.9691	0.9691	0.9283	0.9691	0.9691	0.9671	0.8924	0.9619
600	0.8483	0.8483	0.8483	0.8512	0.8989	0.8989	0.9348	0.8893	0.8989	0.9014	0.9699	0.9017
644	0.9244	0.9300	0.9300	0.9565	0.9596	0.9596	0.9494	0.9596	0.9596	0.9575	0.9526	0.9542
652	0.9207	0.9190	0.9190	0.9608	0.9589	0.9589	0.9276	0.9589	0.9589	0.9622	0.9744	0.9648
689	0.8776	0.8784	0.8784	0.8921	0.9083	0.9083	0.8800	0.9083	0.9083	0.9043	0.9531	0.9005
692	0.9444	0.9444	0.9444	0.9528	0.9541	0.9541	0.8802	0.9540	0.9541	0.9493	0.9272	0.9835
712	0.9035	0.9000	0.9000	0.9370	0.9487	0.9487	0.9407	0.9487	0.9487	0.9519	0.9307	0.9489
713	0.9012	0.8980	0.8980	0.9310	0.9349	0.9349	0.8592	0.9349	0.9349	0.9487	0.8692	0.9557
733	0.9124	0.9112	0.9112	0.9232	0.9254	0.9254	0.8905	0.9254	0.9254	0.9316	0.9354	0.9324
737	0.8881	0.8886	0.8886	0.9067	0.9136	0.9136	0.8865	0.9136	0.9136	0.9053	0.905	0.8871
810	0.9249	0.9194	0.9194	0.9204	0.9163	0.9163	0.9154	0.9163	0.9163	0.9116	0.967	0.9307
832	0.9480	0.9499	0.9499	0.9481	0.9541	0.9541	0.9048	0.9541	0.9541	0.9528	0.934	0.9612
846	0.9145	0.9135	0.9135	0.9546	0.9522	0.9522	0.9536	0.9522	0.9522	0.9596	0.8963	0.9648
852	0.9331	0.9346	0.9346	0.9645	0.9685	0.9685	0.9316	0.9685	0.9685	0.9673	0.9502	0.9598
858	0.8927	0.8914	0.8914	0.9247	0.9323	0.9323	0.914	0.9323	0.9323	0.9339	0.9305	0.9356
859	0.8213	0.8213	0.8213	0.8217	0.9701	0.9701	0.9225	0.9237	0.9701	0.9701	1.0000	0.9700

[a]Los valores promedio AUC[CROC] fueron contrastados frente a la clasificación de aleatoriedad a través de la prueba t-Student de dos colas usando $t_{obs} = \frac{(\bar{X}_{min} - \bar{X}_{max})}{\left(S_{max}/\sqrt{n}\right)}$ con $n - 1 = 9$ g.l., $\bar{X}_{ran} = AUC[CROC]_{ran} = \frac{1}{\delta} - \frac{e^{-\delta}}{1-e^{-\delta}} \approx 0.05$, y $S_{max} = 0.0540$.

La prueba de Friedman y de Iman y Davenport detectaron diferencias significativas entre estos con los conjuntos MUV, y la prueba de Holm posicionó al coeficiente de Pearson como la mejor medida seguido por Tanimoto, Dice, Kulczynski, Coseno, Sokal-Sneath, Bray-Curtis I, y Simpson mostrando en este orden un desempeño similarmente significativo (ver **Tabla 3.13**).

Tabla 3.13. Resultados de la prueba de Holm entre los PMs de la serie 1 sobre los conjuntos de datos MUV.

PM[a]	R_i[b]	Z_s[c]	p_i[d]	α/j[e]
p112	4	-	-	-
p15	4.3235	0.2616	0.793623	0.0500
p16	4.3235	0.2616	0.793623	0.0250
p19	4.3235	0.2616	0.793623	0.0167
p110	4.6765	0.5470	0.584379	0.0125
p18	4.7647	0.6183	0.536347	0.0100
p14	6.2353	1.8075	0.070688	0.0083
p111	6.4706	1.9977	0.045745	0.0071
p_{17}	9.4118	4.3760	1.209E-05*	0.0063
p_{11}	9.5294	4.4711	7.781E-06*	0.0056
p_{13}	9.9412	4.8041	1.555E-06*	0.0050
p_{12}	10.000	4.8517	1.224E-06*	0.0045

[a]Estadístico de Friedman (distribuido según la ji-cuadrado con 11 grados de libertad): 89.7285, valor de p calculado por la prueba de Friedman: 1.884E-14; Estadístico de Iman y Davenport (distribuido según la distribución F de Snedecor con 11 y 176 grados de libertad): 14.7593, valor de p calculado por la prueba de Iman y Daveport: 5.023E-20. PMs en orden decreciente de desempeño con respecto a la mejor medida p_{112}. [b]Ranqueo promedio de la prueba de Friedman. [c]$Z_s = \left[R_j - \left(R_j\right)_{min}\right] / \sqrt{N_{pm}(N_{pm}+1)/6N_{ds}}$ se asume que se distribuye según una normal estandarizada. [d]Probabilidad de dos colas asociada con Z_s; *diferencias significativas. [e]Valores de significación ajustados de acorde a la prueba de Holm.

De forma similar a los resultados previos, la serie 2 de PMs mostraron un desempeño no aleatorio en el rango 82-96% (ver **Tablas 3.11, 3.14 y 3.15**). Sin embargo, en este caso las pruebas por lote no detectaron diferencias significativas entre los PMs.

Tabla 3.14. Exactitud de los modelos de proximidad en el "reconocimiento temprano" a través del AUC [CROC] calculado en los conjuntos de datos MUV: Serie 2 de los PMs

AID	p_{21}	p_{22}	p_{23}	p_{24}	p_{25}	p_{26}	p_{27}	p_{28}	p_{29}
466	0.8845	0.8864	0.8860	0.8869	0.8867	0.8942	0.8978	0.8963	0.8963
548	0.9454	0.9495	0.9380	0.9489	0.9405	0.9408	0.9377	0.9034	0.9034
600	0.8483	0.8488	0.8475	0.8484	0.8474	0.8527	0.8514	0.8933	0.8933
644	0.9300	0.9297	0.9248	0.9233	0.9209	0.9278	0.9280	0.8765	0.8765
652	0.9190	0.9462	0.9344	0.9432	0.9338	0.9579	0.9485	0.8427	0.8427
689	0.8784	0.8796	0.8818	0.8803	0.8824	0.8769	0.8763	0.8879	0.8879
692	0.9444	0.9444	0.9444	0.9444	0.9444	0.9496	0.9496	0.9218	0.9218
712	0.9000	0.8910	0.8921	0.8895	0.8937	0.9376	0.9368	0.9342	0.9342
713	0.8980	0.8607	0.8611	0.8615	0.8623	0.9398	0.9291	0.9152	0.9152
733	0.9111	0.9179	0.9193	0.9172	0.9176	0.9064	0.9036	0.8624	0.8624
737	0.8886	0.8885	0.8887	0.8883	0.8883	0.8971	0.8963	0.9617	0.9617
810	0.9194	0.9439	0.9425	0.9392	0.9396	0.9179	0.9244	0.8582	0.8582
832	0.9499	0.9564	0.9581	0.9546	0.9566	0.922	0.9221	0.8717	0.8717
846	0.9135	0.9200	0.9200	0.9204	0.9218	0.9414	0.9467	0.9273	0.9273
852	0.9346	0.9178	0.9165	0.9177	0.9157	0.9577	0.9457	0.949	0.9490
858	0.8914	0.8861	0.8850	0.8893	0.8878	0.9114	0.9108	0.8624	0.8624
859	0.8213	0.8213	0.8213	0.8213	0.8213	0.8216	0.8216	0.8216	0.8216

[a]Los valores promedio AUC[CROC] fueron contrastados frente a la clasificación de aleatoriedad a través de la prueba t-Student de dos colas usando $t_{obs} = \frac{(\bar{X}_{min} - \bar{X}_{max})}{\left(S_{max}/\sqrt{n}\right)}$ con $n - 1 = 9$ g.l., $\bar{X}_{ran} = AUC[CROC]_{ran} = \frac{1}{\delta} - \frac{e^{-\delta}}{1-e^{-\delta}} \approx 0.05$, y $S_{max} = 0.0560$.

Tabla 3.15. Resultados de la prueba de Holm entre los PMs de la serie 2 sobre los conjuntos de datos MUV.

PM[a]	R_j[b]
p_{26}	3.3529
p_{27}	3.7059
p_{22}	5.0000
p_{28}	5.3235
p_{29}	5.3235
p_{24}	5.4706
p_{21}	5.5882
p_{23}	5.5882
p_{25}	5.6471

[a]Estadístico de Friedman (distribuido según la ji-cuadrado con 8 grados de libertad): 13.4392, valor de p calculado por la prueba de Friedman: 0.0976; Estadístico de Iman y Davenport (distribuido según la distribución F de Snedecor con 8 y 128 grados de libertad): 1.7544, valor de p calculado por la prueba de Iman y Davenport: 0.0920. [b]Ranqueo promedio de la prueba de Friedman.

Las medidas RA excedieron las expectativas ya que mostraron un desempeño no aleatorio en el rango 86-100% (ver **Tablas 3.11, 3.16 y 3.17**). En este caso, las pruebas por lotes detectaron diferencias entre las mismas.

Tabla 3.16. Exactitud de los modelos de proximidad en el "reconocimiento temprano" a través del AUC [CROC] calculado en los conjuntos de datos MUV: Serie 3 de los PMs

AID	p_{32}	p_{33}	p_{35}	p_{37}	p_{38}	p_{39}	p_{310}	p_{311}	p_{312}
466	0.9011	0.9002	0.9290	0.8958	0.8975	0.8828	0.9362	0.9300	0.9300
548	0.9682	0.9682	0.9728	0.9430	0.9478	0.8654	0.9517	0.9580	0.9580
600	0.8987	0.8986	0.9676	0.8824	0.8789	0.8619	0.9328	0.9885	0.9885
644	0.9618	0.9621	0.9588	0.9446	0.9470	0.9205	0.9173	0.9396	0.9396
652	0.9620	0.9621	0.9561	0.9161	0.9205	0.9033	0.8896	0.9410	0.9410
689	0.9111	0.9116	0.9478	0.8674	0.8684	0.8763	0.9255	0.8944	0.8944
692	0.9497	0.9497	0.9112	0.9412	0.9416	0.9083	0.8877	1.0000	1.0000
712	0.9479	0.9462	0.9562	0.8943	0.9124	0.8999	0.8944	0.9469	0.9469
713	0.9405	0.9388	0.9604	0.9270	0.9379	0.8714	0.9120	0.9169	0.9169
733	0.9294	0.9297	0.9446	0.8926	0.8946	0.9091	0.9275	0.9661	0.9661
737	0.9169	0.9169	0.9524	0.8842	0.8981	0.8828	0.8970	0.9761	0.9761
810	0.9406	0.9348	0.9463	0.9233	0.9307	0.9267	0.9193	0.9105	0.9191
832	0.9565	0.9532	0.9422	0.9415	0.9452	0.9048	0.9100	0.9512	0.9512
846	0.9534	0.9536	0.9657	0.9447	0.9518	0.8945	0.9048	0.9545	0.9545
852	0.9660	0.9645	0.9745	0.9496	0.9618	0.9119	0.9315	0.9711	0.9711
858	0.9349	0.9350	0.9646	0.8948	0.9108	0.8581	0.9135	0.9644	0.9644
859	0.9700	0.9700	0.9857	0.9600	0.9600	0.9133	0.9632	0.9879	0.9879

[a]Los valores promedio AUC[CROC] fueron contrastados frente a la clasificación de aleatoriedad a través de la prueba t-Student de dos colas usando $t_{obs} = \frac{(\bar{X}_{min} - \bar{X}_{max})}{\left(S_{max}/\sqrt{n}\right)}$ con $n - 1 = 9$ g.l., $\bar{X}_{ran} = AUC[CROC]_{ran} = \frac{1}{\delta} - \frac{e^{-\delta}}{1-e^{-\delta}} \approx 0.05$, y $S_{max} = 0.0560$.

Tabla 3.17. Resultados de la prueba de Holm entre los PMs de la serie 3 sobre los conjuntos de datos MUV.

PM[a]	R_j[b]	Z_s[c]	$p_{(Npm-j)}$[d]	$\alpha/(N_{pm}-j)$[e]
p35	2.5294	-	-	-
p32	3.4706	1.002	0.3163635	0.0500
p312	3.5294	1.0646	0.2870654	0.0250
p311	3.5882	1.1272	0.2596564	0.0167
p33	3.6471	1.1898	0.2341147	0.0125
p38	6.1471	3.8513	0.0001175*	0.0100
p310	6.5882	4.3209	1.554E-05*	0.0083
p37	7.2647	5.0411	4.629E-07*	0.0071
p39	8.2353	6.0744	1.245E-09*	0.0063

[a]Estadístico de Friedman (distribuido según la ji-cuadrado con 8 grados de libertad): 76.7569, valor de p calculado por la prueba de Friedman: 2.193E-13; Estadístico de Iman y Davenport (distribuido según la distribución F de Snedecor con 8 y 128 grados de libertad): 20.7300, valor de p calculado por la prueba de Iman y Davenport: 7.129E-20. PMs en orden decreciente de desempeño con respecto a la mejor medida p_{35}. Los PMs que fueron retenidos en las comparaciones múltiples por pares están señalados en negrita-cursiva. [b]Ranqueo promedio de la prueba de Friedman. [c]$Z_s = \left[R_j - \left(R_j\right)_{min}\right] / \sqrt{N_{pm}(N_{pm}+1)/6N_{ds}}$ se asume que se distribuye según una normal estandarizada. [d]Probabilidad de dos colas asociada con Z_s; *diferencias significativas. [e]Valores de significación ajustados de acorde a la prueba de Holm.

El procedimiento de Holm posicionó el coeficiente de proporcionalidad corregido como la mejor medida, seguido por la identidad y monotonicidad corregidas y la monotonicidad no corregida. Es notorio que todas las medidas log-transformadas fueron dejadas afuera, lo cual indica que las transformaciones basadas en logaritmo no son necesarias (en tendencia mediana) para modelar las relaciones de similitud molecular latentes en los conjuntos de datos MUV.

Las mejores medidas resultantes de las comparaciones en la serie 1 de PMs fueron comparadas con toda la serie 2 (ver **Tabla 3.18**). Como resultado, las pruebas por lotes detectaron diferencias significativas entre los conjuntos de PMs combinados, y la prueba de Holm retuvo todas las medidas de la serie 1 las cuales superaron radicalmente todas las medidas de la serie 2. En cuanto a las medidas de la serie 1, los resultados concordaron completamente con los obtenidos en el primer experimento de comparación, esto es, el coeficiente de Pearson fue el mejor ranqueado,

seguido por las medidas de similitud Tanimoto, Dice, Kulczynski, Coseno, Sokal-Sneath, Bray-Curtis I, y Simpson.

Tabla 3.18. Resultados de la prueba de Holm en la comparación de las mejores medidas de la serie 1 vs. todas las medidas de la serie 2en los conjuntos de datos MUV.

PM[a]	R_j[b]	Z_s[c]	$p_{(Npm-j)}$[d]	$\alpha/(N_{pm}-j)$[e]
p112	4.7059	-	-	-
p15	4.8529	0.0849	0.9323374	0.0500
p16	4.8529	0.0849	0.9323374	0.0250
p19	4.8529	0.0849	0.9323374	0.0167
p18	5.4118	0.4075	0.6836104	0.0125
p110	5.4412	0.4245	0.671185	0.0100
p14	6.9412	1.2905	0.1968606	0.0083
p111	8.1176	1.9698	0.0488632	0.0071
p26	10.118	3.1245	0.0017812*	0.0063
p27	10.706	3.4641	0.0005320*	0.0056
p22	11.941	4.1773	2.950E-05*	0.0050
p28	12.324	4.3981	1.092E-05*	0.0045
p29	12.324	4.3981	1.092E-05*	0.0042
p24	12.412	4.449	8.627E-06*	0.0038
p23	12.529	4.5169	6.275E-06*	0.0036
p25	12.588	4.5509	5.342E-06*	0.0033
p21	12.882	4.7207	2.350E-06*	0.0031

[a]Estadístico de Friedman (distribuido según la ji-cuadrado con 8 grados de libertad): 76.7569, valor de p calculado por la prueba de Friedman: 2.193E-13; Estadístico de Iman y Davenport (distribuido según la distribución F de Snedecor con 8 y 128 grados de libertad): 20.7300, valor de p calculado por la prueba de Iman y Davenport: 7.129E-20. PMs en orden decreciente de desempeño con respecto a la mejor medida p_{35}. Los PMs que fueron retenidos en las comparaciones múltiples por pares están señalados en negrita y cursiva. [b]Ranqueo promedio de la prueba de Friedman. [c]$Z_s = \left[R_j - \left(R_j\right)_{min}\right] / \sqrt{N_{pm}(N_{pm}+1)/6N_{ds}}$ se asume que se distribuye según una normal estandarizada. [d]Probabilidad de dos colas asociada con Z_s; *diferencias significativas. [e]Valores de significación ajustados de acorde a la prueba de Holm.

En una última ronda de comparaciones, enfrentamos a las mejores medidas del experimento anterior (los mejores PMs de la serie 1 y 2) con las medidas de la serie 3, que son a su vez, las medidas RA propuestas en este trabajo. Los resultados de las pruebas por lote indican que las diferencias significativas se deben a los "peores en desempeño" Simpson y Bray-Curtis I. La prueba *post hoc* de Holm también muestra que la posición más alta fue alcanzada por la medida de proporcionalidad corregida, seguida por las medidas monotonicidad e identidad corregidas, Pearson (linealidad), monotonicidad no corregida, Tanimoto, Dice (identidad no corregida), Kulczynski, Coseno (proporcionalidad no corregida), y Sokal-Sneath (ver **Tabla 3.19**).

Tabla 3.19. Resultados de la prueba de Holm en la comparación de las mejores medidas de las series 1 y 2 vs. las mejores medidas de la serie 3 en los conjuntos de datos MUV.

PM[a]	R_i[b]	Z_s[c]	P_i[d]	α/j[e]	Li[f]
p_{35}**	4.1176	-	-	-	-
p_{312}**	6.2353	1.5853	0.112893	0.0500	0.050
p_{32}**	6.3824	1.6954	0.089997	0.0250	0.047
p_{112}**	6.4118	1.7174	0.085900	0.0167	0.047
p_{311}**	6.5882	1.8495	0.064379	0.0125	0.047
p_{33}	6.8529	2.0477	0.040588	0.0100	0.047
p_{15}	7.0294	2.1798	0.029271	0.0083	0.047
p_{16}	7.0294	2.1798	0.029271	0.0071	0.047
p_{19}	7.0294	2.1798	0.029271	0.0063	0.047
p_{110}	7.2647	2.356	0.018475	0.0056	0.047
p_{18}	7.7059	2.6862	0.007226	0.0050	0.047
p_{111}	8.7647	3.4789	0.0005035*	0.0045	0.047
p_{14}	9.5882	4.0954	4.214E-05*	0.0042	0.047

[a]Estadístico de Friedman (distribuido según la ji-cuadrado con 12 grados de libertad): 22.6367, valor de p calculado por la prueba de Friedman: 0.0310; Estadístico de Iman y Davenport (distribuido según la distribución F de Snedecor con 12 y 192 grados de libertad): 1.9970, valor de p calculado por la prueba de Iman y Daveport: 0.0264. PMs en orden decreciente de desempeño con respecto a la mejor medida p_{35}.[b]Ranqueo promedio de la prueba de Friedman. [c]$Z_s = \left[R_j - \left(R_j\right)_{min}\right] / \sqrt{N_{pm}(N_{pm}+1)/6N_{ds}}$ se asume que se distribuye según una normal estandarizada. [d]Probabilidad de dos colas asociada con Z_s; *diferencias significativas. [e]Valores de significación ajustados de acorde a la prueba de Holm. [f]Valores de significación ajustados de acorde a la prueba en dos etapas de Li.

La lista ordenada de la tabla anterior muestra que las medidas RA corregidas por aleatoriedad ranquearon mejor que las medidas no corregidas. También es notorio que 8 de 11 PMs son medidas RA, y que 5 de 11 PMs son adicionalmente novedosos en Quimioinformática para la búsqueda de similitud. Cuando aplicamos la prueba más potente de Li, en vez de la prueba de Holm, se observa que todas las medidas proceden de la teoría del RA. En la **Tabla 3.19** se muestran estos PMs marcados con ** (p_{35}, p_{312}, p_{32}, p_{112} y p_{311}). Note que cuatro de estos PMs son nuevos en Quimioinformática para el problema de la "recuperación temprana".

En este punto se considera justo resaltar que, en ambas etapas de introducción y validación (preliminar y exhaustiva) de las medidas de acuerdo relacional, algunas de las mismas se situaron entre los 10 mejores modelos ("top-10") para la búsqueda de similitud de repositorios quimioinformáticos de tamaño pequeño (66 casos) a grande (43850 casos), que comprenden resultados de cribados experimentales correspondientes a diversas actividades farmacológicas. Las mejores medidas RA para ser usadas en tareas de búsqueda de similitud resultaron ser: el coeficiente de identidad corregido, e^c ó e'; el coeficiente de aditividad, *a*; el coeficiente de

proporcionalidad corregido, p^c o p'; el coeficiente de linealidad o de Pearson, r y los coeficientes basados en la escala ordinal o de monotonicidad, no corregido y corregido estadísticamente, e^R ó m y e^{Rc} ó m', respectivamente. Estas medidas propuestas pudieran resultar útiles en aplicaciones quimioinformáticas concretas como la búsqueda de nuevas entidades químicas con aplicaciones médicas en enfermedades de alta sensibilidad social como el SIDA y el cáncer.

3.2.2.2. Comportamiento de las medidas RA en un escenario realista.

En esta etapa, se condujo un experimento adicional sobre la calidad de las medidas RA (exceptuando las log transformadas) usando el repositorio NCI y el mismo protocolo experimental. Una vez más, los resultados mostraron que estas medidas son útiles para la búsqueda de similitud ya que rindieron valores de AUC[CROC] entre 0.91-0.94, indicando que existe al menos una probabilidad del 91% que un compuesto anti-VIH ranqueado por una de estas medidas sea encontrado antes que un compuesto que provendría de una distribución exponencial con parámetro $\delta = 20$ (202). También, para disponer de una evaluación visual complementaria, brindamos las partes más interesantes de las curvas CROC "promedios" en el **Gráfico 3.2**, mostrando el hecho que ahora las medidas p_{35} y p_{312} se desempeñan peor que las otras medidas RA para este problema en cuestión. Sin embargo, incluso estas "peores" medidas superan con amplio margen al clasificador aleatorio, recuperando todos los fármacos cuando cerca del 20% de la lista ha sido escaneada.

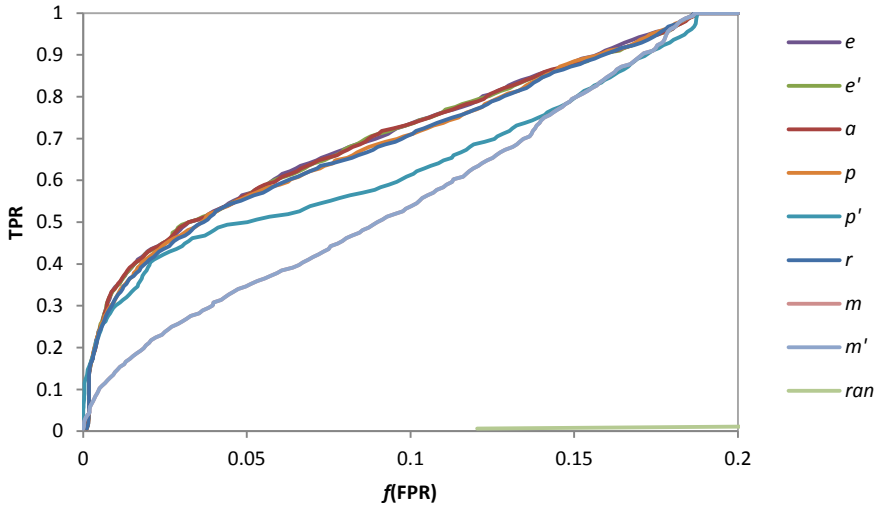


Gráfico 3.2. Curvas CROC de las medidas RA en la recuperación temprana del conjunto de datos anti-VIH del NCI.

3.2.2.3. Interpretación de las medidas basadas en RA.

A través de la inspección de los resultados anteriores resulta claro que los PMs RA se desempeñaron considerablemente bien en tendencia mediana. A partir de la Teoría de la Significatividad, sin embargo, resulta más interesante encontrar el verdadero desacuerdo sin significado o función matemática subyacente que domina las relaciones de similitud entre los biocompuestos en cada repositorio MUV. Considerando el mejor PM en cuanto a su desempeño para cada repositorio químico (ver **Tabla 3.16**), se puede observar que las biomoléculas del conjunto aid644 comparten una relación de "similitud aditiva" mientras que los compuestos de los conjuntos de datos aid548, aid689, aid712, aid713, aid810, aid846, aid852 y aid858 son descritos más adecuadamente por una relación de "similitud proporcional". Por otra parte, las moléculas de los conjuntos de datos aid652 y aid832 pueden ser modelados por una relación de "similitud lineal" más compleja, y finalmente los compuestos de los repositorios aid466, aid600, aid692, aid733, aid737 y aid859 están relacionados a través de una relación de "similitud monótona" de mayor grado polinómico.

3.3. Análisis teórico de la generalización de las medidas de similitud por pares del acuerdo relacional al caso de varios objetos químicos y derivación de nuevas medidas *k*-arias para su uso en la Quimioinformática.

En este epígrafe se muestra que la fórmula general de acuerdo relacional multivariada no corregidas (corregidas) es la media ponderada de la fórmula general de acuerdo relacional bivariada no corregida (corregida). Por último se presentan algunas aplicaciones de los coeficientes multivariados en Quimioinformática.

3.3.1. Generalización de las fórmulas de Zegers-ten Berge y su relación con la Teoría de la Generalizabilidad.

La generalización de (**3.18**) y (**3.20**) es posible considerando $k \geq 2$ mediciones en n objetos, resultando en un arreglo matricial k x n. Siguiendo razonamientos análogos a Zegers y ten Berge (176-177), Fagot (203) desarrolló los coeficientes de acuerdo relacional multivariados no corregidos y corregidos como:

$$G_k = \frac{2\sum_{i<j}^{k}\sum_{t=1}^{n} u_{it}u_{jt}}{(k-1)\sum_{i=1}^{k}\sum_{t=1}^{n} u_{it}^2} \quad (\mathbf{3.30})$$

$$G_k' = \frac{2\left[\sum_{i<j}^{k}\sum_{t=1}^{n} u_{it}u_{jt} - n\sum_{i<j}^{k}\bar{U}_i\bar{U}_j\right]}{(k-1)\sum_{i=1}^{k}\sum_{t=1}^{n} u_{it}^2 - 2n\sum_{i<j}^{k}\bar{U}_i\bar{U}_j} \quad (\mathbf{3.31})$$

Donde, $\sum_{i<j}^{k} V_{ij} = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} V_{ij}$, y $\bar{U}_i = (\sum_{t=1}^{n} u_{it})/n$

Substituyendo las transformaciones uniformadoras correspondientes a las escalas absoluta, diferencia, razón e intervalo en (18), Fagot (203) obtuvo los coeficientes corregidos de identidad (**3.32**), aditividad (**3.33**), proporcionalidad (**3.34**) y linealidad (**3.35**) para mediciones múltiples.

$$I_k' = \frac{2\sum_{i<j}^{k} s_{ij}}{(k-1)\sum_{i=1}^{k} s_i^2 + \sum_{i<j}^{k}\left(\bar{X}_i - \bar{X}_j\right)^2} \quad (\mathbf{3.32})$$

$$A_k = \frac{2\sum_{i<j}^{k} s_{ij}}{(k-1)\sum_{i=1}^{k} s_i^2} \quad (\mathbf{3.33})$$

$$P_k^{'} = \frac{\sum_{i<j}^{k}\left[\frac{\sum_{t=1}^{n} x_{it}x_{jt} - n\bar{X}_i\bar{X}_j}{\left(\sum_{t=1}^{n} x_{it}^2 \sum_{t=1}^{n} x_{jt}^2\right)^{1/2}}\right]}{\frac{k}{2}(k-1) - n\sum_{i<j}^{k}\left[\frac{\bar{X}_i\bar{X}_j}{\left(\sum_{t=1}^{n} x_{it}^2 \sum_{t=1}^{n} x_{jt}^2\right)^{1/2}}\right]} \quad \textbf{(3.34)}$$

$$L_k = \bar{r}_{ij} \quad \textbf{(3.35)}$$

En un trabajo posterior, Fagot (204) derivó el coeficiente multivariado corregido para la escala ordinal *O*' (**3.36**). También, demostró que este coeficiente es equivalente al coeficiente de concordancia de Kendall (*W*) corregido por aleatoriedad.

$$O_k' = \frac{2\left[\sum_{i<j}^{k}\sum_{t=1}^{n} r_{it}r_{jt} - \frac{nk}{8}(k-1)(n+1)^2\right]}{(k-1)\left[\sum_{i=1}^{k}\sum_{t=1}^{n} r_{it}^2 - \frac{nk}{4}(k-1)(n+1)^2\right]} \quad \textbf{(3.36)}$$

3.3.2. Relación funcional entre los coeficientes de Acuerdo Relacional multivariados y bivariados no corregidos por aleatoriedad.

Teorema 1: La fórmula general de acuerdo relacional multivariadas no corregida (**3.30**) es la media ponderada de la fórmula general de acuerdo relacional bivariada no corregida (**3.18**).

Prueba: Reemplazando $2\sum_{t=1}^{n} u_{it}\, u_{jt}$ en (**3.30**) por el numerador de (**3.18**), se obtiene la relación siguiente:

$$G_k = \sum_{i<j}^{k} \frac{\sum_{t=1}^{n} u_{it}^2 + \sum_{t=1}^{n} u_{jt}^2}{(k-1)\sum_{i=1}^{k}\sum_{t=1}^{n} u_{it}^2} g_{ij} \quad \textbf{(3.37)}$$

Ahora, se tiene que probar que los pesos en (**3.37**) suman la unidad, esto es:

$$\sum_{i<j}^{k} \frac{\sum_{t=1}^{n} u_{it}^2 + \sum_{t=1}^{n} u_{jt}^2}{(k-1)\sum_{i=1}^{k}\sum_{t=1}^{n} u_{it}^2} \equiv 1 \quad \textbf{(3.38)}$$

Se comienza notando que la suma doble $\sum_{i<j}^{k}\left(\sum_{t=1}^{n} u_{it}^2 + \sum_{t=1}^{n} u_{jt}^2\right)$ producirá *k-1* términos $u_{1.}^2$, *k-1* términos $u_{2.}^2$,… *k-1* términos $u_{k-1.}^2$,… y *k-1* términos $u_{k.}^2$, de modo que:

$$\sum_{i<j}^{k}\left(\sum_{t=1}^{n} u_{it}^2 + \sum_{t=1}^{n} u_{jt}^2\right) = \sum_{i=1}^{k}(k-1)\sum_{t=1}^{n} u_{it}^2 \quad \textbf{(3.39)}$$

Reemplazando el denominador del miembro izquierdo de (**3.38**) por su forma equivalente dada por (**3.39**) y haciendo algún trabajo algebraico se obtiene la identidad buscada ■

Finalmente, se obtiene una expresión algebraicamente más dócil para (**3.37**) dada por:

$$G_k = \sum_{i<j}^{k} \frac{\sum_{t=1}^{n} u_{it}^2 + \sum_{t=1}^{n} u_{jt}^2}{\sum_{i<j}^{k} \left(\sum_{t=1}^{n} u_{it}^2 + \sum_{t=1}^{n} u_{jt}^2 \right)} g_{ij} \quad \textbf{(3.40)}$$

Corolario 1.1: El coeficiente de identidad multivariado no corregido es la media ponderada del coeficiente de identidad bivariado no corregido (**3.10**).

Prueba: Substituyendo la transformación uniformadora para la escala absoluta $U = X$ en la relación (**3.40**), se obtiene:

$$I_k = \sum_{i<j}^{k} \frac{\sum_{t=1}^{n} x_{it}^2 + \sum_{t=1}^{n} x_{jt}^2}{\sum_{i<j}^{k} \left(\sum_{t=1}^{n} x_{it}^2 + \sum_{t=1}^{n} x_{jt}^2 \right)} e_{ij}$$

La cual es equivalente a:

$$I_k = \sum_{i<j}^{k} \frac{T_i^2 + T_j^2}{\sum_{i<j}^{k} \left(T_i^2 + T_j^2 \right)} e_{ij}$$

Definiendo, $\overline{T_{ij}^2} = \left(T_i^2 + T_j^2 \right)/2$, se obtiene finalmente:

$$I_k = \sum_{i<j}^{k} \frac{\overline{T_{ij}^2}}{\sum_{i<j}^{k} \overline{T_{ij}^2}} e_{ij} \blacksquare$$

Corolario 1.2: El coeficiente de aditividad multivariado no corregido (**3.33**) es la media ponderada del coeficiente de aditividad bivariado no corregido (**3.22**).

Prueba: Substituyendo la transformación uniformadora para la escala de diferencias $U = X - \bar{X}$ en (**3.40**), se obtiene:

$$A_k = \sum_{i<j}^{k} \frac{\sum_{t=1}^{n} (x_{it} - \bar{X}_i)^2 + \sum_{t=1}^{n} \left(x_{jt} - \bar{X}_j \right)^2}{\sum_{i<j}^{k} \left[\sum_{t=1}^{n} (x_{it} - \bar{X}_i)^2 + \sum_{t=1}^{n} \left(x_{jt} - \bar{X}_j \right)^2 \right]} a_{ij}$$

Después de algún trabajo algebraico se obtiene:

$$A_k = \sum_{i<j}^{k} \frac{S_i^2 + S_j^2}{\sum_{i<j}^{k} \left(S_i^2 + S_j^2 \right)} a_{ij}$$

Definiendo $\overline{S_{ij}^2} = \left(S_i^2 + S_j^2 \right)/2$, se obtiene finalmente:

$$A_k = \sum_{i<j}^{k} \frac{\overline{S_{ij}^2}}{\sum_{i<i'}^{k} \overline{S_{ij}^2}} a_{ij} \blacksquare$$

Corolario 1.3: El coeficiente de proporcionalidad multivariado no corregido es el promedio del coeficiente de proporcionalidad bivariado no corregido (**3.14**).

Prueba: Substituyendo la transformación uniformadora para la escala de razón $U = X/T_X$ en (**3.40**), se obtiene:

$$P_k = \sum_{i<j}^{k} \frac{\sum_{t=1}^{n} \left(\frac{x_{it}}{T_i}\right)^2 + \sum_{t=1}^{n} \left(\frac{x_{jt}}{T_j}\right)^2}{\sum_{i<j}^{k} \left[\sum_{t=1}^{n} \left(\frac{x_{it}}{T_i}\right)^2 + \sum_{t=1}^{n} \left(\frac{x_{jt}}{T_j}\right)^2\right]} p_{ij}$$

Luego de algún trabajo algebraico, se obtiene finalmente:

$P_k = \bar{p}_{ij} \blacksquare$

Corolario 1.4: El coeficiente de linealidad multivariado no corregido es el promedio del coeficiente de correlación de Pearson (**3.16**).

Prueba: Substituyendo la transformación uniformadora para la escala de intervalo $U = (X - \bar{X})/S_X$ en (**3.40**), se obtiene:

$$L_k = \sum_{i<j}^{k} \frac{\sum_{t=1}^{n} \left(\frac{x_{it}-\bar{X}_i}{S_i}\right)^2 + \sum_{t=1}^{n} \left(\frac{x_{jt}-\bar{X}_j}{S_j}\right)^2}{\sum_{i<j}^{k} \left[\sum_{t=1}^{n} \left(\frac{x_{it}-\bar{X}_i}{S_i}\right)^2 + \sum_{t=1}^{n} \left(\frac{x_{jt}-\bar{X}_j}{S_j}\right)^2\right]} r_{ij}$$

Después de alguna manipulación algebraica, se obtiene finalmente:

$L_k = \bar{r}_{ij} \blacksquare$

La relación anterior entre los coeficientes de linealidad multivariado y bivariado no corregidos también fue obtenida por Fagot (203) usando una vía alternativa.

Corolario 1.5: El coeficiente de monotonicidad multivariado no corregido es la media ponderada del coeficiente de monotonicidad bivariado no corregido (**3.28**).

Prueba: Substituyendo la transformación uniformadora para la escala ordinal $U = Rank(X)$ [transformada RT-2 de Conover e Iman (186)] en (**3.40**), se obtiene:

$$M_k = \sum_{i<j}^{k} \frac{\sum_{t=1}^{n} r_{it}^2 + \sum_{t=1}^{n} r_{jt}^2}{\sum_{i<j}^{k}\left(\sum_{t=1}^{n} r_{it}^2 + \sum_{t=1}^{n} r_{jt}^2\right)} m_{ij}$$

Esta expresión es equivalente a:

$$M_k = \sum_{i<j}^{k} \frac{TR_i^2 + TR_j^2}{\sum_{i<j}^{k}\left(TR_i^2 + TR_j^2\right)} m_{ij}$$

Definiendo$\overline{TR_{ij}^2} = \left(TR_i^2 + TR_j^2\right)/2$, se obtiene finalmente:

$$M_k = \sum_{i<j}^{k} \frac{\overline{TR_{ij}^2}}{\sum_{i<j}^{k} \overline{TR_{ij}^2}} m_{ij} \blacksquare$$

La relación anterior se reduce a $M_k = \bar{m}_{ij}$ cuando no hay ataduras presentes.

3.3.3. Relación funcional entre los coeficientes de Acuerdo Relacional multivariados y bivariados corregidos por aleatoriedad.

Teorema 2: La fórmula general de acuerdo relacional multivariada corregida por aleatoriedad (**3.31**) es la media ponderada de la fórmula general de acuerdo relacional bivariada corregida por aleatoriedad (**3.20**).

Prueba: Teniendo en cuenta algunos elementos de la prueba para el **Teorema 1**, esto es:

$$\sum_{i<j}^{k}\left(\sum_{t=1}^{n} u_{it}^2 + \sum_{t=1}^{n} u_{jt}^2\right) = (k-1)\sum_{i=1}^{k}\sum_{t=1}^{n} u_{it}^2$$

la expresión (**3.31**) puede ser reescrita como:

$$G_k^{'} = \frac{\sum_{i<j}^{k} 2\left(\sum_{t=1}^{n} u_{it}u_{jt} - n^{-1}\sum_{t=1}^{n} u_{it}\sum_{t=1}^{n} u_{jt}\right)}{\sum_{i<j}^{k}\left(\sum_{t=1}^{n} u_{it}^2 + \sum_{t=1}^{n} u_{jt}^2 - 2n^{-1}\sum_{t=1}^{n} u_{it}\sum_{t=1}^{n} u_{jt}\right)} \qquad (\mathbf{3.41})$$

Reemplazando $2\left(\sum_{t=1}^{n} u_{it}\, u_{jt} - n^{-1}\sum_{t=1}^{n} u_{it}\sum_{t=1}^{n} u_{jt}\right)$ en (**3.41**) por el numerador de (**3.20**), se obtiene la relación siguiente entre los coeficientes generalizados multi y bivariados corregidos por aleatoriedad:

$$G_k^{'} = \sum_{i<j}^{k} \frac{\sum_{t=1}^{n} u_{it}^2 + \sum_{t=1}^{n} u_{jt}^2 - 2n^{-1}\sum_{t=1}^{n} u_{it}\sum_{t=1}^{n} u_{jt}}{\sum_{i<j}^{k}\left(\sum_{t=1}^{n} u_{it}^2 + \sum_{t=1}^{n} u_{jt}^2 - 2n^{-1}\sum_{t=1}^{n} u_{it}\sum_{t=1}^{n} u_{jt}\right)} g_{ij}^{'} \qquad (\mathbf{3.42})$$

Similarmente al caso para la prueba del **Teorema 1**, se debe probar ahora que:

$$\sum_{i<j}^{k} \frac{\sum_{t=1}^{n} u_{it}^2 + \sum_{t=1}^{n} u_{jt}^2 - 2n^{-1}\sum_{t=1}^{n} u_{it}\sum_{t=1}^{n} u_{jt}}{\sum_{i<j}^{k}\left(\sum_{t=1}^{n} u_{it}^2 + \sum_{t=1}^{n} u_{jt}^2 - 2n^{-1}\sum_{t=1}^{n} u_{it}\sum_{t=1}^{n} u_{jt}\right)} \equiv 1$$

Esta identidad puede ser verificada fácilmente notando que el denominador del miembro izquierdo no se afecta por la suma doble que encabeza la expresión completa ■

Corolario 2.1: El coeficiente de identidad multivariado corregido (**3.32**) es la media ponderada del coeficiente de identidad bivariado corregido (**3.21**).

Prueba: Substituyendo la transformación uniformadora para la escala absoluta $U = X$ en (**3.42**), se obtiene:

$$I_k^{'} = \sum_{i<j}^{k} \frac{\sum_{t=1}^{n} x_{it}^2 + \sum_{t=1}^{n} x_{jt}^2 - 2n^{-1}\sum_{t=1}^{n} x_{it}\sum_{t=1}^{n} x_{jt}}{\sum_{i<j}^{k}\left(\sum_{t=1}^{n} x_{it}^2 + \sum_{t=1}^{n} x_{jt}^2 - 2n^{-1}\sum_{t=1}^{n} x_{it}\sum_{t=1}^{n} x_{jt}\right)} e_{ij}^{'}$$

luego de cierta cantidad de trabajo algebraico, se tiene finalmente:

$$I_k^{'} = \sum_{i<j}^{k} \frac{S_i^2 + S_j^2 + \left(\bar{X}_i - \bar{X}_j\right)^2}{\sum_{i<j}^{k}\left[S_i^2 + S_j^2 + \left(\bar{X}_i - \bar{X}_j\right)^2\right]} e_{ij}^{'} \blacksquare$$

Haciendo $\xi_{ij} = S_i^2 + S_j^2 + \left(\bar{X}_i - \bar{X}_j\right)^2$, encontramos que la relación anterior es idéntica a la obtenida por Barnhart y coautores (206).

Corolario 2.2: El coeficiente de aditividad multivariado corregido (**3.33**) es la media ponderada del coeficiente de aditividad bivariado corregido (**3.22**).

Prueba: Substituyendo la transformación uniformadora para la escala de diferencias $U = X - \bar{X}$ en (**3.42**), se obtiene:

$$A_k^{'} = \sum_{i<j}^{k} \frac{\sum_{t=1}^{n}(x_{it} - \bar{X}_i)^2 + \sum_{t=1}^{n}\left(x_{jt} - \bar{X}_j\right)^2 - 2n^{-1}\sum_{t=1}^{n}(x_{it} - \bar{X}_i)\sum_{t=1}^{n}\left(x_{jt} - \bar{X}_j\right)}{\sum_{i<j}^{k}\left[\sum_{t=1}^{n}(x_{it} - \bar{X}_i)^2 + \sum_{t=1}^{n}\left(x_{jt} - \bar{X}_j\right)^2 - 2n^{-1}\sum_{t=1}^{n}(x_{it} - \bar{X}_i)\sum_{t=1}^{n}\left(x_{jt} - \bar{X}_j\right)\right]} a_{ij}^{'}$$

Ya que la desviación promedia de los puntajes $x_{.t}$ alrededor de su media es cero, la corrección para aleatoriedad:

$$a_{ij}^{c} = 2n\frac{\sum_{t=1}^{n}(x_{it} - \bar{X}_{i.})}{n}\frac{\sum_{t=1}^{n}\left(x_{jt} - \bar{X}_{j.}\right)}{n} = 0$$

Debido a la razón anterior, las versiones no corregidas y corregidas de las medidas de aditividad bivariada y multivariada son idénticas, por tanto, se omitirá el resto de la prueba. ■

Corolario 2.3: El coeficiente de proporcionalidad multivariado corregido (**3.34**) es la media ponderada del coeficiente de proporcionalidad bivariado corregido (**3.23**).

Prueba: Substituyendo la transformación uniformadora para la escala de razón $U = X/T_X$ en (**3.42**), se obtiene:

$$P_k^{'} = \sum_{i<j}^{k} \frac{\sum_{t=1}^{n}\left(\frac{x_{it}}{T_i}\right)^2 + \sum_{t=1}^{n}\left(\frac{x_{jt}}{T_j}\right)^2 - 2n^{-1}\sum_{t=1}^{n}\left(\frac{x_{it}}{T_i}\right)\sum_{t=1}^{n}\left(\frac{x_{jt}}{T_j}\right)}{\sum_{i<j}^{k}\left[\sum_{t=1}^{n}\left(\frac{x_{it}}{T_i}\right)^2 + \sum_{t=1}^{n}\left(\frac{x_{jt}}{T_j}\right)^2 - 2n^{-1}\sum_{t=1}^{n}\left(\frac{x_{it}}{T_i}\right)\sum_{t=1}^{n}\left(\frac{x_{jt}}{T_j}\right)\right]} p_{ij}^{'}$$

Después de cierta manipulación algebraica se tiene:

$$P_k^{'} = \sum_{i<j}^{k} \frac{1 - \bar{X}_i\bar{X}_j/T_iT_j}{\sum_{i<j}^{k}\left(1 - \bar{X}_i\bar{X}_j/T_iT_j\right)} p_{ij}^{'} \blacksquare$$

Corolario 2.4: El coeficiente de linealidad multivariado corregido (**3.35**) es el promedio del coeficiente de correlación de Pearson (**3.16**).

Prueba: Substituyendo la transformación uniformadora para la escala de intervalo $U = (X - \bar{X})/S_X$ en (**3.42**), se obtiene:

$$L_k^{'} = \sum_{i<j}^{k} \frac{\sum_{t=1}^{n}\left(\frac{x_{it}-\bar{X}_i}{S_i}\right)^2 + \sum_{t=1}^{n}\left(\frac{x_{jt}-\bar{X}_j}{S_j}\right)^2 - 2n^{-1}\sum_{t=1}^{n}\left(\frac{x_{it}-\bar{X}_i}{S_i}\right)\sum_{t=1}^{n}\left(\frac{x_{jt}-\bar{X}_j}{S_j}\right)}{\sum_{i<j}^{k}\left[\sum_{t=1}^{n}\left(\frac{x_{it}-\bar{X}_i}{S_i}\right)^2 + \sum_{t=1}^{n}\left(\frac{x_{jt}-\bar{X}_j}{S_j}\right)^2 - 2n^{-1}\sum_{t=1}^{n}\left(\frac{x_{it}-\bar{X}_i}{S_i}\right)\sum_{t=1}^{n}\left(\frac{x_{jt}-\bar{X}_j}{S_j}\right)\right]} r_{ij}^{'}$$

Ya que la desviación promedia de los puntajes $x_{.t}$ alrededor de su media es cero, la corrección para aleatoriedad:

$$r_{ij}^{c} = \frac{2n}{S_i S_j} \frac{\sum_{t=1}^{n}(x_{it} - \bar{X}_i)}{n} \frac{\sum_{t=1}^{n}(x_{jt} - \bar{X}_j)}{n} = 0$$

Debido a la razón anterior, las versiones no corregidas y corregidas de las medidas de linealidad bivariada y multivariada son idénticas, por tanto, aquí también se omitirán el resto de la prueba. ■

Corolario 2.5: El coeficiente de monotonía multivariado corregido (**3.36**) es la media ponderada del coeficiente de monotonía bivariado corregido (**3.29**).

Prueba: Substituyendo la transformación uniformadora para la escala ordinal $U = Rank(X)$ [transformada RT-2 de Conover e Iman (186)] en (**3.42**), se obtiene:

$$M_k^{'} = \sum_{i<j}^{k} \frac{\sum_{t=1}^{n} r_{it}^2 + \sum_{t=1}^{n} r_{jt}^2 - 2n^{-1} \sum_{t=1}^{n} r_{it} \sum_{t=1}^{n} r_{jt}}{\sum_{i<j}^{k}\left(\sum_{t=1}^{n} r_{it}^2 + \sum_{t=1}^{n} r_{jt}^2 - 2n^{-1} \sum_{t=1}^{n} r_{it} \sum_{t=1}^{n} r_{jt}\right)} m_{ij}^{'}$$

En el caso de mediciones basadas en rango, el promedio $(\sum_{t=1}^{n} r_{it})/n = (n+1)/2$ no es afectado por la presencia de puntajes atados; sin embargo, este no es el caso para $TR_X^2 = (\sum_{t=1}^{n} r_t^2)/n$. Cuando hay ausencia total de ataduras $TR_X^2 = \frac{(n+1)(2n+1)}{6}$, la expresión anterior se reduce al coeficiente de monotonía bivariado promedio o al CCR de Spearman promedio. ■

Los coeficientes multivariados no corregidos estadísticamente de identidad (I_k), proporcionalidad (P_k), monotonicidad (M_k) son reportados por primera vez en este trabajo y por ende son nuevos en la literatura científica (en la extensión del conocimiento del autor de la tesis) ya que no fueron reportados por Fagot (203).

La relación entre los coeficientes de acuerdo relacional multivariados para la escala nominal y su relación con sus contrapartes de orden inferior ha sido trabajada por otros autores (210-211).

Los resultados obtenidos en las subsecciones **3.3.2** y **3.3.3** sobre la relación entre los coeficientes de acuerdo relacional multivariados y bivariados no corregidos (corregidos), así como las fórmulas

algebraicas equivalentes de la literatura, pruebas de contraste estadísticas y los rangos de definición de las medidas multivariadas se resumen en la **Tabla 3.20**.

3.3.4. Coeficientes de Acuerdo Relacional aplicados como medidas de similitud biomoleculares ($k \geq 2$)- arias.

Con el objetivo de mostrar que las medidas RA multivariadas también pueden discriminar entre compuestos activos e inactivos, estas fueron calculadas como medidas de similitud biomolecular en dos conjuntos de datos diseñados *ad hoc* a partir de los conjuntos de datos quimioinformáticos MUV comprendiendo el subconjunto de activos solamente y el conjunto completo (activos y decoys), respectivamente. La prueba de Wilcoxon confirmó que los valores de similitud para compuestos activos son significativamente mayores que para activos + decoys (ver **Tabla 3.21**). Esto sugiere que todas las medidas multivariadas RA satisfacen la propiedad deseada $S_a >> S_{a+i}$ (los resultados de la prueba no son presentados). Sin embargo, las medidas RA ($k > 2$)-arias deben ser integradas en técnicas operativas análogas de VS basadas en similitud con el objetivo de evaluar correctamente el mejoramiento en el desempeño e interpretación de estas medidas sobre sus contrapartes binarias. Como consecuencia de su funcionalidad, estas medidas dependen linealmente del conjunto de las medidas RA por pares. En este sentido sería interesante encontrar nuevas familias de coeficientes de acuerdo relacional basadas en modelos de distancia de mayor grado polinómico y para nuevas escalas métricas, incluida la nominal. Para los problemas quimio(bio)informáticos el volumen V del poliedro formado por k biomoléculas-punto representa la distancia de mayor grado polinómico (65).

Tabla 3.20. Coeficientes multivariados de acuerdo relacional y algunas de sus propiedades.

Escala	Nombre	Fórmula	Fórmula equivalente	Prueba de inferencia	Rango
Absoluta	Identidad no corregida	$I_k = \sum_{i<j}^{k} \frac{\overline{T_{ij}^2}}{\sum_{i<j}^{k} \overline{T_{ij}^2}} e_{ij}$	Nuevo	-	[-1/(*k*-1), +1]
	Identidad corregida	$I_k^{'} = \sum_{i<j}^{k} \frac{\xi_{ij}}{\sum_{i<j}^{k} \xi_{ij}} e_{ij}^{'}$	- Coeficiente de identidad corregido de Fagot *I'*(203) - *Caso 3 ICC(A,1)* de McGraw y Wong (208) - *OCCC* de Barnhart y coautores (206) - Coeficiente Kappa de Cohen ponderado cuadráticamente (212)	- Estadístico *F* (208) - Enfoque GEE (206)	[-1/(*k*-1), +1]
Diferencia	Aditividad no corregida y corregida	$A_k = \sum_{i<j}^{k} \frac{\overline{S_{ij}^2}}{\sum_{i<j}^{k} \overline{S_{ij}^2}} a_{ij}$	- "Punto de anclaje" de Winner (179) - *ICC(3,1)* de Shrouty Fleiss (205) - Coeficiente de aditividad *A* de Fagot (203) - *Caso 3 ICC(C,1)* de McGraw y Wong (208)	Estadístico *F* (208)	[-1/(*k*-1), +1]
Razón	Proporcionalidad no corregida	$P_k = \bar{p}_{ij}$	Nuevo	-	[-1/(*k*-1), +1]
	Proporcionalidad corregida	$P_k^{'} = \sum_{i<j}^{k} \frac{\left(1 - \bar{X}_i \bar{X}_j / T_i T_j\right)}{\sum_{i<j}^{k} \left(1 - \bar{X}_i \bar{X}_j / T_i T_j\right)} p_{ij}^{'}$	-	NA	[-1/(*k*-1), +1]
Intervalo	Linealidad no corregida y corregida	$L_k = \bar{r}_{ij}$	-	NA	[-1/(*k*-1), +1]
Ordinal	Monotonicidad no corregida	$M_k = \sum_{i<j}^{k} \frac{\overline{TR_{ij}^2}}{\sum_{i<j}^{k} \overline{TR_{ij}^2}} m_{ij}$	Nuevo	-	[+0.75,+1]
	Monotonicidad corregida	$M_k^{'} = \overline{m^{'}}_{ij}$	- Coeficiente R_r de Haggard (213) - *W* de Kendall corregida por aleatoriedad (154) - Coeficiente ordinal *O'* de Fagot (204)	- Estadístico χ^2 de Friedman (214) - Estadístico *z* de Kendall y Babington (154) - Estadístico *F* de Iman y Daventport (153)	[-1/(*k*-1), +1]

Tabla 3.21. Resultados de las medidas RA multivariadas calculados en los conjuntos de datos MUV utilizando los compuestos: *i*) activos solamente (A) e *ii*) activos + inactivos (AI)

AID	I_k-A	I_k-AI	I'_k-A	I'_k-AI	A_k-A	A_k-AI	P_k-A	P_k-AI	P'_k-A	P'_k-AI	L_k-A	L_k-AI	M_k-A	M_k-AI	M'_k-A	M'_k-AI
466	0.0472	-0.0032	0.0484	-0.0037	0.0388	-0.0030	0.0132	-0.0024	0.5676	-0.0023	0.0249	-0.0024	0.8461	-0.0008	0.2523	-0.0021
548	0.3600	-0.0020	0.2822	-0.0041	0.2947	-0.0028	0.3888	-0.0037	0.1523	-0.0017	0.2897	-0.0022	0.8472	-0.0008	0.2361	-0.0016
600	0.0918	-0.0041	0.0856	-0.0040	0.0874	-0.0049	0.0637	-0.0027	0.0514	-0.0022	0.1030	-0.0040	0.8560	-0.0008	0.2080	-0.0023
644	0.2454	-0.0016	0.2326	-0.0012	0.2475	-0.0013	0.2552	-0.0018	0.2073	-0.0014	0.2631	-0.0016	0.8385	-0.0008	0.2731	-0.0015
652	0.1000	-0.0040	0.0942	-0.0025	0.0976	-0.0034	0.1168	-0.0047	0.1999	-0.0030	0.1167	-0.0029	0.8331	-0.0008	0.261	-0.0043
689	0.0408	-0.0017	0.0354	-0.0021	0.0296	-0.0017	0.0231	-0.0035	0.0629	-0.0025	0.0162	-0.0018	0.8359	-0.0008	0.1468	-0.0016
692	0.5124	-0.0029	0.4480	-0.0025	0.4841	-0.0019	0.5618	-0.0029	0.1162	-0.0028	0.6239	-0.0026	0.9624	-0.0009	0.6239	-0.0026
712	0.1135	-0.0021	0.1094	-0.0021	0.1162	-0.0022	0.1229	-0.0016	0.1530	-0.0031	0.1325	-0.0022	0.8077	-0.0008	0.1588	-0.0049
713	0.0635	-0.0048	0.0472	-0.0030	0.0452	-0.0049	0.0620	-0.0010	0.0089	-0.0025	0.0421	-0.0036	0.8421	-0.0008	0.2107	-0.0016
733	0.0618	-0.0022	0.0846	-0.0043	0.0741	-0.0039	0.0386	-0.0020	0.0293	-0.0019	0.0756	-0.0040	0.8385	-0.0008	0.1117	-0.0023
737	0.1667	-0.0019	0.1483	-0.0013	0.1780	-0.0019	0.1519	-0.0020	0.1136	-0.0010	0.2817	-0.0026	0.8960	-0.0008	0.3758	-0.0018
810	0.1288	-0.0027	0.1821	-0.0012	0.1813	-0.0047	0.1362	-0.0028	0.1075	-0.0028	0.1950	-0.0040	0.8347	-0.0008	0.2295	-0.0014
832	0.1356	-0.0018	0.1642	-0.0032	0.1632	-0.0022	0.1373	-0.0028	0.1430	-0.0027	0.1752	-0.0022	0.8298	-0.0008	0.1490	-0.0022
846	0.2826	-0.0021	0.2755	-0.0021	0.2873	-0.0021	0.3184	-0.0021	0.2374	-0.0036	0.3338	-0.0021	0.862	-0.0008	0.3446	-0.0009
852	0.2356	-0.0021	0.1944	-0.0024	0.2095	-0.0033	0.2251	-0.0046	0.1816	-0.0015	0.2187	-0.0039	0.8551	-0.0008	0.2031	-0.0029
858	0.0548	-0.0039	0.0430	-0.0040	0.0419	-0.0047	0.0448	-0.0010	0.0227	-0.0032	0.0434	-0.0012	0.8317	-0.0008	0.1826	-0.0017
859	0.3516	-0.0024	0.4053	-0.0031	0.3880	-0.0030	0.1996	-0.0018	0.0351	-0.0026	0.3014	-0.0020	0.9135	-0.0009	0.5244	-0.0038

3.4. Conclusiones parciales del capítulo.

En la primera fase de la investigación se compararon cuatro medidas de proximidad en la búsqueda de similitud de repositorios químico-médicos. Los resultados de la efectividad de las medidas y las pruebas estadísticas correspondientes revelaron la jerarquía Euclídea = Correlación = Ruzicka > Bray-Curtis II; por tanto, los resultados sugieren que la disimilitud de Ruzicka es equivalentemente útil a las medidas establecidas en la literatura para estos propósitos.

En la segunda fase se introdujeron nueve medidas de similitud bivariadas procedentes de la Teoría Estadística de las Mediciones, que permiten describir el tipo de relación funcional que existe entre los vectores de representación biomoleculares. En una etapa primaria, dichas medidas se integraron y compararon, junto a otras 12 medidas reportadas en la literatura, a un algoritmo de búsqueda de similitud. Los resultados preliminares, usando conjuntos químico-médicos de tamaño mediano a grande, mostraron que los modelos de proximidad basados en el acuerdo relacional se desempeñan relativamente superior 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 (monotonicidad no corregida), e^{Rc} (monotonicidad corregida), p^c (proporcionalidad corregida), a (aditividad) y e^c (identidad corregida) cuyo desempeño está por encima del obtenido con el coeficiente de Tanimoto, el cual había mostrado los mejores resultados en estudios previos.

En una segunda etapa de validación las nueve medidas de similitud del acuerdo relacional se integraron a un algoritmo de autoentrenamiento y compararon con un panel extenso de 21 modelos de proximidad, usando conjuntos de datos para validación y aplicación más apropiados y pruebas estadísticas más potentes para la comparación. Los resultados finales confirmaron que las medidas basadas en el acuerdo relacional se desempeñan superiormente a las que no proceden de esta teoría y que cuatro de las primeras cinco medidas más potentes son del acuerdo relacional, dentro de las cuales no está incluido el coeficiente de Tanimoto, estas fueron: p'_{XY} (proporcionalidad corregida),

m'_{XY} (monotonicidad corregida), e'_{XY} (identidad corregida) y m_{XY} (monotonicidad no corregida), serie de medidas que concuerda aproximadamente con la obtenida en la primera etapa.

En la tercera fase se obtuvieron algunos resultados teóricos interesantes. Se demostró que la fórmula general de acuerdo relacional multivariada no corregida (corregida) es la media ponderada de la fórmula general de acuerdo relacional bivariada no corregida (corregida). Como consecuencia, se probó que los coeficientes de acuerdo relacional multivariados no corregidos (corregidos) específicos para siete escalas de medición se pueden expresar como la media ponderada de sus contrapartes bivariadas no corregidas (corregidas). Usando esta dependencia funcional se proponen en este trabajo tres medidas de acuerdo relacional multivariados: I_k (identidad no corregida), P_k (proporcionalidad no corregida) y M_k (monotonicidad no corregida).

CONCLUSIONES

CONCLUSIONES

1. Se implementaron herramientas para el cribado virtual de conjuntos quimio-/bio-informáticos que consisten en 30 medidas de similitud para datos numéricos acopladas a un algoritmo de búsqueda rápida y la técnica de fusión de grupos (MAX-SIM). Se implementó también la técnica de validación cruzada de diez pliegues y el área bajo la curva CROC para la validación de modelos de proximidad.
2. Se introdujeron dos medidas de disimilitud de la Ecología en Quimioinformática, una de las cuales (Ruzicka) mostró un desempeño similar a las medidas estándares en la búsqueda de similitud de repositorios químico-médicos representados por descriptores numéricos.
3. Se introdujeron nueve medidas de similitud de la Teoría Estadística de las Mediciones, cinco de las cuales ($e_{XY}^{'}$, a_{XY}, $p_{XY}^{'}$, m_{XY},$m_{XY}^{'}$) mostraron un desempeño superior a las medidas reportadas (incluyendo a Tanimoto) en la búsqueda de similitud de repositorios quimioinformáticos, representados por descriptores numéricos relevantes. Se logró una interpretación cualitativa del tipo de funcionalidad latente entre los vectores de representación.
4. Se probó que las fórmulas generales de acuerdo relacional multivariadas, se pueden expresar como las medias ponderadas de las respectivas fórmulas bivariadas. Como consecuencia, se dedujeron tres nuevos coeficientes de acuerdo relacional multivariados para su uso en Quimioinformática y Estadística en sentido general.

RECOMENDACIONES

RECOMENDACIONES

1. Desarrollar nuevas medidas de similitud biomoleculares a partir del acuerdo relacional empleando otros modelos de distancia bivariados de partida y otras escalas métricas (incluida la nominal), así como otros modelos de distancia multivariados de grado polinómico superior para lograr una descripción más ajustada a las complejas relaciones de semejanza biomoleculares.
2. Integrar las medidas de similitud biomoleculares a las técnicas de búsqueda de similitud para convertirlas en técnicas de VS operativas que además permitan la evaluación del mejoramiento con respecto a sus contrapartes bivariadas usando datos biomoleculares simulados y reales característicos de la Quimio(Bio)informática contemporáneos.
3. Desarrollar nuevos algoritmos de búsqueda que sean más eficientes que los reportados en la literatura para lograr una recuperación más rápida en los enormes repositorios de datos quimio-/bio-informáticos.

REFERENCIAS BIBLIOGRÁFICAS

REFERENCIAS BIBLIOGRÁFICAS

1. Maggiora G, Shanmugasundaram V. Molecular similarity measures. In: Bajorath J., editor. Chemoinformatics and Computational Chemical Biology. Volume 672, Methods in Molecular Biology. New York: Humana Press; 2011. p 77-84.
2. Ágoston V, Kaján L, Carugo O, Hegedüs Z, Vlahovicek K, Pongor S. Concepts of similarity in bioinformatics. In: Moss DS, Jelaska S, Pongor S, editors. Essays in Bioinformatics. Volume 368, NATO Science Series, I: Life and Behavioural Sciences. The Netherland: IOS Press; 2005. p 11-31.
3. Johnson MA, Maggiora GM. Concepts and applications of molecular similarity. Johnson MA, Maggiora GM, editors. New York: Wiley; 1990.
4. Holyoak KJ, Thagard P. The analogical mind. *Am. Psychol.* 1997; **52**: 35-44.
5. Martin YC, Kofron JL, Traphagen LM. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* 2002; **45**: 4350-4358.
6. Skolnick J, Fetrow JS. From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends Biotechnol.* 2000; **18**: 34-39.
7. Dalkilic MM, Costello JC, Clark WT, Radivojac P. From protein-disease associations to disease informatics. *Front. Biosci.* 2008; **13**: 3391-3407.
8. Maldonado AG, Doucet JP, Petitjean M, Fan B-T. Molecular similarity and diversity in chemoinformatics: from theory to applications. *Mol. Divers.* 2006; **10**: 39-79.
9. Laskowski RA, Thornton JM. Understanding the molecular machinery of genetics through 3D structures. *Nat. Rev. Genet.* 2008; **9**: 141-151.
10. Hinz U, Consortium TU. From protein sequences to 3D-structures and beyond: the example of the UniProt Knowledgebase. *Cell. Mol. Life Sci.* 2010; **67**: 1049-1064.
11. Medina-Franco JL. Scanning structure–activity relationships with structure–activity similarity and related maps: from consensus activity cliffs to selectivity switches. *J. Chem. Inf. Model.* 2012; **52**: 2485-2493.
12. Punta M, Ofran Y. The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. *PLoS Comput. Biol.* 2008; **4**: e1000160.
13. Sturm N, Desaphy J, Quinn RJ, Rognan D, Kellenberger E. Structural insights into the molecular basis of the ligand promiscuity. *J. Chem. Inf. Model.* 2012; **52**: 2410-2421.

14. Rentzsch R, Orengo CA. Protein function prediction – the power of multiplicity. *Trends Biotechnol.* 2009; **27**: 210-219.
15. Hofstadter D. Epilogue: analogy as the core of cognition. In: Gentner D, Holyoak KJ, Kokinov BN, editors. The Analogical Mind: Perspectives from Cognitive Science, Bradford Books. Cambridge, Massachusetts: The MIT Press; 2001. p 499-538.
16. Juthe A. Argument by analogy. *Argumentation* 2005; **19**: 1-27.
17. Willett P. Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* 2005; **48**: 4183-4199.
18. Kaján L, Vlahovicek K, Carugo O, Ágoston V, Hegedüs Z, Pongor S. Comparison of sequences, protein 3D structures and genomes. In: Moss DS, Jelaska S, Pongor S, editors. Essays in Bioinformatics. Volume 368, NATO Science Series, I: Life and Behavioural Sciences. The Netherland: IOS Press; 2005. p 32-45.
19. Meinel T. Function and homology of proteins similar in sequence: Phylogenetic profiling. In: Daskalaki A, editor. Handbook of Research on Systems Biology Applications in Medicine. Volume 1. Hershey, New York: Medical information science reference; 2009. p 143-166.
20. Maggiora GM. On outliers and activity cliffs – why qsar often disappoints. *J. Chem. Inf. Model.* 2006; **46**: 1535-1535.
21. Stumpfe D, Bajorath J. Exploring activity cliffs in medicinal chemistry. *J. Med. Chem.* 2012; **55**: 2932-2942.
22. Bajorath J, Li R, Stumpfe D, Vogt M, Geppert HC. Development of a method to consistently quantify the structural distance between scaffolds and to assess scaffold hopping potential. *J. Chem. Inf. Model.* 2011; **51**: 2507-2514.
23. Koch MA, Waldmann H. Protein structure similarity clustering and natural product structure as guiding principles in drug discovery. *Drug Discov. Today* 2005; **10**: 471-483.
24. Agrafiotis DK. Diversity of chemical libraries. In: Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer III HF, Schreiner PR, editors. The Encyclopedia of Computational Chemistry. Volume 1. Chichester: John Wiley and Sons; 1998. p 742-761.
25. Snarey M, Terrett NK, Willett P, Wilton DJ. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graphics Modell.* 1997; **15**: 372-385.
26. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Šali A. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Bioph. Biom. Struc.* 2000; **29**: 291-325.

27. Vogt M, Bajorath J. Introduction of the conditional correlated Bernoulli model of similarity value distributions and its application to the prospective prediction of fingerprint search performance. *J. Chem. Inf. Model.* 2011; **51**: 2496-2506.
28. Geppert H, Bajorath J. Advances in 2D fingerprint similarity searching. *Expert. Opin. Drug. Discov.* 2010 **5**: 529-542.
29. Haranczyk M, Holliday J. Comparison of similarity coefficients for clustering and compound selection. *J. Chem. Inf. Model.* 2008; **48**: 498-508.
30. Trepalin S, Yarkov A. Hierarchical clustering of large databases and classification of antibiotics at high noise levels. *Algorithms* 2008; **1**: 183-200.
31. Downs GM, Willett P, Fisanick W. Similarity searching and clustering of chemical-structure databases using molecular property data. *J. Chem. Inf. Comput. Sci.* 1994; **34**: 1094-1102.
32. Al Khalifa A, Haranczyk M, Holliday J. Comparison of nonbinary similarity coefficients for similarity searching, clustering and compound selection. *J. Chem. Inf. Model.* 2009; **49**: 1193-1201.
33. Siegel S, Castellan NJ. Nonparametric statistics for the behavioral sciences. New York, USA: McGraw-Hill; 1988.
34. Talavera L. Dependency-based feature selection for clustering symbolic data. *Intell. Data Anal.* 2000; **4**: 19-28.
35. Manoranjan D, Choi K, Scheuermann P, Huan L. Feature selection for clustering: A filter solution. 2nd IEEE International Conference on Data Mining (ICDM'02); 2002 December 09-12; Maebashi City, Japan. IEEE Press. p 115-122.
36. Liu T, Liu S, Chen Z, Ma W-Y. An evaluation on feature selection for text clustering. In: Fawcett T, Mishra N, editors. 20th International Conference on Machine Learning (ICML-2003); 2003 August 21-24th; Washington DC AAAI Press, Menlo Park, California. p 488-495.
37. Law MHC, Figueiredo MAT, Jain AK. Simultaneous feature selection and clustering using mixture models. *IEEE T. Pattern Anal.* 2004; **26**: 1-13.
38. Yanjun L. Text clustering with feature selection by using statistical data. *IEEE T. Knowl. Data.* 2008; **20**: 641-652.
39. Böcker A, Derksen S, Schmidt E, Teckentrup A, Schneider G. A hierarchical clustering approach for large compound libraries. *J. Chem. Inf. Model.* 2005; **45**: 807-815.

40. Bender A, Mussa HY, Glen RC. Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier. *J. Chem. Inf. Comput. Sci.* 2004; **44**: 170-178.
41. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* 1996; **39** 3049-3059.
42. Nikolova N, Jaworska J. Approaches to measure chemical similarity - a review. *QSAR Comb. Sci.* 2003; **22** 1006-1026.
43. Biggs JB. The role of meta-learning in study process. *Brit. J Educ. Psychol.* 1985; **55**: 185-212.
44. Downs GM, Barnard JM. Clustering methods and their uses in computational chemistry. In: Lipkowitz KB, Boyd DB, editors. Reviews in Computational Chemistry Volume 18. Hoboken, New Jersey, USA: John Wiley and Sons; 2002. p 1-40.
45. Willett P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* 2006; **11**: 1046-1053.
46. Wolpert DH. The supervised learning no-free-lunch theorems. 6th Online World Conference on Soft Computing in Industrial Applications (WSC6); 2001. pp. 1-20. [online] http://ti.arc.nasa.gov/profile/dhw/statistical/ (visitado el 7 de octubre de 2013).
47. Holyoak KJ, Gentner D, Kokinov BN. Introduction: The place of analogy in cognition. In: Gentner D, Holyoak KJ, Kokinov BN, editors. The Analogical Mind: Perspectives from Cognitive Science, Bradford Books. Cambridge, Massachusetts: The MIT Press; 2001. p 1-19.
48. Grau Ábalo R. La lógica informal. Incompatibilidad y preludio de la investigación científica. COMPUMAT 2011; Santa Clara, Cuba.
49. Marraud H. La analogía como transferencia argumentativa. *Theoria* 2007; **59**: 167-188.
50. Adler J. Asymmetrical analogical arguments. *Argumentation* 2007; **21**: 83-92.
51. Rouvray DH. Definition and role of similarity concepts in the chemical and physical sciences. *J. Chem. Inf. Comput. Sci.* 1992; **32**: 580-586.
52. Quine WV. Ontological Relativity and Other Essays. New York: Columbia University Press; 1969.
53. Hahn U, Chater N, Richardson LB. Similarity as transformation. *Cognition* 2003; **87**: 1-32.

54. Goldstone RL, Son JY. Similarity. In: Holyoak KJ, Morrison R, editors. Cambridge Handbook of Thinking and Reasoning. Cambridge: Cambridge University Press; 2005. p 13-36.

55. Borg I, Groenen PJF. Modern Multidimensional Scaling: Theory and Applications. New York, USA: Springer; 2005.

56. Tversky A. Features of similarity. *Psychol. Rev.* 1977; **84**: 327-352.

57. Gentner D. Structure-mapping: a theoretical framework for analogy. *Cognitive Sci.* 1983; **7**: 155-170.

58. Gower JC, Legendre P. Metric and Euclidean properties of dissimilarity coefficients. *J. Classif.* 1986; **3**: 5-48.

59. Tversky A, Gati I. Similarity, separability, and the triangle inequality. *Psychol. Rev.* 1982; **89**: 123-154.

60. Cuadras CM. Distancias estadísticas. *Estadística Española* 1989; **30**: 295-378.

61. Deza E, Deza MM. Dictionary of distances. Oxford, UK: Elsevier; 2006.

62. Warrens MJ. Similarity Coefficients for Binary Data. Properties of Coefficients, Coefficient Matrices, Multi-way Metrics and Multivariate Coefficients [PhD.]. Leiden: Universiteit Leiden; 2008.

63. Cruz Monteagudo, M. *Métodos de Correlación Estructura-Actividad Multiobjetivos Aplicados al Desarrollo Racional de Fármacos.* Dr., Universidad Central "Marta Abreu" de Las Villas, 2009.

64. Hann, M.; Green, R. Chemoinformatics--a new name for an old problem? *Curr Opin Chem Biol*, 1999; 3: 379-383.

65. Sabitov IK. The volume as a metric invariant of polyhedra. *Discrete Comput. Geom.* 1998; **20**: 405-425.

66. Rivera Borroto, O. M.; Hernández Díaz, Y.; García de La Vega, J. M.; Grau Ábalo, R.; Marrero Ponce, Y. Novel similarity measures for the effective and efficient retrieval of pharmacological datasets. *Afinidad, 2011;* 68.

67. Gaifullin AA. Generalization of Sabitov's theorem to polyhedra of arbitrary dimensions. arXiv preprint arXiv:12105408 2012.

68. Xu J, Hagler A. Chemoinformatics and drug discovery. *Molecules* 2002; **7**: 566-700.

69. Chen WL. Chemoinformatics: past, present, and future. *J. Chem. Inf. Model.* 2006; **46**: 2230-2255.

70. Gasteiger J. Chemoinformatics: a new field with a long tradition. *Anal. Bioanal. Chem.* 2006; **384**: 57-64.

71. Warr WA. Some trends in chem (o) informatics. *Methods Mol. Biol.* 2011; **672**: 1-37.

72. Cohen J. Bioinformatics - an introduction for computer scientists. *ACM Comput. Surv.* 2004; **36**: 122-158.

73. Reddy AS, Pati SP, Kumar PP, Pradeep H, Sastry GN. Virtual screening in drug discovery-a computational perspective. *Curr. Protein Pept. Sci.* 2007; **8**: 329-351.

74. Seifert MHJ, Wolf K, Vitt D. Virtual high-throughput *in silico* screening. *Biosilico* 2003; **1**: 143-149.

75. Bajorath J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* 2002; **1**: 882-894.

76. Jorgensen W. The many roles of computation in drug discovery. *Science* 2004; **303**: 1813-1818.

77. Scior T, Bender A, Tresadern G, Medina-Franco JL, Martínez-Mayorga K, Langer T, Cuanalo-Contreras K, Agrafiotis DK. Recognizing pitfalls in virtual screening: a critical review. *J. Chem. Inf. Model.* 2012; **52** 867-881.

78. Brown RD. Descriptors for diversity analysis. *Perspect. Drug Disc. Design.* 1997; **7**: 31.

79. Maggiora GM, Shanmugasundaram V. Molecular Similarity Measures. In: Bajorath J, editor. Chemoinformatics. Volume 275: Humana Press; 2004. p 1-50.

80. Todeschini R, Consonni V. Molecular Descriptors for Chemoinformatics. Mannhold R, Kubinyi H, Folkers G, editors. Weinheim, Germany: WILEY-VHC; 2009.

81. Willett P, Barnard JM, Downs GM. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* 1998; **38**: 983-996.

82. National Center for Biotechnology Information. PubChem. Bethesda, Maryland, USA. http://pubchem.ncbi.nlm.nih.gov/ (visitado el 7 de octubre de 2013).

83. National Institutes of Health. National Cancer Institute. Bethesda, Maryland, USA. https://resresources.nci.nih.gov/resources/ (visitado el 7 de octubre de 2013).

84. The Cheminformatics and QSAR Society. New Hampshire, U.S.A. http://www.qsar.org (visitado el 7 de octubre de 2013).

85. International Academy of Mathematical Chemistry. Dubrovnik, Croatia. http://www.iamc-online.org/(visitado el 7 de octubre de 2013).

86. Daylight Chemical Information Systems. WDI. Laguna Niguel, CA, USA. . http://www.daylight.com(visitado el 7 de octubre de 2013).

87. Sunset Molecular Discovery. WOMBAT. New Mexico, USA. http://sunsetmolecular.com(visitado el 7 de octubre de 2013).

88. Baykoucheva S. A new era in chemical information: PubChem, DiscoveryGate, and Chemistry Central. Online 2007; **31 Issue , p16**: 16-20.

89. Bender A. Compound bioactivities go public. *Nature Chem. Biol.* 2010 **6**: 309.

90. Sheridan RP, Kearsley SK. Why do we need so many chemical similarity search methods? *Drug Discov. Today* 2002; **7**: 903-911.

91. Johnson MA. A review and examination of mathematical spaces underlying molecular similarity analysis. *J. Math. Chem.* 1989 **3**: 117-145.

92. Agrafiotis DK, Bandyopadhyay D, Wegner JK, van Vlijmen H. Recent Advances in chemoinformatics. *J. Chem. Inf. Model.* 2007; **47**: 1279-1293.

93. Wegner JK, Fröhlich H, Mielenz HM, Zell A. Data and graph mining in chemical space for ADME and activity data sets. *QSAR Comb. Sci.* 2006; **25**: 205-220.

94. Rivera Borroto OM. Estrategias QSAR combinadas, TOMOCOMD-CARDD y quimiométricas, para el Descubrimiento de candidatos a fármacos nuevos/novedosos frente a *trichomonas vaginalis* [MSc.]. Santa Clara: Universidad Central "Marta Abreu" de Las Villas; 2008.

95. Janecek A, Gansterer W, Demel M, Ecker G. On the relationship between feature selection and classification accuracy. In: Saeys Y, Liu H, Inza I, Wehenkel L, Van de Peer Y, editors; 2008 September 15; Antwerp, Belgium. JMLR: Workshop and Conference Proceedings. p 90-105.

96. Steinbach M, Ertöz L, Kumar V. The challenges of clustering high dimensional data. In: Wille LT, editor. New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition. Berlin: Springer-Verlag; 2000. p 273-307.

97. John GH, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. In: Cohen WW, Hirsh H, editors; 1994 July 10-13; Rutgers University, New Brunswick, NJ, USA. Morgan Kaufman. p 121-129.

98. Watanabe S. Knowing and guessing: A quantitative study of inference and information. New York: John Wiley & Sons Inc; 1969.

99. Böcker A, Schneider G, Teckentrup A. Status of HTS data mining approaches. *QSAR Comb. Sci*. 2004; **23**: 207-213.

100. Selwood DL, Livingstone DJ, Comley JCW, O'Dowd AB, Hudson AT, Jackson P, Jandu KS, Rose VS, Stables JN. Structure-activity relationships of antifilarial antimycin analogues, a multivariate pattern recognition study. *J. Med. Chem.* 1990; **33**: 136.

101. Zheng W, Tropsha A. Novel variable selection quantitative structure-property relationship approach based on the *k* nearest neighbor principle. *J. Chem. Inf. Comput. Sci.* 2000 **40**: 185-194.

102. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J. Mach. Lear. Research* 2003; **3**: 1157-1182.

103. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *SIGKDD Explor Newsl* 2009 **11**: 10-18. El software Weka esta disponible del Grupo de la Universidad de Waikato en: http://www.cs.waikato.ac.nz/ml/weka/(visitado el 7 de octubre de 2013).

104. Nicholls A. What do we know and when do we know it? *J. Comput.-Aided Mol. Des.* 2008; **22**: 239-255.

105. Truchon J, Bayly CI. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* 2007; **47**: 488-508.

106. Clark RD, Webster-Clark DJ. Managing bias in ROC curves. *J. Comput.-Aided Mol. Des.* 2008; **22**: 141-146.

107. Jain AK, Dubes RC. Algorithms for clustering data. Englewood Cliffs, New Jersey: Prentice-Hall; 1988.

108. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. *ACM Comput. Surv.* 1999; **31**: 264-323.

109. Arco García L. Agrupamiento basado en el concepto de intermediación diferencial y la aplicación de la teoría de los conjuntos aproximados para valorar resultados de agrupamientos [PhD.]. Santa Clara: Universidad Central "Marta Abreu" de Las Villas; 2008. 162 p.

110. Anderberg MR. Cluster analysis for applications. New York: Wiley; 1973.

111. Lance GN, Williams WT. A general theory of classificatory sorting strategies: 1. Hierarchical systems. *Comput. J.* 1967; **9**: 373-380.

112. Jambu M, Lebeaux MO. Classification automatique pour l'analyse des données (1. Méthodes et algorithmes. 2. Logiciels). Paris, France: Dunod; 1978. 310-400 p.

113. Jambu M, Lebeaux MO. Cluster analysis and data analysis. Amsterdam, Oxford: North-Holland 1983.

114. Dubien JL, Warde WD. A mathematical comparison of the members of an infinite family of agglomerative clustering algorithms. *Can. J. Stat.* 1979; **7**: 29-38.

115. Podani J. New combinatorial clustering methods. *Vegetatio* 1989; **81**: 61-77.

116. Batagelj V. Generalized Ward and related clustering problems. In: Bock HH, editor. Classification and Related Methods of Data Analysis. Amsterdam: North-Holland; 1988. p 67-74.

117. Hubálek Z. Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biol. Rev.* 1982; **57**: 669-689.

118. Murtagh F. A survey of recent advances in hierarchical clustering algorithms. *Comput. J.* 1983; **26**: 354-359.

119. Willet P. Clustering tendency in chemical classifications. *J. Chem. Inf. Comput. Sci.* 1985; **25**: 78-80.

120. Lawson RG, Jurs PC. New index for clustering tendency and its application to chemical problems. *J. Chem. Inf. Comput. Sci.* 1990; **30**: 36-41.

121. Everitt BS. Graphical Techniques for Multivariate Data. New York, NY: North Holland; 1978.

122. Fernández Pierna JA, Massart DL. Improved algorithm for clustering tendency. *Anal. Chim. Acta.* 2000; **408**: 13-20.

123. Massey L. Determination of clustering tendency with ART neural networks. 4th International Conference on Recent Advances in Soft Computing; 2002 12-13 December; Nottingham, U.K.

124. Veenman CJ, Reinders MJT, Backer E. A maximum variance cluster algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.* 2002; **24**: 1273-1280.

125. Forina M, Lanteri S, Esteban Díez I. New index for clustering tendency. *Anal. Chim. Acta.* 2001; **446**: 59-70.

126. Hodes L. Limits of classification. 2. Comment on Lawson and Jurs. *J. Chem. Inf. Comput. Sci.* 1992; **32**: 157-166.

127. Rządca K, Ferri F. Incrementally assessing cluster tendencies with a maximum variance cluster algorithm. Pattern Recognition and Image Analysis. Volume 2652, Lecture Notes in Computer Science: Springer Berlin / Heidelberg; 2003. p 868-875.

128. Cleveland WS. Visualizing Data. Summit, New Jersey: Hobart Press; 1993.

129. Tukey JW. Exploratory Data Analysis. Reading, MA Addison-Wesley; 1977.

130. Bezdek JC, Hathaway RJ. VAT: A tool for visual assessment of (cluster) tendency. 2002 International Joint Conference on Neural Networks (IJCNN'02); 2002; Piscataway, N. J. IEEE Press. p 2225-2230.
131. Willett P. Similarity methods in chemoinformatics. *Annu. Rev. Inf. Sci. Technol.* 2009; **43**: 1-117.
132. Willett P. Some heuristics for nearest-neighbor searching in chemical structure files. *J. Chem. Inf. Comput. Sci.* 1983; **23**: 22-25.
133. Friedman JH, Bentlev JL, Finkel RA. An algorithm for finding best matches in-logarithmic expected time. *ACM Trans. Math. Softw.* 1977; **3**: 209.
134. Bentley JL, Weide BW, Yao AC. Optimal expected time algorithms for closest point problems. *ACM Trans. Math. Softw.* 1980; **6**: 563.
135. Smeaton AF, Van Rijsbergen CJ. The nearest neighbour in information retrieval. An algorithm using upperbounds. *ACM SIGIR Forum* 1981; **16**: 83-87.
136. Murtagh F. A very fast, exact nearest neighbour algorithm for use in information retrieval. *Inf. Technol.: Res. Deu.* 1982; **1**: 275-283.
137. Van Marlen G, Van Den Hende JH. Search strategy and data compression for a retrieval system with binary-coded mass spectra. *Anal. Chim. Acta.* 1979; **112**: 143-150.
138. Rasmussen GT, Isenhour TL, Marshall JC. Mass spectral library searches using ion series data compression. *J. Chem. Inf. Comput. Sci.* 1979; **19**: 98-104.
139. Baldi P, Hirschberg DS, Nasr RJ. Speeding up chemical database searches using a proximity filter based on the logical exclusive OR. *J. Chem. Inf. Model.* 2008; **48**: 1367-1378.
140. Cao Y, Jiang T, Girke T. Accelerated similarity searching and clustering of large compound sets by geometric embedding and locality sensitive hashing. *Bioinformatics* 2010; **26**: 953-959.
141. Willett P. Data fusion in ligand-based virtual screening. *QSAR Comb. Sci.* 2006; **25**: 1143-1152.
142. Podani J. SYN-TAX2000. Computer Programs for Data Analysis in Ecology and Systematics. User's Manual Scientia Publishing, Budapest 2001. El programa paquete SYN-TAX esta diseñado para análisis de datos multivariados en SYNbiología (o Ecología) y TAXonomía (o Sistemática) y esta disponible del profesor JánosPodani en http://ramet.elte.hu/~podani/subindex.html(visitado el 7 de octubre de 2013).
143. Sutherland JJ, O'Brien LA, Weaver DF. A comparison of methods for modeling quantitative structure−activity relationships. *J. Med. Chem.* 2004; **47**: 5541-5554.

144. JChem for Excel. 5.3.8 (166). Budapest, Hungary; 2010. JChemfor Excel es una herramienta integrada a Microsoft Excel que le permite al científico gestionar y analizar estructuras químicas y datos numéricos. El software está disponible de ChemAxonKft. en http://www.chemaxon.com(visitado el 7 de octubre de 2013).
145. Sadowski J, Gasteiger J, Klebe G. Comparison of automatic three-dimensional model builders using 639 X-Ray structures. *J. Chem. Inf. Comput. Sci.* 1994; **34**: 1000-1008. El software generador de estructuras 3D CORINA está disponible de Molecular Networks GmbH en http://www.molecular-networks.com(visitado el 7 de octubre de 2013).
146. DRAGON for Windows 5.5. Milano, Italy; 2007. El software para cálculo de descriptores numéricos DRAGON está disponible de Talete srl en http://www.talete.mi.it(visitado el 7 de octubre de 2013).
147. Hall MA. Correlation-based Feature Subset Selection for Machine Learning [PhD.]. Hamilton, New Zealand: The University of Waikato; 1998.
148. Podani J. A method for generating consensus partitions and its application to community classification. *Coenoses* 1989; **4**: 1-10.
149. Podani J. Explanatory variables in classifications and the detection of the optimum number of clusters. In: Hayashi C, Ohsumi N, Yajima K, Tanaka Y, Bock HH, editors. Data Science, Classification and Related Methods. Tokyo, Japan: Springer; 1998. p 125-132.
150. Baldi P, Brunak S, Yves C, Andersen CAF, Nielsen H. Assesing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 2000; **16**: 412-424.
151. Rohrer SG, Baumann K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.* 2009; **49**: 169-184.
152. Swamidass SJ, Azencott CA, Daily K, Baldi P. A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. *Bioinformatics* 2010; **26**: 1348-1356.
153. Iman RL, Davenport JM. Approximations of the critical region of the Friedman statistic. *Commun. Stat. Theory* 1980; **9**: 571-595.
154. Kendall MG, Smith BB. The Problem of *m* Rankings. *Ann. Math. Statist.* 1939; **10**: 275-287.
155. García S, Fernández A, Luengo J, Herrera F. A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Comput.* 2009; **13**: 959-977.
156. Li J. A two-step rejection procedure for testing multiple hypotheses. *J. Stat. Plan. Infer.* 2008; **138**: 1521-1527.

157. Demšar J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Lear. Research* 2006; **7**: 1-30.
158. García S, Herrera F. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *J. Mach. Lear. Research* 2008 **9** 2677-2694.
159. Milligan GW. Ultrametric hierarchical clustering algorithms. *Psychometrika* 1979; **44**: 343-346.
160. Batagelj V. Note on ultrametric clustering algorithms. *Psychometrika* 1981; **46**: 351-352.
161. Diday E. Inversions en classification hiérarchique: application á la construction adaptive d'indices d'agrégation. *Rev. Stat. Appl.* 1983; **31**: 45-62.
162. Bruce CL, Melville JL, Pickett SD, Hirst JD. Contemporary QSAR classifiers compared. *J. Chem. Inf. Model.* 2007; **47**: 219-227.
163. Sönströd C, Johansson U, Norinder U. Generating comprehensible QSAR models. In: Johansson R, van Laere J, Mellin J, editors. 3rd Skövde Workshop on Information Fusion Topics 2009 (SWIFT 2009); 2009 Oct 12-13; Skövde, Sweden. University of Skövde. p 44-48.
164. Johansson U, Löfström T, Norinder U. Evaluating ensembles on QSAR classification. In: Johansson R, van Laere J, Mellin J, editors. 3rd Skövde Workshop on Information Fusion Topics 2009 (SWIFT 2009); 2009 Oct 12-13; Skövde, Sweden. Univeristy of Skövde. p 49-54.
165. Ivanciuc O. Applications of support vector machines in chemistry. In: Lipkowitz KB, Cundari TR, editors. Reviews in Computational Chemistry. Volume 23: Wiley-VCH, Weinheim; 2007. p 291-400.
166. Huband JM, Bezdek JC, Hathaway RJ. Revised visual assessment of (cluster) tendency (reVAT). In: Dick S, Kurgan L, Musilek P, Pedrycz W, Reformat M, editors. IEEE Annual Meeting of the Fuzzy Information Processing (NAFIPS '04); 2004 Jun 27-30; IEEE, Piscataway. NJ, USA p 101-104.
167. Huband JM, Bezdek JC, Hathaway RJ. bigVAT: Visual assessment of cluster tendency for large data sets. *Pattern Recog.* 2005; **38**: 1875-1886.
168. Hathaway RJ, Bezdek JC, Huband JM. Scalable visual assessment of cluster tendency for large data sets. *Pattern Recog.* 2006; **39**: 1315-1324.
169. Hu Y, Hathaway RJ. Tendency curves for visual clustering assessment. In: Demiralp M, Mikhael W, Caballero A, Abatzoglou N, Tabrizi M, Leandre R, Garcia-Planas M, Choras R,

editors. WSEAS International Conference on Applied Computing; 2008; Stevens Point, Wisconsin. World Scientific and Engineering Academy and Society (WSEAS). p 274-279.

170. Havens T, Bezdek J, Keller J, Popescu M, Huband J. Is VAT really single linkage in disguise? *Ann. Math. Artif. Intell.* 2008; **55**: 237-251.
171. Stein B, Meyer zu Eissen S, Wißbrock F. On cluster validity and the information need of users. In: Hanza MH, editor. 3rd IASTED International Conference on Artificial Intelligence and Applications (AIA'03); 2003 Sep 8-10; Benalmádena, Spain. ACTA Press. p 216-221.
172. Havens TC, Bezdek JC, Keller JM, Popescu M. Dunn's cluster validity index as a contrast measure of VAT images. 19th International Conference on Pattern Recognition (ICPR'08); 2008 Dec 8-11; Tampa, FL. p 1-4.
173. Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. Cybernetics* 1974; **3**: 32-57.
174. García V, Sánchez JS, Mollineda RA. An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. In: Rueda L, Mery D, Kittler J, editors. CIARP: Lecture Notes in Computer Science. Volume 4756. Heidelberg: Springer-Verlag; 2007. p 397-406.
175. Rivera Borroto OM, Hernández Llanes D, Marrero Ponce Y, Grau Ábalo RC, Rodríguez Machín L, García de la Vega JM, López Fernández R, González-Jonte Cruz R. Comparación de algoritmos de clústeres combinatorios novedosos en Quimioinformática. In: Bello Pérez R, editor. Conferencia Internacional de Ciencias Computacionales e Informáticas (CICCI' 2011); 2011 7-11 Febrero; La Habana, Cuba. CITMATEL. p CCI012.
176. Zegers FE, ten Berge JMF. A family of association coefficients for metric scales. *Psychometrika* 1985; **50**: 17-24.
177. Zegers FE. A family of chance-corrected association coefficients for metric scales. *Psychometrika* 1986; **51**: 559-562.
178. Stine WW. Interobserver Relational Agreement. *Psychol. Bull.* 1989; **106**: 341-347.
179. Winer BJ. Statistical principles in experimental design. New York: McGraw-Hill; 1971.
180. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A. Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbour information. *J. Med. Chem.* 2005; **48**: 7049-7054.
181. Sawilowsky SS. Fermat, Schubert, Einstein, and Behrens-Fisher: the probable difference between two means when σ_1^2 ≠ σ_2^2. *J. Mod. App. Stat.* 2002; **1**: 461-472.

182. Rivera-Borroto OM, Marrero-Ponce Y, García-de la Vega JM, Grau-Ábalo RC. Comparison of combinatorial clustering methods on pharmacological data sets represented by machine learning-selected real molecular descriptors. *J. Chem. Inf. Model.* 2011; **51**: 3036-3049.

183. Rivera-Borroto OM, Rabassa-Gutiérrez M, Grau-Ábalo RC, Marrero-Ponce Y, García-de la Vega JM. Dunn's index for cluster tendency assessment of pharmacological data sets. *Can. J. Physiol. Pharmacol.* 2012; **90**: 425-433.

184. Mojena R. Hierarchical grouping methods and stopping rules: an evaluation. *Comput. J.* 1977; **20**: 359-363.

185. Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurements. *J. Biopharm. Stat.* 2007; **17**: 529-569.

186. Conover WJ, Iman RL. Rank transformations as a bridge between parametric and nonparametric statistics. *Am. Stat.* 1981; **35**: 124-129.

187. Stine WW. Meaningful inference: the role of measurement in statistics. *Psychol. Bull.* 1989; **105**: 147-155.

188. Varin T, Bureau R, Mueller C, Willett P. Clustering files of chemical structures using the Székely–Rizzo generalization of Ward's method. *J. Mol. Graphics Mod.* 2009; **28**: 187-195.

189. Nasr RJ, Swamidass SJ, Baldi PF. Large scale study of multiple-molecule queries. *J. Cheminform.* 2009; **1**: 1-19.

190. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* 2004; **44**: 1177-1185.

191. Jobson J. A coefficient of equality for questionnaire items with interval scales. *Educ. Psychol. Meas.* 1976; **36**: 271-274.

192. Lin LI-KL. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; **45**: 255-268.

193. King TS, Chinchilli VM. A generalized concordance correlation coefficient for continuous and categorical data. *Stat. Med.* 2001; **20**: 2131-2147.

194. McDonald RP. Linear versus nonlinear models in item response theory. *Appl. Psychol. Meas.* 1982; **6**: 379-396.

195. Cureton EE. The definition and estimation of test reliability. *Educ. Psychol. Meas. 1958;* **18**:715-738.

196. Mehta J, Gurland J. Some properties and an application of a statistic arising in testing correlation. *Ann. Math. Statist.* 1969; **40**: 1736-1745.

197. Kristof W. On a statistic arising in testing correlation. Psychometrika 1972; **37**: 377-384.
198. Burt C. The factorial study of temperamental traits. *Brit. J. Psychol.* 1948; **1**: 178-203.
199. Tucker LR. A method for synthesis of factor analysis studies. Priceton, NJ: Eucational testing servise; 1951. Rep. Num. 984.
200. Sjöberg L, Holley JW. A measure of similarity between individuals when scoring directions of variables are arbitrary. *Multivar. Behav. Res.*1967; **2**: 377-384.
201. Kendall MG, Kendall SFH, Smith BB. The distribution of Spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times. *Biometrika* 1939; **30**: 251-273.
202. Truchon J-F, Bayly CI. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* 2007; **47**: 488-508.
203. Fagot RF. A generalized family of coefficients of relational agreement for numerical scales. *Psychometrika* 1993; **58**: 357-370.
204. Fagot RF. An ordinal coefficient of relational agreement for multiple judges. *Psychometrika* 1994; **59**: 241-251.
205. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* 1979; **86**: 420-428.
206. Barnhart HX, Haber M, Song J. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics* 2004; **58**: 1020-1027.
207. Haber M, Barnhart HX. Coefficients of agreement for fixed observers. *Stat. Methods Med. Res.* 2006; **15**: 255-271.
208. McGraw KO, Wong S. Forming inferences about some intraclass correlation coefficients. *Psychol. Methods* 1996; **1**: 30-46.
209. Fagot R, Mazo R. Association coefficients of identity and proportionality for metric scales. *Psychometrika* 1989; **54**: 93-104.
210. Warrens MJ. Cohen's kappa is a weighted average. *Stat. Methodol.* 2011; **8**: 473-484.
211. Warrens MJ. Cohen's linearly weighted kappa is a weighted average. *Adv. Data Anal. Classif.* 2012; **6**: 67-79.
212. Schuster C. A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educ. Psychol. Meas.* 2004; **64**: 243-253.
213. Haggard EA. The application of intraclass correlation to data in the form of ranks. Intraclass correlation and the analysis of variance. New York: Dryden Press; 1958 p123-163.

214. Friedman M. A comparison of alternative tests of significance for the problem of m rankings. *Ann. Math. Statist.* 1940; **11**: 86-92.