REVISIÓN

Perspectiva general sobre el proceso de desarrollo de fármacos y las técnicas de cribado virtual basadas en la similitud molecular

Óscar Miguel Rivera Borroto[1, 3*], Yoandy Hernández Díaz[1], José Manuel García de la Vega[2], Ricardo Grau[1], Yovani Marrero Ponce[3], Maikel Cruz Monteagudo[4]

[1]Laboratorio de Bioinformática, Centro de Estudios de Informática, Facultad de Matemática, Física y Computación, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, 54830 Villa Clara, Cuba. [2]Departamento de Química Física Aplicada, Facultad de Ciencias, Universidad Autónoma de Madrid (UAM), 28049 Madrid, España. [3]Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatics Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba. [4]Applied Chemistry Research Center-Faculty of Chemistry and Pharmacy, Molecular Simulation and Drug Design Group, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, 54830, Cuba.
*e-mail: oscarrb@uclv.edu.cu

Recibido el 2 de octubre de 2013 *An. Real Acad. Farm. Vol 79, Nº 4 (2013), pag. 530-561*

RESUMEN

El desarrollo de fármacos es una tarea en extremo compleja pero también muy apreciada por la sensibilidad que genera el impacto negativo de las enfermedades en la sociedad moderna. En este trabajo de revisión se tratarán las características generales del paradigma tradicional del proceso de desarrollo de fármacos. Posteriormente, se abordarán las técnicas de cribado virtual basadas en el concepto de similitud molecular como alternativa racional y complementaria a las primeras fases dicho proceso. En este sentido, se hará énfasis en la búsqueda de similitud y sus componentes esenciales.

Palabras clave: Proceso de desarrollo de fármacos; Técnicas de cribado virtual; Similitud molecular; Búsqueda de similitud.

ABSTRACT

Overview on the drug development process and molecular similarity-based virtual screening techniques

Drug development is a very complex task but also very appreciated by the sensibility that generates the negative impact of diseases in modern society. In this review, we will address the general characteristics of the traditional paradigm of drug development pipeline. Later, virtual screening techniques will be

introduced as a rational and complementary alternative to the early stages of this process. In this sense, we will focus on similarity searching and its key components.

Keywords: Drug development process; Virtual screening techniques; Molecular similarity; Similarity searching.

1. INTRODUCCIÓN

El desarrollo de una terapia para una patología específica es un proceso usualmente estructurado en tres pasos. El primer paso -identificación de la diana biológica o terapéutica- consiste en la identificación de una molécula biológica, mayormente proteínas, involucrada en algún mecanismo que participa en algún proceso patológico. El propósito del segundo paso es identificar una molécula con un perfil biológico interesante, capaz de interferir con el blanco terapéutico antes mencionado. Eventualmente, antes de que el candidato a fármaco entre al mercado, en el tercer paso -validación clínica- debe probar su eficiencia y seguridad a través de una evaluación extensiva en animales y humanos (1, 2).

1.1. Identificación de la diana biológica o terapéutica

El objetivo principal en la investigación terapéutica es interferir alguna vía o señal metabólica responsable de una enfermedad o proceso patológico. Las vías o señales metabólicas son cascadas de reacciones químicas intracelulares que llevan respectivamente a la formación de un producto metabólico que es usado por la célula, o a una alteración de la expresión de un gen debido a la activación de factores de transcripción. La tarea de la investigación terapéutica es encontrar una molécula de fármaco capaz de modificar esta vía mediante la alteración de una entidad clave involucrada en la cascada de reacciones correspondiente: el blanco terapéutico. La identificación del blanco involucra conocimientos tanto biológicos como químicos, con el objetivo de descubrir blancos potenciales y conocer en qué medida este puede ser alterado por una molécula de fármaco (2). Previo a la fase de descubrimiento de fármacos, el blanco terapéutico identificado debe ser validado con el objetivo de demostrar su papel determinante en la enfermedad. Esta validación usualmente involucra experimentos *in vitro* e *in vivo* (3).

1.2. Descubrimiento de fármacos

En este segundo paso, el objetivo es encontrar una molécula pequeña, denominada ligando, capaz de unirse mediante fuerzas intermoleculares al blanco biológico y alterar su funcionamiento normal. Esta interacción se dice que es directa cuando el fármaco se une al sitio activo del blanco y compite con su sustrato natural, o indirecta si el fármaco se une a un sitio secundario e induce cambios en la conformación química del blanco, modulando así su afinidad con el ligando natural (4). Para cuantificar la actividad del ligando, correspondiente al

grado de interacción con el blanco, se debe diseñar un procedimiento experimental denominado ensayo de la actividad biológica. La actividad de las moléculas candidatas puede ser subsecuentemente ensayada con el objetivo de encontrar candidatos a fármacos, o compuestos líderes, capaces de interferir con el blanco a bajas concentraciones (1). La identificación de candidatos prometedores en esta vasta (casi infinita) cantidad de moléculas depende fuertemente de la pericia bioquímica y que tradicionalmente se logra en un proceso iterativo, denominado como el *ciclo de descubrimiento de fármacos* que alterna entre los pasos de selección, síntesis y ensayo biológico de los candidatos, guiando este último al próximo paso de selección (5).

Durante los ensayos biológicos iniciales del ciclo de descubrimiento de fármacos son identificados las entidades noveles o "hits". A esta fase de generación de *hits* le sigue la fase de generación de cabezas de serie, líderes o "leads", donde los *hits* identificados son validados mediante ensayos confirmativos y refinados estructuralmente con el objetivo de incrementar su potencia con respecto al blanco. De lograrse una potencia suficiente, se pueden realizar ensayos biológicos adicionales para asegurar que el compuesto líder no interacciona con proteínas homólogas al blanco, con el fin de limitar sus efectos secundarios (6).

Hasta este punto, es posible identificar compuestos líderes con perfiles de unión al blanco adecuado. Sin embargo, el fármaco no solo debe interferir con el blanco terapéutico, sino que además debe poseer un perfil biológico favorable, específicamente una toxicidad baja, de manera que no sea dañino para el organismo, y propiedades farmacocinéticas adecuadas. De manera general, la farmacocinética está relacionada con el comportamiento de un fármaco en el organismo, tales como su capacidad de pasar al torrente circulatorio y alcanzar el blanco, y ser posteriormente destruido y eliminado por el organismo. Las principales propiedades farmacocinéticas se resumen en el acrónimo ADME, que incluye los procesos de Absorción, Distribución, Metabolismo y Excreción (7, 8).

De este modo, el desarrollar un medicamento exitoso es el resultado del descubrimiento del mejor compromiso entre numerosos objetivos que muy a menudo compiten entre sí. El fracaso de un candidato a fármaco con una potencia adecuada durante el proceso de desarrollo es debido principalmente a una pobre biodisponibilidad, y/o toxicidad (9). De forma simplificada, el fármaco ideal debería tener la mayor eficacia terapéutica y biodisponibilidad, y la mínima toxicidad posible, lo que evidencia la naturaleza multiobjetiva del proceso de descubrimiento de fármacos (ver Figura 1). Lo anterior sugiere que en la fase de optimización del líder, la capacidad de mejorar el perfil terapéutico del candidato seleccionado basándose solamente en su actividad farmacológica se ha sobreestimado lo que refuerza, durante la fase de identificación del líder, considerar las propiedades toxicológicas y farmacocinéticas del candidato

paralelamente a sus propiedades farmacológicas en etapas anteriores a la optimización (10). Todo lo anterior ha llevado tanto a la academia como a la industria farmacéutica a una reconsideración del paradigma secuencial del proceso de descubrimiento de fármacos en favor a un enfoque multiobjetivos del proceso del mismo (11, 12).

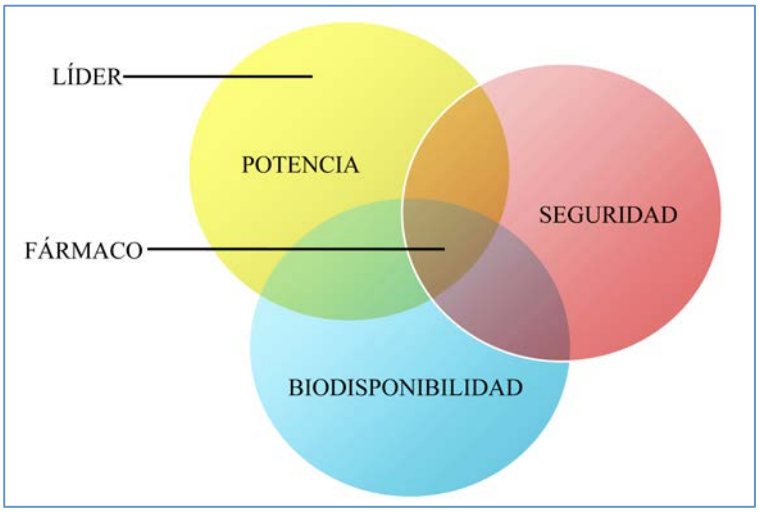


Figura 1.- Ilustración gráfica del compromiso entre eficacia terapéutica (potencia), biodisponibilidad (propiedades ADME) y toxicidad (seguridad) requerido para alcanzar un fármaco exitoso.

La fase final del descubrimiento de fármacos es la fase de optimización del líder, donde se refina la estructura química del mismo de manera que cumpla con los criterios requeridos para convertirse en un fármaco. Este proceso de optimización es altamente iterativo y se considera la fase más crítica del proceso de descubrimiento de fármacos ya que es aquí donde ocurre la mayor cantidad de fallas. Una vez descubierto un compuesto líder con características de fármaco prometedoras, el paso final hacia la puesta en el mercado del fármaco es la fase de validación clínica (11).

1.3. Validación clínica

Previo a la puesta en el mercado, el candidato a fármaco debe ser validado durante una fase de prueba extensiva, dirigida a demostrar su eficacia y seguridad para el organismo humano: la *validación clínica*. Esta fase comienza con la realización de pruebas preliminares de seguridad en animales, la etapa preclínica, y es subsecuentemente articulada en tres fases (11):

• Fase I (1 a 2 años): Inicialmente se llevan a cabo pruebas de seguridad con un número limitado (< 100) de personas sanas.

• Fase II (1 a 2 años): Seguidamente se llevan a cabo pruebas de seguridad y eficacia a una muestra mayor compuesta por cientos de personas que incluye grupos de sanos y enfermos.

• Fase III (2 a 3 años): Finalmente, el estudio se completa con la realización de pruebas de eficacia a gran escala, las que involucran una muestra mucho mayor de personas (miles) de diferentes áreas demográficas.

Eventualmente, una vez previstos los resultados de este estudio clínico y concedida la aprobación gubernamental, entonces puede comenzar la explotación comercial del fármaco. La aprobación gubernamental es concedida, por ejemplo, por la Administración de Alimentos y Medicamentos (FDA, del inglés Food and Drugs Administration) en los Estados Unidos de América, por la Agencia Europea de Medicamentos (EMA, del inglés European Medicines Agency) en la comunidad Europea, por la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS) en España o para el caso particular de Cuba por el Centro para el Control Estatal de la Calidad de los Medicamentos (CECMED).

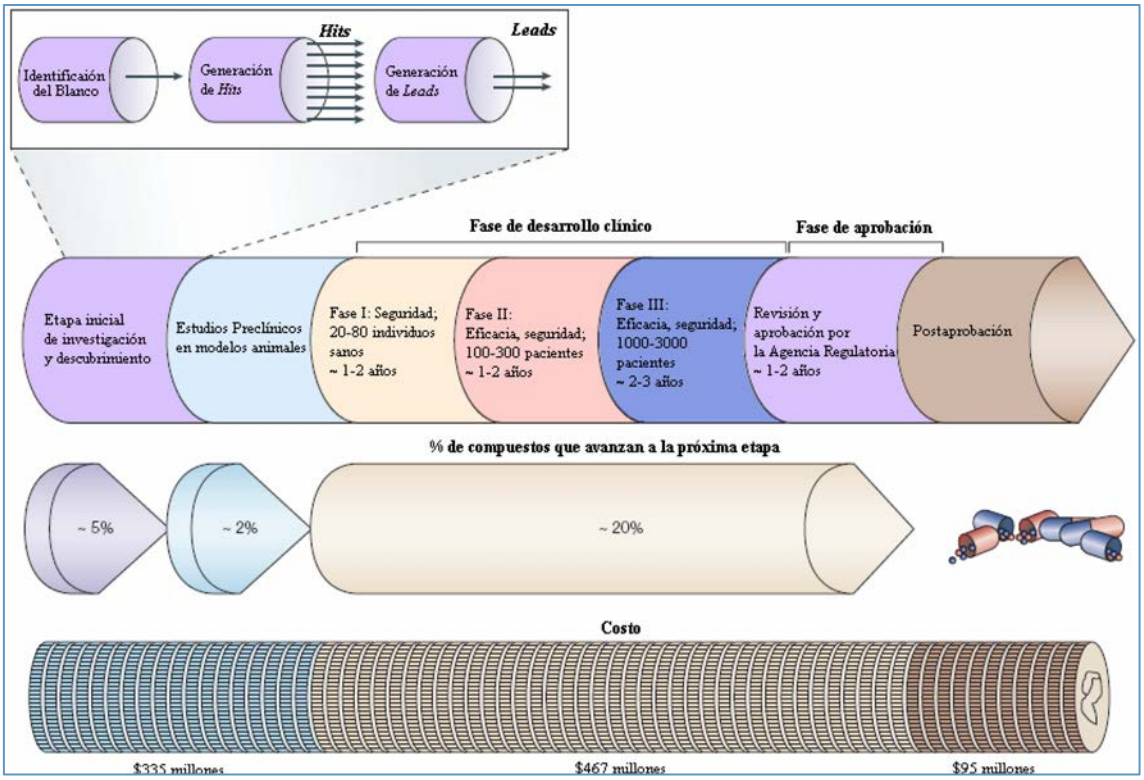


Figura 2.- Representación esquemática general del proceso de descubrimiento y desarrollo de fármacos.

Un estudio relativamente reciente llevado a cabo por el *Boston Consulting Group* (BCG) y que involucró a 50 compañías e instituciones académicas, mostró que el coste de desarrollo de un nuevo medicamento desde la identificación de su diana farmacológica, el descubrimiento y optimización de uno de los nuevos compuestos líderes, el desarrollo de los ensayos clínicos requeridos y su uso autorizado en terapéutica es como promedio de 880 millones de dólares (USD) y se necesita también como promedio un período de 15 años de investigación [ver figura 2] (13)

1.4. Necesidad de nuevos paradigmas

Hasta los años 80, el paso de generación de *hits* o candidatos potenciales (moléculas que muestran una determinada actividad química pero que no necesariamente cumplen con los requerimientos de eficiencia de un *lead* o compuesto líder) constituía el principal paso limitante del proceso de descubrimiento y desarrollo de nuevos fármacos (DDDP, del inglés Drug Discovery and Development Process) debido al costo de la síntesis y evaluación de nuevas moléculas (10). Durante esta etapa las esperanzas de resolver el problema del DDDP fueron puestas en el desarrollo de las tecnologías de alto rendimiento (14) y la química combinatoria (15), a través de una paralelización masiva del proceso. En la práctica, se evidenció que si no eran utilizadas cuidadosamente, el uso indiscriminado de estas técnicas podría conducir a un aumento dramático del número de moléculas o candidatos, de manera que el descubrimiento de un nuevo fármaco sería como hallar una aguja en un pajar. Mientras que el número de *hits* identificados pudo ser incrementado sustancialmente, se observó que no existía una correspondencia con el crecimiento del número de fármacos que entraban al mercado, dejando esto claro que el verdadero paso limitante del descubrimiento de fármacos no era la generación de *hits*, sino los pasos de identificación y optimización del compuesto líder (10). Como resultado, este tipo de solución a gran escala ha sido abandonada progresivamente en los últimos años, favoreciéndose una racionalización del proceso, en la que los métodos computacionales han ganado una importancia creciente (10).

2. MÉTODOS COMPUTACIONALES O *IN SILICO*

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 (7). Para unificar la combinación de los métodos computacionales y la Química Medicinal, 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"* (16). Actualmente, este concepto aparece definido de una manera más amplia para considerar la quimioinformática como "*la aplicación de métodos informáticos para resolver problemas químicos*" (17). Esta definición general engloba múltiples aspectos como la representación, almacenamiento, recolección y análisis de la información química en un sistema informático. Algunos de los frentes de trabajo abiertos en esta área relativamente joven continúan siendo la minería de textos químicos, los estudios QSAR, el diseño de fármacos basado en estructura y el diseño de fármacos basado en fragmentos (18).

2.1. Cribado virtual

El cribado virtual (VS, del inglés Virtual Screening) *in silico* 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 (19). Este paradigma "más racional" puede verse como una alternativa al cribado de alto rendimiento (HTS, del inglés High Throughput Screening) con las ventajas de que pueden ser evaluadas *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 (20). En esencia, los enfoques HTS y VS poseen una naturaleza complementaria entre sí, por cuanto se han introducido varios y diversos conceptos y métodos computacionales para analizar datos de cribado experimental, extraer conocimiento de los experimentos HTS y derivar modelos predictivos de actividad (21).

En la práctica, el cribado virtual 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 cribado virtual son los estudios de acoplamiento o "docking", los 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 (6). Este enfoque de cribado virtual suele brindar los resultados más confiables y al mismo tiempo resulta atractivo por el gran número (alrededor de 500) de dianas farmacológicas disponibles (ver Figura 3).

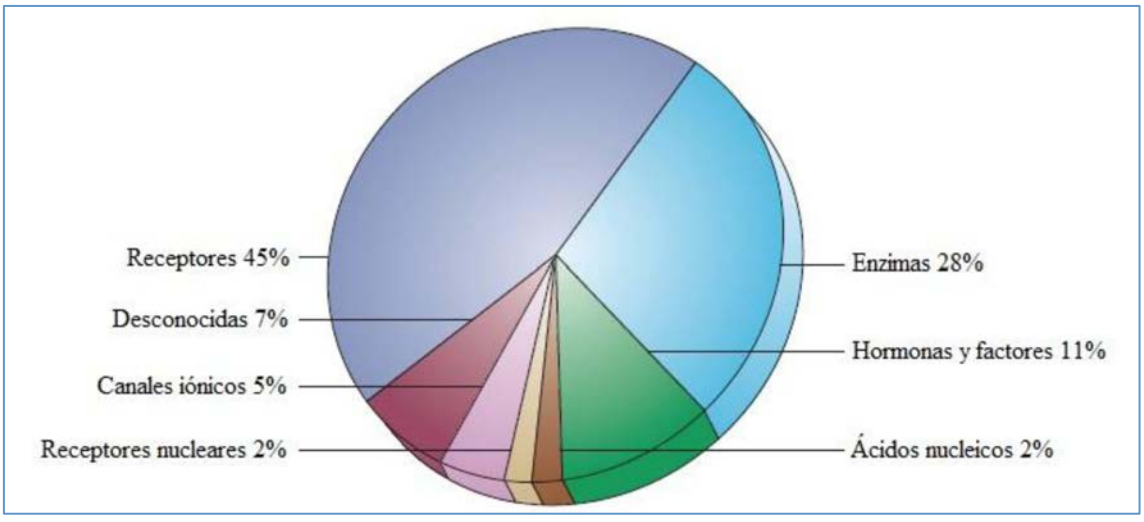


Figura 3.- Dianas farmacológicas distribuidas en siete clases bioquímicas principales, donde las enzimas y receptores representan la parte mayoritaria.

Si se desconoce la estructura del blanco, los métodos de cribado virtual pueden derivarse de un grupo de compuestos con actividad conocida obtenidos de ensayos experimentales previos. Estos métodos se conocen como enfoques de

cribado virtual basados en ligandos, en oposición al enfoque anterior basado en la estructura del blanco. Alternativamente, el conjunto de compuestos activos puede usarse para derivar un *modelo farmacóforo* que puede usarse como un filtro para eliminar aquellos compuestos que no cumplan con las condiciones de actividad necesarias (22).

Una de las herramientas más simples y populares del cribado virtual de conjuntos de datos quimio(bio)informáticos lo constituye la *búsqueda de similitud*, la cual es ampliamente utilizada en las etapas más tempranas de los programas de descubrimiento de líderes. Su función principal es identificar los compuestos activos que más se asemejan a la estructura de referencia que luego pueden servir de base para más estudios detallados de cribado virtual que emplean técnicas más refinadas (23). Dentro de las técnicas más usadas en el *análisis de diversidad* de bibliotecas de cribado y combinatorias se encuentran los algoritmos de agrupamiento, cuya idea esencial consiste en subdividir el conjunto de moléculas en grupos o clústeres de modo que la similitud intra cluster sea máxima mientras que la similitud inter clúster sea mínima; los algoritmos de partición, que consisten en subdividir el rango de valores de un pequeño grupo de características, relevantes a la unión del ligando al receptor y previamente identificadas por el investigador, en sub rangos, cuya combinación genera una malla *n*-dimensional de celdas a las cuales son asignadas las moléculas del repositorio, de modo que los valores de las características estudiadas en las mismas concuerden con aquellos "encerrados" en una celda específica; los algoritmos basados en disimilitud, que a diferencia de los anteriores buscan identificar directamente un subconjunto diverso mediante la selección iterativa de compuestos que son lo más diferentes posible a aquellos que han sido seleccionados previamente; y los algoritmos basados en optimización, que parten de definir una medida de diversidad cuantitativa y entonces la selección del conjunto más diverso posible se formula en términos de los problemas de optimización combinatoria (24).

Algunos estudios previos sugieren que los algoritmos de agrupamiento brindan un mejor balance entre representatividad y diversidad que otras técnicas basadas en disimilitud para el análisis de diversidad (25,26). Un estudio muy reciente que aborda la comparación de varios de los algoritmos de agrupamiento más exitosos en Quimioinformática (i.e., la clase de los algoritmos no superpuestos, jerárquicos, aglomerativos, secuenciales y combinatorios, CSAHN) puede encontrarse en (27). En este contexto, atención especial merecen las técnicas rápidas de tendencia al agrupamiento que permiten obtener una evaluación de la "predisposición" de los datos químicos a ser agrupados antes de ejecutarse la técnica de agrupamiento en sí. La importancia práctica de estas técnicas radica en evitar formarse una idea errónea acerca de la organización de los datos como estructurados en clústeres cuando en realidad provienen de una única población

aleatoria, además que evitan el malgasto de recursos computacionales y de tiempo (28).

Un enfoque más exacto de los métodos de cribado virtual consiste en la construcción de un modelo que correlacione la estructura de las moléculas con sus respectivas actividades biológicas a partir de un grupo de moléculas previamente evaluadas. Este problema se conoce como la modelización de la relación estructura-actividad (REA) más comúnmente conocido por sus siglas en inglés QSAR, acrónimo de Quantitative Structure-Activity Relationship, e involucra métodos de los campos de la Estadística y el Aprendizaje Automático (29, 30). La utilidad práctica de este enfoque se ha constatado, por ejemplo, en estudios de identificación de nuevas entidades anti protozoarias (*Trichomona vaginalis*) mediante la técnica estadística Análisis Discriminante Lineal usando descriptores moleculares definidos por el Prof. Dr. Yovani Marrero Ponce (31-33). Este enfoque también se ha aplicado en el descubrimiento de fármacos frente a la *enfermedad de Chagas* (34). Algunos de los algoritmos más populares en el área de la minería de datos pueden estudiarse en (35).

3. GENERALIDADES DE LA SIMILITUD MOLECULAR

El concepto de similitud ha ganado un espacio cada vez más importante en la quimioinformática debido fundamentalmente al *principio de similitud*, el cual plantea que *moléculas con estructuras similares tienden a exhibir propiedades similares* (36). Este principio parece ser una adaptación de un proceso que, según algunos autores, es el reflejo directo del núcleo del sistema cognitivo humano, el *razonamiento por analogía* (37), y ha sido apoyado por un buen número de resultados experimentales [ver por ejemplo referencia (38)]. Sin embargo, otros hallazgos han sugerido que *eventualmente* moléculas estructuralmente similares exhiben comportamientos disimilares, así como moléculas estructuralmente disimilares exhiben comportamientos similares (39). 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* , i.e., i-) *moléculas estructuralmente similares es muy plausible que tengan actividades similares*, ii-) *moléculas estructuralmente similares es plausible que tengan actividades disimilares*, iii-) *moléculas estructuralmente disimilares es plausible que tengan actividades similares*, iv-) *moléculas estructuralmente disimilares es muy plausible que tengan actividades disimilares* (40). Las hipótesis i-) y iv-) conforman la lógica de base para técnicas como la búsqueda de similitud y los algoritmos de agrupamiento para la selección de compuestos intra clúster (41, 42). La hipótesis ii-) conforma la lógica de base de un grupo de técnicas novedosas para el análisis y visualización de los *acantilados de actividad* y una de sus aplicaciones potenciales es la identificación de pequeños cambios moleculares responsables de un cambio

abrupto en la actividad, que de por sí conlleva un gran interés (43). Por último, la hipótesis iii-) conforma la lógica de base de técnicas basadas en diversidad para la búsqueda de estructuras patrones o "Scaffold Hopping", que se refiere a la capacidad para identificar clases estructurales diferentes de compuestos activos a través del cribado computacional y constituye el criterio de éxito más importante en las aplicaciones de cribado virtual prospectivo (44).

3.1. Técnica de búsqueda de similitud

La búsqueda de similitud es una de las técnicas de cribado virtual más simples (*vide supra*), 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 (45). En la literatura se han reportado varios estudios comparativos entre técnicas de búsqueda de similitud resaltando sus meritos y deficiencias [ver por ejemplo (46)]. Sin embargo, como Sheridan y Kearsley (2002) han señalado, es muy poco probable que un solo mecanismo de búsqueda pueda comportarse consistentemente superior a los demás en todos los problemas (47). 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* (48).

3.2. Componentes de la búsqueda de similitud

La búsqueda de similitud molecular comprende cuatro componentes esenciales: el conjunto de datos estructurales químicos, que cubre cierta región del espacio químico a explorar; las estructuras de referencia o consulta, que contienen la información química de interés a recuperar; la representación matemática de los compuestos químicos, a través de descriptores moleculares; la medida de (di)similitud, que cuantifica el grado y tipo de semejanza entre dos compuestos químicos; y el algoritmo de emparejamiento o "matching", cuya función es buscar y recuperar los compuestos más parecidos a la molécula de referencia (49).

3.2.1. Conjuntos de datos químicos

El desempeño de los índices de similitud, descriptores moleculares e, incluso, enfoques de validación, es altamente dependiente de las bases de datos de entrenamiento y prueba. Actualmente existe un número considerable de conjuntos de datos estructurales para la evaluación práctica de las técnicas de cribado virtual, de entre los más populares se encuentran: la mega base de datos del proyecto PubChem, disponible gratuitamente (50); la base de datos de los cribados anti-VIH y anti cancerígeno del Instituto Nacional del Cáncer (NCI, del inglés National Cáncer Institute), disponible gratuitamente (51); los repositorios de datos de la Sociedad de Quimioinformática y QSAR, disponibles gratuitamente (52); los

conjuntos de datos de la Academia Internacional de Química Matemática, disponibles gratuitamente (53); la base de datos MDDR (MDL Drug Data Report), comercial; la base de datos WDI (World Drug Index), comercial (54); y la base de datos WOMBAT (World of Molecular Bioactivity Data), comercial (55). La tendencia actual de las bases de datos quimioinformáticas es pasar al dominio público (56, 57).

Especial atención merecen los conjuntos de datos para propósitos de comparación de nuevas herramientas de cribado. En la literatura se recomienda el uso de los conjuntos de datos MUV diseñadas por Rohrer et al. (2009). Estos conjuntos de datos de compuestos activos y señuelos de activos "*decoys*" (inactivos confirmados) fueron construidos usando herramientas estadísticas de diseño experimental basadas en la técnica del *análisis refinado de los vecinos más cercanos* y están orientadas a minimizar problemas encontrados con el uso de las métricas de desempeño (*vide infra*) en otros conjuntos de validación como el *enriquecimiento artificial*, donde la clasificación es causada por diferencias en propiedades simples y usualmente irrelevantes entre activos y decoys; el *sesgo de análogos*, causada por la tendencia de los conjuntos de datos a sobre representar las clases de activos y deriva en una clasificación sobreestimada de los mismos. Estos dos problemas se tienden a englobar en el problema denominado *sesgo de conjuntos de datos de referencia*. El último problema de este tipo se refiere a la *varianza de los resultados de validación*, causada por usar conjuntos indebidamente desbalanceados que conducen al *efecto de saturación* de las curvas ROC correspondientes (58). En los últimos años, algunos autores han alertado acerca de otro tipo de problemas más sutiles que concierne la calidad de conjuntos de datos altamente referenciados como son los errores estructurales, presencia de compuestos duplicados, errores de correspondencia de los datos estructurales con las mediciones experimentales, falta de reproducibilidad en las mediciones experimentales, etc. Los hallazgos sugieren que el tener estructuras erróneas representadas por descriptores erróneos deriva en un efecto perjudicial para el desempeño y la fiabilidad de las predicciones de los modelos de cribado. Para solucionar estos problemas los investigadores proponen se utilicen un buen número de potentes herramientas de software libre así como una última etapa de inspección "manual" (59).

Hasta el momento, la comunidad científica internacional no ha adoptado ningún conjunto de datos estándar para la comparación de medidas de similitud, probablemente por la imposibilidad de encontrar un grupo único de moléculas que reagrupe todas las necesidades de cribado de la Quimioinformática moderna (39). 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 (47).

3.2.2. 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 un conjunto de relaciones asociadas (similitudes, disimilitudes, distancias) entre las moléculas, lo cual le da al espacio una "estructura" (60).

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 descriptores moleculares (DMs) y, por tanto, tiene una posición absoluta en un espacio multidimensional. La dimensión de este espacio se especifica por el número de rasgos no correlacionados (descriptores de complejidad, descriptores de solubilidad, huellas dactilares o "fingerprints", tripletes de farmacóforos, u otro vector de descriptores). 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 (descriptores de pares de átomos, árboles de rasgos, enfoques de Subestructura Máxima Común) (61-63).

Cuatro tipos de objetos matemáticos se utilizan normalmente para representar las moléculas, estos son: conjuntos, grafos, vectores y funciones. Los conjuntos son los objetos más generales y, básicamente, la base de los otros tres. Normalmente, los químicos representan moléculas como "grafos químicos" (64), que están estrechamente relacionados con los tipos de grafos tratados por los matemáticos en el campo de la teoría de grafos (65).

Los grafos químicos proporcionan una metáfora potente e intuitiva para la comprensión de muchos aspectos de la química, pero sin embargo tienen sus limitaciones, especialmente cuando se trata de cuestiones de interés en la quimiometría y quimioinformática.

En estos campos de información molecular se representan normalmente los vectores de características, donde cada componente corresponde a una función local o global característica de una molécula. Las características locales incluyen fragmentos moleculares (subestructuras), farmacóforos (66), varios índices topológicos (67), y cargas atómicas parciales, entre otras. Las características globales incluyen características tales como el peso molecular, logP, la superficie polar, varios BCUTs y el volumen molecular (49).

Más recientemente, con el aumento significativo de la potencia de los ordenadores, incluso en PCs de escritorio, los métodos para identificar directamente los rasgos de las moléculas 3D se han vuelto más frecuentes. Las características aquí se refieren generalmente a diversos tipos de campos moleculares, algunos, como la densidad electrónica ("estérica"), otros como los campos potenciales eléctricos (26) y también como campos potenciales lipofílicos (68). Los campos moleculares son generalmente representados como funciones continuas. Los campos discretos también se han utilizado aunque algo menos frecuente (69).

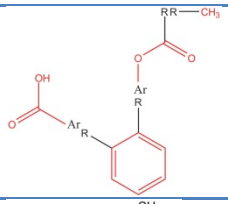
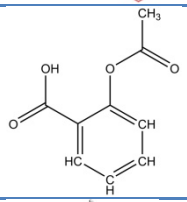
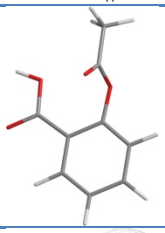
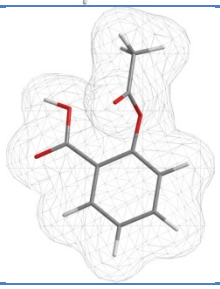
De acuerdo a la naturaleza en su definición y a la complejidad de los rasgos moleculares estructurales que se codifican, los DMs se clasifican de forma general según las dimensiones que abarcan en: DMs-0D (Descriptores Constitucionales), DMs-1D (Descriptores Unidimensionales), DMs-2D (Descriptores Bidimensionales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

Los DMs-0D son descriptores que se obtienen directamente de la fórmula molecular y son independientes de cualquier conocimiento sobre la estructura molecular, por ejemplo, el número de átomos (A), el peso molecular (MW), conteo de átomos-tipo (Nx) o cualquier función de las propiedades atómicas. Los DMs-1D están basados en la representación unidimensional de la molécula (o representación que consiste en una lista de fragmentos estructurales de la molécula), aunque no requieren del conocimiento completo de la estructura molecular, tal es el caso de los descriptores de búsqueda y análisis subestructural, como los Descriptores de Conteo de Fragmentos.

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 molécula.

Finalmente los DMs-4D 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, tales como los Valores de la Energía de Interacción (39).

Tabla 1.- Relación entre la dimensionalidad de los descriptores y la complejidad de la representación que describen.

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

Otra clasificación de los DMs que aunque no se menciona explícitamente en los textos quimioinformáticos tiene una importancia trascendental a la hora de aplicar la modelización estadística y/o de aprendizaje automático es la de acorde a la naturaleza numérica de definición de los mismos, esto es, en continuos y discretos. Por ejemplo, la mayoría de los descriptores implementados en el software DRAGON son continuos, las principales excepciones son los bloques *constitucionales*, donde se pueden encontrar varios descriptores con valores discretos, y todos resultan ser los "contadores" (number of atoms, etc); los bloques de *grupos funcionales* y *fragmentos centrados en átomos*, todos son contadores y por ende tienen valores discretos; algunos descriptores de *propiedades moleculares*, los descriptores tipo-fármacos que comienzan en LAI y terminan en Infective-50 son binarios o *booleanos* (1/0); las *huellas dactilares binarias 2D* todas son binarias (1/0); las *huellas dactilares de frecuencia 2D*, todos tienen valores discretos.

Desde el punto de vista estadístico, de acorde a la fortaleza de la medición de las variables o DMs, estos pueden clasificarse en las escalas de proporción, intervalo, ordinal y categórica (el caso binario para dos categorías). Una práctica común en quimioinformática consiste en transformar descriptores continuos y discretos (proporción/intervalo) en binarios (categórica) a través de un valor de corte como la mediana, o simplemente trabajar con huellas dactilares, para aumentar la eficiencia de los algoritmos de clasificación/predicción; sin embargo, esta práctica también conduce a una pérdida de información estadística que se traduce en la aparición de ataduras en los valores de similitud y disminución de la potencia de las técnicas, resultando además en una menor versatilidad de las mismas (70).

La presentación que se muestra en la Tabla 1 está lejos de ser representativa, por lo que para una presentación detallada los lectores interesados pueden referirse a la última versión del manual de descriptores moleculares de Todeschini y Consonni (2009) donde se trata este tema con profundidad (71). El número de descriptores moleculares propuestos en la literatura hasta el momento es realmente amplio, para ello recientemente se han desarrollado sistemas para el cálculo de grandes conjuntos de descriptores algunos de ellos son el software DRAGON, comercial (72); PaDEL, disponible gratuitamente (73); y MODEL, en plataforma web y disponible gratuitamente (74). Una lista más ampliada de programas para este fin puede encontrarse en el sitio web de la ref. (75).

3.2.3. Selección de rasgos

Actualmente, existe un número realmente grande de descriptores desarrollados que pueden ser usados en los cálculos de similitud (76). Sin embargo, a medida que la dimensionalidad de los datos incrementa, muchos tipos de análisis de datos y problemas de clasificación se vuelven computacionalmente difíciles. En ocasiones, también los datos se vuelven crecientemente dispersos en el espacio que ocupan. Esto puede conducir a grandes problemas para ambos, para el aprendizaje supervisado y no supervisado. En la literatura este fenómeno se refiere como *la maldición de la dimensionalidad* (77). 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 (78). 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 (79), el caso extremo de este fenómeno se ilustra en *el teorema del patito feo* de Watanabe; basicamente, si uno considera el universo de rasgos de los objetos y no tiene algún sesgo

cognitivo acerca de cuales de ellos son mejores, no importa cuales dos objetos uno compare, todo resultará igualmente similar (disimilar) (80). En este sentido, algunos investigadores de la química medicinal han planteado que no tiene sentido hablar de diversidad sin un sistema de referencia, que está dado en este caso por el ensayo biológico (81). Una estrategia para solucionar esta dificultad es seleccionar un conjunto de descriptores en particular para los cuales se demostró que funcionan bien en un cierto problema. 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. Un enfoque diferente es dejar que la computadora escoja la combinación óptima de descriptores para el problema en cuestión (82).

Numerosos métodos automáticos han sido propuestos en quimioinformática para la selección de rasgos, por ejemplo, la técnica paso a paso de los procesos de integración hacia adelante o eliminación hacia atrás y el análisis de componentes principales (83); también ha sido propuesto el uso de los *k*-vecinos más cercanos (84). Otros métodos de selección más usados en la modelación REA se encuentran la selección secuencial hacia delante (Sequential Feature Forward Selection), la eliminación secuencial hacia atrás (Sequential Feature Backward Elimination), el recocido simulado (Simulated Annaeling) y la selección basada en algoritmos genéticos, siendo esta última una de las más eficientes en el campo de modelación REA (85).

En el pasado, algunos enfoques estaban directamente relacionados con las Redes Neuronales Artificiales, como son: división de los pesos (86), correlación en cascada (87), mapas de Kohonen (88), determinación de la relevancia automática (89), etc. También han sido presentados en la literatura especializada los Sistemas Artificiales de Colonias de Hormigas y Enjambres (90). También ha sido evaluada la eficiencia de algunos algoritmos de poda (91).

En resumen, existe una amplia variedad de descriptores moleculares y métricas usadas en los métodos de similitud molecular; parece ser, sin embargo, que el mejor rendimiento se logra adaptando dicha combinación al problema estudiado (92).

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 (93). Un buen número de estas técnicas aparecen implementadas en el software de aprendizaje automático y minería de datos Weka (94), que también puede usarse para la modelización QSAR. Este producto es uno de los más populares en el área del Aprendizaje Automático, es de código abierto y se encuentra disponible gratuitamente (95).

3.2.4. Medidas de similitud

El concepto de similitud es fundamental para varios aspectos del razonamiento y análisis químicos, de hecho, es tal vez la premisa fundamental de la química médica, y cae bajo la rúbrica general de análisis de similitud molecular. La determinación de la similitud de un "objeto molecular" con otro es básicamente un ejercicio de comparación de patrones químicos.

El resultado de este ejercicio es un valor, *la medida de similitud*, que caracteriza el grado de concordancia, de asociación, proximidad, semejanza, alineamiento, porcentaje de identidad o similitud entre pares de moléculas manifestada por sus “patrones moleculares”, que se componen de conjuntos de rasgos.

La terminología de “proximidad” a veces se utiliza en un sentido más general para referirse a la similitud, disimilitud, o la distancia entre los pares de moléculas. Las medidas de similitud son funciones que hacen corresponder pares de representaciones moleculares de la misma forma matemática con números reales que usualmente, pero no siempre, yacen en el intervalo unitario [0,1] (61). La similitud es generalmente considerada como una propiedad simétrica, es decir, “A” es tan similar a “B” como “B” a “A”, y la mayoría de los estudios se basan en esta propiedad. Tversky (96), sin embargo, ha argumentado persuasivamente que ciertas similitudes son inherentemente asimétricas.

Aunque su trabajo se orientó hacia la psicología, este tiene aplicabilidad además en los estudios de similitud molecular (97). Por otra parte, cuando se aplican los conceptos de similitud y diversidad en química, es necesario definir similitudes globales y locales; las similitudes locales se centran en parte en un objeto (átomo, grupo funcional, las cadenas de proteínas, cadena de ADN, etc.), mientras que las similitudes globales la semejanza se mide entre dos objetos enteros (moléculas, proteínas, etc.) (98).

Consideremos dos objetos químicos arbitrarios A y B descritos mediante vectores $\boldsymbol{X}$ e $\boldsymbol{Y}$, respectivamente, de n atributos, de modo que $\boldsymbol{X} = (x_1, x_2, x_3, \dots, x_n)$.e $\boldsymbol{Y} = (y_1, y_2, y_3, \dots, y_n)$. En la Tabla 2 se muestra un grupo de medidas de (di)similitud de amplio uso en quimioinformática extraídas de la revisión de Ellis et al (99). Otro trabajo de revisión excelente sobre medidas de similitud puede encontrarse en (100).

Tabla 2.- Algunas de las medidas de proximidad más usadas en la búsqueda de similitud

Medida	Fórmula[a]	Tipo[b]
Manhattan Media	$MM_{XY} = \frac{\sum_{j=1}^{n}\lvert x_j - y_j\rvert}{n}$	D
Euclidiana Media	$EM_{XY} = \frac{\sqrt{\sum_{j=1}^{n}\lvert x_j - y_j\rvert^2}}{n}$	D
Bray/Curtis	$BC_{XY} = \frac{\sum_{j=1}^{n}\lvert x_j - y_j\rvert}{\sum_{j=1}^{n}(\lvert x_j\rvert + \lvert y_j\rvert)}$	D
Tan	$T_{XY} = \frac{\sum_{j=1}^{n} x_j y_j}{\sum_{j=1}^{n} x_j^2 + \sum_{j=1}^{n} y_j^2 - \sum_{j=1}^{n} x_j y_j}$	A
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
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
Cosine/Ochiai	$Cos_{XY} = \frac{\sum_{j=1}^{n} x_j y_j}{\sqrt{\sum_{j=1}^{n} x_j^2 \sum_{j=1}^{n} y_j^2}}$	A
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

[a]xj (y_j) representa el valor del descriptor del vector ***X*** (A) e ***Y*** (B) en el atributo *j*; [b]Clasificación de las medidas de proximidad acorde a su naturaleza de definición. D, coeficientes de distancia: están basados en la suma de diferencias, sus valores varían en proporción inversa con el grado de similitud; A, coeficientes de asociación: se basan en el producto interno, y sus valores varían en proporción directa con el grado de similitud, por lo que una mayor similitud se indica por el aumento de los valores; C, coeficientes de correlación: Los coeficientes de correlación se basan en una tercera función más compleja: la suma de los productos de la diferencias entre cada valor-atributo y la media de todos los valores de los atributos de cada uno de los dos vectores. Los valores de estos por lo general varían de 1 (lo que indica que cualquier cambio en los atributos de un objeto sería acompañado por un cambio idéntico en los atributos del otro) a -1 (que indica que un cambio en uno y sería acompañado por un cambio igual y opuesto en el otro).

Cuando los valores de atributo se limitan a 0 y 1, las expresiones utilizadas por varias similitudes y medidas de distancia pueden a menudo ser simplificadas considerablemente. Si los objetos A y B que se caracterizan por vectores $\boldsymbol{X}$ e $\boldsymbol{Y}$ que contienen n valores binarios (tales como huellas digitales) 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)

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

(2)

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

(3)

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

(4)

Por tanto, $n = a + b - c + d$

(5)

Estas cantidades anteriores también se pueden expresar en notación de teoría de conjuntos dando lugar a otras formulaciones basadas en este tipo de representación (101).

Como ejemplo ilustrativo tenemos el coeficiente de Tanimoto para el caso binario dado por:

$T_{XY} = c/[a + b - c]$

(6)

Este coeficiente aplicado a las huellas dactilares 2D constituye actualmente la medida de elección de los sistemas de software comerciales para la gestión de la información química. También forma parte de sistemas de acceso público importantes como el PubChem (50).

En un artículo revisión reciente Willet (2006) resume los resultados de los estudios de comparación y combinación de coeficientes de similitud usando huellas dactilares en conjuntos de datos apropiados. Estos resultados muestran que algunos coeficientes se comportan monotónicamente entre sí, lo que significa que producen clasificaciones u ordenamientos idénticos o muy similares de los compuestos de la base de datos frente a una estructura de referencia determinada, a pesar de que los valores del coeficiente real son diferentes. También se ha mostrado que algunos coeficientes tienen una marcada preferencia a funcionar bien en la búsqueda de moléculas activas de un tamaño determinado dado aproximadamente por el número de bits activos en el vector de representación; por ejemplo, el coeficiente de Russel-Rao “muestra preferencia” por moléculas

bioactivas de tamaño relativamente grande, el coeficiente de Tanimoto por moléculas bioactivas de tamaño mediano y el coeficiente de Forbes por moléculas bioactivas de tamaño pequeño (102).

Aún cuando el coeficiente de Tanimoto continua siendo la medida de similitud estándar en la industria y se ha usado en innumerables trabajos de investigación, la evidencia indica que ningún modelo de proximidad es universalmente superior a los demás, sino que su utilidad práctica depende del problema o grupo de problemas a tratado (92). Esta conclusión parece estar de acorde a la dialéctica resultante de la complementación de los teoremas *Ningún Almuerzo es Gratis* (NFL, del inglés No Free Lunch) (103, 104), y la *Longitud Mínima de la Descripción* [MDL del inglés Minimun Description Length] (105), correspondientemente.

3.3. Algoritmos de emparejamiento o "matching"

El concepto de emparejamiento "match" exacto y parcial, y los algoritmos de búsqueda de emparejamiento son ampliamente utilizados en sistemas de información química basados en ordenadores con el fin 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 recupera la estructura(s) más similar a una estructura de consulta, donde la similitud se define sobre la base de alguna función de coeficiente de similitud o de distancia que refleja el número de fragmentos comunes de la consulta y de una molécula en el fichero. La búsqueda del mejor par es la base para la clasificación del *k-(ésimo) vecino más cercano* (kNN, del inglés *k*-Nearest Neighbor) y juega un papel importante en el uso de árboles de expansión y técnicas de clasificación automática (106).

El problema general de encontrar las mejores pares se define por Friedman et al. (107) como: "... dado un fichero de *m* instancias (cada uno de los cuales es descrito por *n* atributos con valores reales) y una medida de similitud/disimilitud, encontrar las *k* instancias más cercanas a la instancia de consulta (es posible que no esté dentro del fichero) con los atributos especificados". Es obvio que el algoritmo de fuerza bruta para la búsqueda del mejor par es calcular la distancia entre la consulta y cada uno de las instancias del fichero y luego elegir las *m* distancias más cortas, este algoritmo tiene una complejidad temporal *O(mn)* para el caso de una consulta simple, pero en el caso de consulta múltiple sería un *O(mnc),* siendo *c* el número de consultas con igual cantidad de atributos *n*, el cual consume demasiado tiempo para ficheros considerablemente grandes.

Un algoritmo eficiente del vecino más cercano será uno que evite el cálculo de la mayoría de las distancias, calculando solamente las distancias de las escasas instancias que rodean la instancia o estructura de consulta. Existen varios tipos de

criterios para reducir el número de cálculos necesarios, incluyendo la proyección de las instancias *d*-dimensionales en un espacio de dimensión menor, de forma tal que varias instancias puedan ser buscadas, o eliminadas desde una búsqueda, simultáneamente (108). En este sentido, varios de los algoritmos reportados pueden no ser directamente aplicables a la búsqueda de los mejores pares en contextos químicos ya que los primeros asumen que los atributos son variables continuas, mientras que las estructuras químicas son descritas frecuentemente por fragmentos de ocurrencia binaria. En estos casos, cada una de las estructuras en un archivo se representa por una cadena de bits en el que se establece el bit *i-ésimo* si el fragmento correspondiente está presente en la estructura. Además, a menudo se supone que las instancias se encuentran en un espacio de dimensión *d* pequeña, por lo general 2 o 3; sin embargo, para el caso de la representación química binaria, *d* puede ser del orden de 10^2 o 10^3 (el número de bits en la cadena de bits), y por ende estos algoritmos resultan ser poco factibles. Por ejemplo, el procedimiento *O(nlog N)* debido a Friedman et al. (1977) implica una constante de proporcionalidad alrededor de 1.6^d (107), mientras que el método de búsqueda de Bentley et a1. (1980) implica la inspección de todas las 3^d - 1 celdas adyacentes a una celda dada en un espacio *d*-dimensional (108).

Alternativamente, otros investigadores han centrado su atención en los algoritmos de búsqueda basados en la representación binaria. Smeaton y Van Rijsbergen (1981) tienen en cuenta que un *archivo invertido* puede ser utilizado para aumentar la eficiencia de la búsqueda de emparejamiento a una consulta en documentos donde tiene al menos un término en común. A partir de aquí, estos autores describen experimentos usando un procedimiento de *límite superior* que permite que la búsqueda de la mejor pareja se termine antes de que todos los documentos en la lista de los ficheros invertidos correspondientes a la consulta hayan sido inspeccionados (109). Murtagh (1982) describe una extensión de este algoritmo en el que son calculados otros límites superiores, posibilitando una mayor reducción en el número de documentos que necesitan ser comparados con una consulta (110).

Van Marlen y Van den Hende (1979), y Rasmussen et al. (1979) han descrito algoritmos de recuperación de las mejores parejas para el uso de ficheros informáticos con espectros de masa, donde la estructura es caracterizada por una cadena de bits correspondientes a los picos observados en el espectro de masa molecular (111, 112), mientras que otros autores han estudiado la búsqueda del mejor emparejamiento en los sistemas de recuperación de información molecular (106).

Baldi et al. (2008) plantean un algoritmo diferente a los demás, el cual 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 siendo n potencia de 2 (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 XOR (eXclusive OR, del inglés) al vector $\vec{A}$. Esta información adicional puede ser vista como una guía que precede al vector $\vec{A}$, la cual puede ser usada para derivar los límites útiles en las medidas de similitud lo cual permite explorar menos del 50% de la base de datos y acelera la búsqueda significativamente (113). Más recientemente, Cao et al. (2010) han reportado un algoritmo de búsqueda y agrupamiento acelerado basado en técnicas de empotramiento 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 et al. (2008) 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 (114).

3.4. Fusión de datos

La fusión de datos se utilizó por primera vez en la búsqueda de similitud a finales de los años noventa (115,116). 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 (48). Actualmente, las nuevas técnicas de búsqueda de similitud son validadas usando la técnica fusión de datos *embebida* en algún mecanismo de validación cruzada. Para ello, una vez obtenidas las listas de recuperación como producto de aplicar las *multi consultas*, es necesario combinar dicha información para derivar un puntaje fusionado y útil para cada molécula del repositorio que permita el ordenamiento final del conjunto de datos. En este sentido Hert et al. (2004) introdujeron la regla de fusión MAX-SIM (máxima similitud) que por su probada alta efectividad se ha usado durante varios años como el multi clasificador *de facto* para los estudios quimioinformáticos por su eficacia y simplicidad matemática y computacional en el cribado de conjunto de datos farmacológicos (117, 118). Básicamente, el algoritmo MAX-SIM es uno de los

métodos más simples para el cribado virtual por el cual una molécula es punteada con su similitud más alta a una molécula activa de la multi consulta. Formalmente, si una consulta múltiple de activos es denotada por $\{x_1, x_2, \dots, x_q\}$, el puntaje asignado a una molécula del conjunto de datos x_n viene dado por:

$$z_1(x_n) = max|_1^q\{S(x_n, x_i)\} \quad (7)$$

Donde, $S(x_n, x_i)$ es la similitud de la molécula del conjunto de datos x_n a la referencia x_i de la multi consulta, S es la función de similitud y algunas de ellos han demostrado ser eficaces en la operación. Sin embargo, en un estudio abarcador Chen et al. (2010) mostraron recientemente que la regla "suma de rangos inversos" se comporta superiormente a la regla MAX-SIM en los dominios de datos examinados, esto es:

$$z_2(x_n) = \sum_{i=1}^{q} 1/r[S(x_n, x_i)] \quad (8)$$

Donde, r es el "ranking" asignado al puntaje de similitud $S(x_n, x_i)$, relativo a los puntajes de las moléculas del conjunto con respecto a una consulta especifica.

Esta regla de fusión procede del área de Recuperación de Información y su efectividad se debe a la cercana relación que existe entre el rango reciproco de la estructura de la base de datos con respecto a una búsqueda de similitud simple y la probabilidad de que esta estructura comparta la misma actividad que la estructura de referencia (119).

Como alternativa 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. Algunas técnicas de ponderación de rasgos se basan en el análisis de frecuencia de bits en huellas dactilares o "fingerprints" de molecular activas y/o inactivas, perfilando, escalando y promediando los *fingerprints* para derivar en el cálculo de los *fingerprints de consenso.* Un grupo de técnicas más reciente se basan en el *acallado de bits* "bit silencing" y difiere de los enfoques estadísticos en que monitorean directamente el cambio en la calidad de la recuperación cuando se omiten *bits* individuales en moléculas de referencia activas (120). 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 por la facilidad con que pueden calcularse, manipularse y almacenarse los descriptores binarios. Por otra parte, estas técnicas también pueden ser extendidas al caso no binario.

3.5. 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 cribado virtual retrospectivos. Una métrica popular es el "factor de enriquecimiento", que es intuitivo y sencillo de interpretar. Un problema asociado con el cálculo de los factores de enriquecimiento simples es la dependencia de un valor de corte elegido, por lo general el 1 o 5% de la base de datos para cribado. Nicholls (2008) aboga firmemente por el uso de medidas estándares, incluyendo la curva de la Característica en Operación del Receptor (ROC, del inglés Receiver Operating Characteristics) y el área bajo la curva AUC[ROC] (121), que se aplican habitualmente en otros campos que emplean el análisis estadístico, minería de datos, o las técnicas de aprendizaje automático (122). Sin embargo, Truchon y Bayly (2007) detectaron que la curva ROC no tiene en cuenta explícitamente el llamado "problema de la detección temprana", i.e., la propiedad de un método para recuperar compuestos activos "tempranamente", i.e., al principio de la lista de clasificación. Específicamente, este fenómeno es ejemplificado 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 aunque, si solo algunos pocos primeros hits pueden ser probados experimentalmente, el caso 1-) es claramente mejor que el caso 2-) que, a su vez, es mejor que el caso 3-). En este sentido, los autores desarrollaron un mejoramiento de la curva ROC a través de la métrica Discriminación Mejorada por (la distribución de) Boltzmann de la ROC (BEDROC, del inglés Boltzmann-Enhanced Discrimination of ROC), que utiliza una ponderación exponencial para asignar mayor peso a la detección temprana (123). Esta medida es esencialmente una versión normalizada de la medida Mejora Inicial Robusta (RIE, del inglés Robust Initial Enhancement) (124). Del mismo modo, se ha sugerido el escalado semilogarítmico de la ROC, pROC (125). Sin embargo, Nicholls (2008) también presenta evidencias de una fuerte correlación entre el AUC[ROC] y AUC[BEDROC], lo que sugiere a AUC[ROC] como una medida suficiente para evaluar la eficiencia de cribado virtual. Este mismo autor recomienda se aplique un ponderado exponencial a la curva ROC preferentemente a los rangos individuales de los compuestos activos dentro de los inactivos para mejorar algunas de las deficiencias de las métricas AUC[RIE] y AUC[BEDROC] (121).

3.5.1. Curva ROC concentrada

Basados en la idea de Nicholls (2008), aunque no lo citan explícitamente, Swamidass et al. (2010) proponen 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*) e *Y* representa la razón de verdaderos positivos (*tpr*)] a través de una transformación de magnificación suave ya sea exponencial, de potencia o logarítmica. La lógica de su trabajo se basa en el "comportamiento del usuario" que se observa en la recuperación de páginas web donde se conoce, como promedio, la frecuencia con que el primero, segundo, …, *n-ésimo* registro son pinchados ("cliqueados"); la curva decreciente correspondiente de cuán relevante es cada rango provee información valiosa para los niveles de intervalo y magnificación requeridos; a partir de aquí es razonable requerir que el factor de magnificación local sea proporcional a la relevancia correspondiente. Por la analogía de estos sistemas con los sistemas de recuperación en el descubrimiento de fármacos, se propone se emplee una relevancia exponencialmente decreciente del "ranqueo" final. Finalmente, a través de 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 Curva de Acumulación Concentrada (CAC, del inglés Concentrated Acumulation Curve), pROC y ROC (126).

La variante más potente de la curva CROC se obtiene aplicando una transformación de magnificación exponencial del eje *X* (*fpr*) de la curva ROC dada por:

$$h(x) = \frac{1-e^{-\alpha x}}{1-e^{-\alpha}}$$

(**9**)

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

Una vez establecida la función de magnificación $h(x)$, el área bajo la curva CROC puede calcularse fácilmente como el promedio de los valores de *fpr* transformados correspondientes a las posiciones de las instancias positivas en la lista de recuperación como:

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

(**10**)

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

Por último, valores del área bajo CROC se pueden comparar con el valor correspondiente al clasificador aleatorio a través de la formula:

$$AUC[CROC]_{aleat} = \frac{1}{\alpha} - \frac{e^{-\alpha}}{1-e^{-\alpha}}$$

(**11**)

Donde, α = 20 la métrica del clasificador aleatorio toma el valor $AUC[CROC]_{aleat} = 0.2809$

4. CONCLUSIONES

El proceso tradicional de descubrimiento y desarrollo de nuevos fármacos es muy costoso en términos de recursos materiales y de tiempo. Una alternativa viable y complementaria a este paradigma es el método de cribado virtual *in silico*, cuya esencia radica en manipular de forma racional en términos explicativos, de diseño y predictivos el gran volumen de información procedente del cribado de alto rendimiento y quimiotecas virtuales. Una de las técnicas que resaltan por su alta eficiencia y comprobada efectividad es la búsqueda de similitud, que contando solamente con un ordenador potente, un conjunto de datos químicos *in silico*, una medida de similitud, un algoritmo de emparejamiento e información acerca de una única molécula bioactiva de consulta, o al menos unos pocos rasgos estructurales de interés, es capaz de recuperar las moléculas más parecidas a la referencia, que a su vez tienen la mayor probabilidad de exhibir la bioactividad estudiada. El panorama actual brinda una magnífica oportunidad para el uso y explotación de estas técnicas en la solución de problemas de la química medicinal ya que, al igual que en el caso de la bioinformática, los recursos quimioinformáticos siguen pasando aceleradamente al dominio público. A pesar de ello, se debe seguir velando por la rigurosidad y calidad de los modelos y soluciones puesto que los productos finales serán usados en humanos, demás animales, plantas y el medio ambiente en general. Para validar y usar nuevas técnicas (nuevos descriptores, medidas de similitud, algoritmos de búsqueda) recomendamos usar conjuntos de datos no sesgados, curados y representativos del contexto bioactivo a investigar; usar representaciones moleculares eficientes pero informativas; usar técnicas de selección de rasgos (automática) cada vez que sea posible y usar estos rasgos seleccionados en las búsquedas "no supervisadas" de otros repositorios grandes; emplear validación cruzada, cuando sea apropiado, para obtener un estimado promedio del desempeño en las distintas regiones del espacio de entrenamiento, y finalmente, comprobar la calidad de las predicciones a través de evaluaciones experimentales de la actividad *in vitro* e *in vivo*. Esperamos que en los años venideros, con la disponibilidad de mayores recursos virtuales gratuitos, un mayor grado en la comprensión del enigma encantador de la similitud molecular y los mapas de similitud, y contando con algoritmos de búsqueda eficientes y ordenadores veloces, seremos capaces de adentrarnos cada vez más en el "espacio astronómico químico" descubriendo otras "galaxias de compuestos líderes" y aportando soluciones eficaces en términos de entidades farmacológicas noveles en favor de una mayor calidad de vida y longevidad del ser humano.

5. AGRADECIMIENTOS

El primer autor (O.M.R.B.) quisiera agradecer a sus colegas y amigos Noel Ferro, de la Universidad de Hannover (Alemania); Nelaine Mora-Diez, de la Universidad Thomson Rivers (Canadá) y Lourdes Casas-Cardoso, de la Universidad de Cádiz (España) por proveerle gentilmente con materiales bibliográficos útiles. También, quisiera reconocer el trabajo altamente eficiente del consejo editorial científico de la revista Anales de la Real Academia Nacional de Farmacia. Esta investigación fue financiada parcialmente por el Programa de Colaboración entre la UCLV y la institución belga VLIR-IUS. El programa de becas entre la Universidad Autónoma de Madrid y la UCLV también financió parte de esta investigación.

6. REFERENCIAS

1. Drews, J. Drug discovery: A historical perspective. *Science* **2000**, *287*, 1960.
2. Kubinyi, H. Strategies and recent technologies in drug discovery. *Pharmazie* **1995**, *50*, 647.
3. Chanda, S.; & Caldwell, J. Fulfilling the promise: Drug discovery in the postgenomic era. *Drug Discov Today* **2003**, *8*, 168.
4. Ren, J.; & Stammers, D. HIV reverse transcriptase structures: Designing new inhibitors and understanding mechanisms of drug resistance. *Trends Pharmacol Sci* **2005**, *26*, 4.
5. Manly, C.; Louise-May, S.; & Hammer, J. The impact of informatics and computational chemistry on synthesis and screening. *Drug Discov Today* **2001**, *6*, 1101.
6. Jorgensen, W. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813.
7. Xu, J.; & Hagler, A. Chemoinformatics and drug discovery. *Molecules* **2002**, *7*, 566.
8. Boobis, A.; Gundert-Remy, U.; Kremers, P.; Macheras, P.; & Pelkonen, O. In silico prediction of ADME and pharmacokinetics. Report of an expert meeting organised by COST B15. *Eur J Pharm Sci* **2002**, *17*, 183.
9. Ekins, S.; Boulanger, B.; Swaan, P.; & Hupcey, M. Towards a new age of virtual ADME/TOX and multidimensional drug discovery. *J Comput Aided Mol Des* **2002**, *16*, 381.
10. Bleicher, K.; Bohm, H.; Muller, K.; & Alanine, A. Hit and lead generation: Beyond high-throughput screening. *Nat Rev Drug Discov* **2003**, *2*, 369.
11. DiMasi, J.; Hansen, R.; & Grabowski, H. The price of innovation: New estimates of drug development costs. *J Health Econ* **2003**, *22*, 151.
12. Cruz-Monteagudo, M.; Borges, F.; & Cordeiro, M. N. D. S. Jointly handling potency and toxicity of antimicrobial peptidomimetics by simple rules from desirability theory and chemoinformatics. *J Chem Inf Model* **2011**, *51*, 3060.
13. Tollman, P.; Guy, P.; Altshuler, J.; Flanagan, A.; & Steiner, M. *Revolution in R&D, How Genomics and Genetics are Transforming the Biopharmaceutical Industry*; Group, B. C.; Massachusetts, 2001.
14. Bajorath, J. Integration of virtual and high-throughput screening. *Nat Rev Drug Discov* **2002**, *1*, 882.
15. Lazo, J.; & Wipf, P. Combinatorial chemistry and contemporary pharmacology. *J Pharmacol Exp Ther* **2000**, *293*, 705.
16. Chen, W. L. Chemoinformatics: past, present, and future. *J Chem Inf Model* **2006**, *46*, 2230.
17. Gasteiger, J. Chemoinformatics: a new field with a long tradition. *Anal Bioanal Chem* **2006**, *384*, 57.
18. Warr, W. A. Some trends in chem (o) informatics. *Methods Mol Biol* **2011**, *672*, 1.
19. Reddy, A. S.; Pati, S. P.; Kumar, P. P.; Pradeep, H.; & Sastry, G. N. Virtual screening in drug discovery-A computational perspective. *Curr Protein Pept Sc* **2007**, *8*, 329.
20. Seifert, M. H. J.; Wolf, K.; & Vitt, D. Virtual high-throughput in silico screening. *Biosilico* **2003**, *1*, 143.
21. Bajorath, J. Integration of virtual and high-throughput screening. *Nat Rev Drug Discovery* **2002**, *1*, 882.

22. Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; & Agrafiotis, D. K. Recognizing pitfalls in virtual screening: A critical review. *J Chem Inf Model* **2012**, *52* 867–881.
23. Willett, P. In *Chemoinformatics: concepts, methods, and tools for drug discovery*; Bajorath, J., Ed.; Humana Press; Totowa, New Jersey, 2004; p 51.
24. Agrafiotis, D. K. Diversity of chemical libraries. *ECC* **1998**, *1*, 742.
25. Taylor, R. Simulation analysis of experimental design strategies for screening random compounds as potential new drugs and agrochemicals. *J Chem Inf Comput Sci* **1995**, *35*, 59.
26. Bayada, D. M.; Hamersma, H.; & Van Geerestein, V. J. Molecular diversity and representativity in chemical databases. *J Chem Inf Comput Sci* **1999**, *39*, 1.
27. Rivera-Borroto, O. M.; Marrero-Ponce, Y.; García-de la Vega, J. M.; & Grau-Ábalo, R. d. C. Comparison of combinatorial clustering methods on pharmacological data sets represented by machine learning-selected real molecular descriptors. *J Chem Inf Model* **2011**, *51*, 3036.
28. Rivera-Borroto, O. M.; Rabassa-Gutiérrez, M.; Grau-Ábalo, R. d. C.; Marrero-Ponce, Y.; & García-de la Vega, J. M. Dunn's index for cluster tendency assessment of pharmacological data sets. *Can J Physiol Pharmacol* **2012**, *90*, 425.
29. Doweyko, A. QSAR: Dead or alive? *J Comput -Aided Mol Des* **2008**, *22*, 81.
30. Deardena, J. C.; Cronina, M. T. D.; & Kaiserb, K. L. E. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). *SAR QSAR Environ Res* **2009**, *20*, 241.
31. Marrero-Ponce, Y.; Meneses-Marcel, A.; Rivera-Borroto, O. M.; García-Domenech, R.; De Julián-Ortiz, J. V.; Montero, A.; Escario, J. A.; Barrio, A. G.; Pereira, D. M.; & Nogal, J. J. Bond-based linear indices in QSAR: Computational discovery of novel anti-trichomonal compounds. *J Comput -Aided Mol Des* **2008**, *22*, 523.
32. Meneses-Marcel, A.; Rivera-Borroto, O. M.; Marrero-Ponce, Y.; Montero, A.; Tugores, Y. M.; Escario, J. A.; Barrio, A. G.; Pereira, D. M.; Nogal, J. J.; & Kouznetsov, V. V. New antitrichomonal drug-like chemicals selected by bond (edge)-based TOMOCOMD-CARDD descriptors. *J Biomol Screening* **2008**, *13*, 785.
33. Rivera-Borroto, O. M.; Marrero-Ponce, Y.; Meneses-Marcel, A.; Escario, J. A.; Gómez Barrio, A.; Arán, V. J.; Martins Alho, M. A.; Montero Pereira, D.; Nogal, J. J.; & Torrens, F. Discovery of novel trichomonacidals using LDA-driven QSAR models and bond-based bilinear indices as molecular descriptors. *QSAR Comb Sci* **2009**, *28*, 9.
34. Campillo, N. E.; González-Naranjo, P.; & Páez, J. A. Presente y futuro en el descubrimiento de fármacos para la enfermedad de Chagas. *An R Acad Nac Farm* **2012**, *78*, 34.
35. Wu, X.; Kumar, V.; Ross Quinlan, J.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G. J.; Ng, A.; Liu, B.; & Yu, P. S. Top 10 algorithms in data mining. *Knowl Inf Syst* **2008**, *14*, 1.
36. Johnson, M. A.; & Maggiora, G. M. *Concepts and applications of molecular similarity*; Wiley; New York, 1990.
37. Hofstadter, D. In *The analogical mind: Perspectives from cognitive science*; Gentner, D., Ed.; The MIT Press; Cambridge, Massachusetts, 2001; p 541.
38. Martin, Y. C.; Kofron, J. L.; & Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J Med Chem* **2002**, *45*, 4350.
39. Maldonado, A. G.; Doucet, J. P.; Petitjean, M.; & Fan, B.-T. Molecular similarity and diversity in chemoinformatics: From theory to applications. *Mol Div* **2006**, *10*, 39.
40. Maggiora, G.; & Shanmugasundaram, V. In *Chemoinformatics and Computational Chemical Biology*; Bajorath, J., Ed.; Humana Press; New York, 2011; p 77.
41. Willett, P. Searching techniques for databases of two- and three-dimensional chemical structures. *J Med Chem* **2005**, *48*, 4183.
42. Willett, P. Chemoinformatics-similarity and diversity in chemical libraries. *Curr Opin Biotechnol* **2000**, *11*, 85.
43. Stumpfe, D.; & Bajorath, J. Exploring activity cliffs in medicinal chemistry. *J Med Chem* **2012**.
44. Bajorath, J.; Li, R.; Stumpfe, D.; Vogt, M.; & Geppert, H. C. Development of a method to consistently quantify the structural distance between scaffolds and to assess scaffold hopping potential. *J Chem Inf Model* **2011**.
45. Willett, P. Similarity methods in chemoinformatics. *Annu Rev Inf Sci Technol* **2009**, *43*, 1.

46. Rivera Borroto, O. M.; Hernández Díaz, Y.; García de la Vega, J. M.; Grau Ábalo, R. d. C.; & Marrero Ponce, Y. Novel similarity measures for the effective and efficient retrieval of pharmacological data sets. *Afinidad* **2011**, *68*, 50.
47. Sheridan, R. P.; & Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discov Today* **2002**, *7*, 903.
48. Willett, P. Data fusion in ligand-based virtual screening. *QSAR Comb Sci* **2006**, *25*, 1143.
49. Brown, R. D. Descriptors for diversity analysis. *Perspect Drug Disc Design* **1997**, *7*, 31.
50. National Center for Biotechnology Information. PubChem. http://pubchem.ncbi.nlm.nih.gov/ (visitado el 1 de octubre de 2013).
51. National Institutes of Health. National Cancer Institute. https://resresources.nci.nih.gov/resources/ (visitado el 1 de octubre de 2013).
52. The Cheminformatics and QSAR Society. http://www.qsar.org (visitado el 1 de octubre de 2013).
53. International Academy of Mathematical Chemistry. http://www.iamc-online.org/ (visitado el 1 de octubre de 2013).
54. Daylight Chemical Information Systems. WDI. http://www.daylight.com (visitado el 1 de octubre de 2013).
55. Sunset Molecular Discovery. WOMBAT. http://sunsetmolecular.com (visitado el 1 de octubre de 2013).
56. Baykoucheva, S. A new era in chemical information: PubChem, DiscoveryGate, and Chemistry Central. *Online* **2007**, *31 Issue , p16*, 16.
57. Bender, A. Compound bioactivities go public. *Nature Chem Biol* **2010** *6*, 309.
58. Rohrer, S. G.; & Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J Chem Inf Model* **2009**, *49*, 169.
59. Fourches, D.; Muratov, E.; & Tropsha, A. Trust, But Verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model* **2010**, *50*, 1189.
60. Johnson, M. A. A review and examination of mathematical spaces underlying molecular similarity analysis. *J Math Chem* **1989** *3*, 117.
61. Maggiora, G. M.; & Shanmugasundaram, V. In *Chemoinformatics*; Bajorath, J., Ed.; Humana Press; 2004; p 1.
62. Agrafiotis, D. K.; Bandyopadhyay, D.; Wegner, J. K.; & van Vlijmen, H. Recent advances in chemoinformatics. *J Chem Inf Model* **2007**, *47*, 1279.
63. Wegner, J. K.; Fröhlich, H.; Mielenz, H. M.; & Zell, A. Data and graph mining in chemical space for ADME and activity data sets. *QSAR Comb Sci* **2006**, *25*, 205.
64. Cuissart, B.; Touffet, F.; Cremilleux, B.; Bureau, R.; & Rault, S. The maximum common substructure as a molecular depiction in a supervised classification context: experiments in quantitative structure/biodegradability relationships. *J Chem Inf Comput Sci* **2002**, *42*, 1043.
65. Adamson, G. W.; & Bush, J. A. A method for the automatic classification of chemical structures. *Inf Stor Retriev* **1973**, *9*, 561.
66. Willett, P.; & Winterman, V. A comparison of some measures for the determination of inter-molecular structural similarity. *Quant Struct-Activ Relat* **1986**, *5*, 18.
67. Brown, R. D.; & Martin, Y. C. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J Chem Inf Comput Sci* **1996**, *36*, 572.
68. Matter, H.; & Potter, T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J Chem Inf Comput Sci* **1999**, *39*, 1211.
69. Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; & Weinberger, L. E. Neighbourhood behaviour: A useful concept for validation of "molecular diversity" descriptors. *J Med Chem* **1996**, *39*, 3049.
70. Siegel, S.; & Castellan, N. J. *Nonparametric statistics for the behavioral sciences*; McGraw-Hill; New York, USA, 1988.
71. Todeschini, R.; & Consonni, V. *Molecular Descriptors for Chemoinformatics*; 2nd ed.; WILEY-VHC; Weinheim, Germany, 2009.
72. DRAGON for Windows 5.5; Milano, Italy, 2007. Este software se encuentra disponible en: http://www.talete.mi.it (visitado el 1 de octubre de 2013).
73. PaDEL-Descriptor, 1.0; Singapore, 2010. Este software se encuentra disponible en: http://padel.nus.edu.sg/software/padeldescriptor (visitado el 1 de octubre de 2013).

74. Li, Z.; Han, L.; Xue, Y.; Yap, C.; Li, H.; Jiang, L.; & Chen, Y. MODEL—molecular descriptor lab: A web-based server for computing structural and physicochemical features of compounds. *Biotechnol Bioeng* **2007**, *97*, 389. Este software se encuentra disponible en: http://jing.cz3.nus.edu.sg/cgi-bin/model/model.cgi (visitado el 1 de octubre de 2013).
75. Molecular descriptors: The free online resource. http://www.moleculardescriptors.eu/index.htm (visitado el 1 de octubre de 2013).
76. Bender, A.; & Glen, R. C. Molecular similarity: A key technique in molecular informatics. *Org Biomol Chem* **2004**, *2*, 3204.
77. Janecek, A.; Gansterer, W.; Demel, M.; & Ecker, G. In *Proceedings of the Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery (FSDM 2008)*; Saeys, Y., Liu, H., Inza, I., Wehenkel, L., Van de Peer, Y., Eds.; JMLR: Workshop and Conference Proceedings; Antwerp, Belgium, 2008; p 90.
78. Steinbach, M.; Ertöz, L.; & Kumar, V. In *New directions in statistical physics: econophysics, bioinformatics, and pattern recognition*; Wille, L. T., Ed.; Springer-Verlag; Berlin, 2000; p 273.
79. John, G. H.; Kohavi, R.; & Pfleger, K. In *Eleventh International Conference on Machine Learning (ICML)* Cohen, W. W., Hirsh, H., Eds.; Morgan Kaufman; Rutgers University, New Brunswick, NJ, USA, 1994; p 121.
80. Watanabe, S. *Knowing and guessing: A quantitative study of inference and information*; John Wiley & Sons Inc; New York, 1969.
81. Roth, H. J. There is no such thing as 'diversity'! *Curr Opin Chem Biol* **2005**, *9*, 293.
82. Böcker, A.; Schneider, G.; & Teckentrup, A. Status of HTS data mining approaches. *QSAR Comb Sci* **2004**, *23*, 207.
83. Selwood, D. L.; Livingstone, D. J.; Comley, J. C. W.; O'Dowd, A. B.; Hudson, A. T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; & Stables, J. N. Structure-activity relationships of antifilarial antimycin analogues, a multivariate pattern recognition study. *J Med Chem* **1990**, *33*, 136.
84. 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.
85. Dudek, A. Z.; Arodz, T.; & Gálvez, J. Computational methods in developing quantitative structure-activity relationships (QSAR): A review. *Comb Chem High Throughput Screen* **2006**, *9*, 1.
86. Nath, R.; Rajagopalan, B.; & Ryker, R. Determining the saliency of input variables in neural networks classifiers. *Comput Ops Res* **1997**, *24*, 767.
87. Koivalishyn, V.; Tetko, V. I.; Luik, A. I.; Kholodovych, V. V.; Villa, A. E. P.; & Livingstone, D. J. Neural networks studies. Variable selection in the cascade-correlation learning architecture. *J Chem Inf Comput Sci* **1998**, *38*, 651.
88. Todeschini, R.; Galvagni, D.; Vilchez, J. L.; Del Olmo, M.; & Navas, N. Kohonen artificial neural networks as a tool for wavelength selection in multicomponent spectrofluorimetric PLS modeling: application to phenol, o-cresol, m-cresol and p-cresol mixtures. *Trends Anal Chem* **1999**, *18*, 93.
89. Burden, F. D.; Ford, M. G.; Whitley, D. C.; & Winkler, D. A. Use of automatic relevance determination in QSAR studies using Bayesian neural networks. *J Chem Inf Comput Sci* **2000**, *40*, 1423.
90. Agrafiotis, D. K.; & Cedeno, W. Feature selection for structureactivity correlation using binary particle swarms. *J Med Chem* **2002**, *45*, 1098.
91. Tetko, I. V.; Villa, A. E.; & Livingstone, D. J. Neural network studies. Variable selection. *J Chem Inf Comput Sci* **1996**, *36*, 794.
92. Glen, R. C.; & Adams, S. E. Similarity metrics and descriptor spaces – Which combinations to choose? *QSAR Comb Sci* **2006**, *25*, 1133.
93. Guyon, I.; & Elisseeff, A. An introduction to variable and feature selection. *J Mach Lear Research* **2003**, *3*, 1157.
94. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; & Witten, I. H. The WEKA Data Mining Software: An Update. *SIGKDD Explor Newsl* **2009** *11*, 10.
95. Machine Learning Group. Weka. http://www.cs.waikato.ac.nz/ml/weka/ (visitado el 1 de octubre de 2013).
96. Tversky, A. Features of similarity. *Psychol Rev* **1977**, *84*, 327.
97. Chen, X.; & Brown, F. K. Asymmetry of chemical similarity. *ChemMedChem* **2007**, *2*, 180

98. Ágoston, V.; Kaján, L.; Carugo, O.; Hegedüs, Z.; Vlahovicek, K.; & Pongor, S. In *Essays in Bioinformatics*; Moss, D. S., Jelaska, S., Pongor, S., Eds.; IOS Press; The Netherland, 2005; p 11.
99. Ellis, D.; Furner-Hines, J.; & Willett, P. Measuring the degree of similarity between objects in text retrieval systems. *Perspect Inf Manag* **1994**, *3*, 128.
100. Cuadras, C. M. Distancias estadísticas. *Estadística Española* **1989**, *30*, 295.
101. Willett, P.; Barnard, J. M.; & Downs, G. M. Chemical similarity searching. *J Chem Inf Comput Sci* **1998**, *38*, 983.
102. Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov Today* **2006**, *11*, 1046.
103. David, H. W.; & William, G. M. *No Free Lunch Theorems for Search*, 1995.
104. Wolpert, D. H.; & Macready, W. G. No free lunch theorems for optimization. *IEEE T Evolut Comput* **2002**, *1*, 67.
105. Grünwald, P. In *Advances In Minimum Description Length: Theory And Applications*; Grünwald, P. D., Myung, I. J., Pitt, M. A., Eds.; MIT Press; Cambridge, Massachusetts, 2005; p 3.
106. Willett, P. Some heuristics for nearest-neighbor searching in chemical structure files. *J Chem Inf Comput Sci* **1983**, *23*, 22.
107. Friedman, J. H.; Bentlev, J. L.; & Finkel, R. A. An algorithm for finding best matches in-logarithmic expected time. *ACM Trans Marh Softw* **1977**, *3*, 209.
108. Bentley, J. L.; Weide, B. W.; & Yao, A. C. Optimal expected time algorithms for closest point problems. *ACM Trans Marh Softw* **1980**, *6*, 563.
109. Smeaton, A. F.; & Van Rijsbergen, C. J. The nearest neighbour in information retrieval. an algorithm using upperbounds. *ACM SIGIR Forum* **1981**, *16*, 83.
110. Murtagh, F. A very fast, exact nearest neighbour algorithm for use in information retrieval. *Inf Technol: Res Deu* **1982**, *1*, 275.
111. Van Marlen, G.; & Van Den Hende, J. H. Search strategy and data compression for a retrieval system with binary-coded mass spectra. *Anal Chim Acra* **1979**, *112*, 143.
112. Rasmussen, G. T.; Isenhour, T. L.; & Marshall, J. C. Mass spectral library searches using ion series data compression. *J Chem Inf Comput Sci* **1979**, *19*, 98.
113. Baldi, P.; Hirschberg, D. S.; & Nasr, R. J. Speeding up chemical database searches using a proximity filter based on the logical exclusive OR. *J Chem Inf Model* **2008**, *48*, 1367.
114. 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.
115. Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; & Sheridan, R. P. Chemical similarity using physiochemical property descriptors. *J Chem Inf Comput Sci* **1996**, *36*, 118.
116. Ginn, C. M. R.; Willett, P.; & Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect Drug Discov Des* **2000**, *20*, 1.
117. Hert, J.; Willett, P.; Wilton, D. J.; 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.
118. Nasr, R. J.; Swamidass, S. J.; & Baldi, P. F. Large scale study of multiple-molecule queries. *J Cheminf* **2009**, *1*, 1.
119. Chen, B.; Mueller, C.; & Willett, P. Combination rules for group fusion in similarity-based virtual screening. *Mol Inf* **2010**, *29*, 533
120. Geppert, H.; & Bajorath, J. Advances in 2D fingerprint similarity searching. *Expert Opin Drug Discov* **2010** *5*, 529.
121. Nicholls, A. What do we know and when do we know it? *J Comput-Aided Mol Des* **2008**, *22*, 239.
122. Witten, I. H.; & Frank, E. *Data Mining - Practical Machine Learning Tools and Techniques*; 2nd ed.; Morgan Kaufmann; San Francisco, CA, 2005; 161-176.
123. Truchon, J.; & Bayly, C. I. Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. *J Chem Inf Model* **2007**, *47*, 488.
124. Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; & Kearsley, S. K. Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *J Chem Inf Model* **2001**, *41*, 1395.

125. Clark, R.; & Webster-Clark, D. Managing bias in ROC curves. *J Comput-Aided Mol Des* **2008**, *22*, 141.
126. Swamidass, S. J.; Azencott, C.-A.; Daily, K.; & Baldi, P. A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. *Bioinformatics* **2010**, *26*, 1348