Modelos de proximidad novedosos para el cribado virtual de conjuntos de datos quimioinformáticos

Colectivo de Autores

Edición: José Angel Morejón

Corrección: Estrella Pardo Rodríguez

Editorial Feijóo, 2012

ISBN: 978-959-250-811-8

EDITORIAL
Feijóo

Editorial Samuel Feijóo, Universidad Central “Marta Abreu” de Las Villas, Carretera a Camajuaní, km 5 ½, Santa Clara, Villa Clara, Cuba. CP 54830

RESUMEN

La búsqueda de similitud es una prestación importante en los sistemas modernos de gestión de la información química para acceder a la rica información contenida en los enormes repositorios químicos modernos. Básicamente, dadas una representación molecular, una medida de similitud y un algoritmo de búsqueda, la salida de la técnica devuelve una lista ordenada de moléculas del conjunto de datos en orden decreciente de similitud con respecto a la molécula consulta especificada por el usuario. Como consecuencia, los investigadores han puesto su interés en la eficacia de las representaciones y medidas de similitud en estas tareas. Sin embargo, sus estudios se han enfocado predominantemente en representaciones binarias y las medidas de semejanza correspondientes, y poco se ha trabajado en otros tipos de descripción numérica. También se han aplicado técnicas del Aprendizaje Automático en la selección de rasgos, aunque no de forma consistente con el *principio de vecindad*. Estos precedentes junto a la necesidad de nuevos métodos apropiados para cada contexto químico, constituyen la motivación para este trabajo. El mismo comprende la implementación computacional en el ambiente *Java* de 21 modelos de proximidad, 9 de los cuales son novedosos en Quimioinformática, proceden del área de la Psicología y están basados en el concepto *acuerdo relacional*, y otros doce son medidas ya establecidas de la literatura especializada. Posteriormente, las nuevas medidas de similitud fueron comparadas y validadas en la "recuperación temprana" usando nueve conjuntos farmacológicos de la Química Medicinal de interés internacional, representados por descriptores numéricos, seleccionados por Aprendizaje Automático, y un algoritmo de búsqueda eficiente. Los resultados muestran que en tendencia promedia los nuevos modelos se comportan superiormente a los de referencia y que más de la mitad de los mismos se sitúan entre los diez modelos más potentes.

ÍNDICE

CAPÍTULO 1. REVISIÓN BIBLIOGRÁFICA 10

1.1 Generalidades del Proceso de Descubrimiento de Fármacos 10

1.2 Métodos Computacionales en el Proceso de Descubrimiento de Fármacos 14

1.3 Generalidades de la Similitud Molecular 17

1.4 Consideraciones finales del capítulo 31

CAPÍTULO 2. MATERIALES Y MÉTODOS 33

2.1 Similitud Molecular 33

2.1.1 Medidas de Semejanza Molecular Reportadas 34

2.1.2 Acuerdo Relacional como Medidas de Similitud Molecular Novedosas 36

2.2 Conjuntos de Datos para Validación 43

2.3 Representación Molecular y Obtención de las Matrices de Datos 46

2.4 Métricas de Rendimiento 49

2.5 Diseño Experimental y Análisis Estadístico 51

2.6 Diseño e Implementación Computacionales 52

2.7 Conclusiones Parciales del Capítulo 58

CAPÍTULO 3. RESULTA1DOS Y DISCUSIÓN 60

3.1 Análisis de Dimensionalidad de los Datos 60

3.2 Comparación de los Modelos de Proximidad 67

3.3 Conclusiones Parciales 74

CONCLUSIONES 76

RECOMENDACIONES 77

REFERENCIAS BIBLIOGRÁFICAS 78

INTRODUCCIÓN

El acceder a bases de datos de estructuras químicas es una de las facilidades más importantes en los sistemas modernos para la gestión de la información química. Los primeros sistemas se enfocaron en la provisión de prestaciones para la búsqueda subestructural, las cuales fueron complementadas en los finales de los ochenta con la búsqueda de similitud (Carhart *et al*., 1985, Willett *et al.,* 1986). La diferencia entre ambas técnicas es que con la primera se recuperan aquellas moléculas que contienen la *subestructura de consulta* (ya sea 2D o 3D), mientras que la segunda persigue identificar aquellas moléculas de la base de datos que son más similares a una *estructura de referencia* definida por el usuario, usando alguna definición cuantitativa de similitud estructural intermolecular (Willett *et al.,* 1998a, Sheridan and Kearsley, 2002c). La estructura de referencia se caracteriza a través de uno o más descriptores moleculares, y este conjunto se compara con los conjuntos de descriptores correspondientes de cada molécula de la base de datos. Estas comparaciones permiten el cálculo de una medida de similitud entre la estructura de referencia y cada una de las estructuras de la base de datos, las cuales se ordenan posteriormente en orden decreciente de similitud con la referencia. De esta forma, la salida de cada búsqueda es una lista ranqueada en la cual las estructuras que se calculan como las más similares a la estructura referencia, los *vecinos más cercanos*, están localizados al principio de la lista. Estos vecinos forman la salida inicial de la búsqueda y serán aquellos que tienen la mayor probabilidad de ser de interés al usuario, dada una medida apropiada de la similitud estructural intermolecular (Willett, 2005).

Por consiguiente, la opción del tipo de representación molecular y la medida de similitud correspondiente tiene una influencia determinante en la técnica de búsqueda de similitud (Glen and Adams, 2006a). Varios estudios comparativos del desempeño de las medidas de similitud en la recuperación efectiva y eficiente en bases de datos quimioinformáticas han aparecido en la literatura, resaltando sus méritos y deficiencias

relativos (Chen and Reynolds, 2002, Bender *et al*., 2009, Hert *et al*., 2004b). La mayoría de estas investigaciones se han enfocado en el uso de cadenas binarias para la representación molecular, reportando el coeficiente de Tanimoto como el más eficaz para estas tareas (Willett, 2006b). Otros trabajos han empleado representaciones no binarias pero sólo aprovechan una fracción de la rica información contenida en las entidades moleculares (Whittle et al., 2003), que puede ser accedida actualmente en una magnitud apreciable a través de los softwares de optimización y cálculo de descriptores moleculares (Todeschini and Consonni, 2009b). A pesar de que estas investigaciones se han llevado a cabo en escenarios supervisados, la selección de los descriptores ha sido guiada por la experiencia de los investigadores; aunque algunos pocos trabajos han introducido técnicas de Aprendizaje Automático para este fin (Bender *et al.,* 2004b), no lo han hecho de forma consistente con el *principio de similitud* (Nikolova and Jaworska, 2003b). Por último, ninguna medida de similitud se comporta igualmente superior con independencia del contexto químico en consideración (Holliday *et al.,* 2002), lo cual deja una puerta abierta para trabajos de desarrollo y comparación de medidas de semejanza novedosas.

Por todo lo anterior se plantea el siguiente ***Problema Científico:***

A pesar de la gran cantidad de estudiosde comparación de medidas de similitud basados en la representación binaria reportados, poca atención se le ha prestado a las medidas de semejanza basadas en otro tipo de representación numérica. La mayoría de lossoftwarespara la búsqueda de similitud no están disponibles de manera gratuita y su adquisición es costosa. No existe consenso en la literatura quimioinformática en cuanto a quétipo y cantidad de descriptores moleculares se deben calcular, así como cuáles de ellos se deben seleccionar para obtener una "detección temprana" efectiva usando la técnica de búsqueda de similitud.

Para dar respuesta a la problemática científica planteada, se propone el siguiente ***Objetivo General***:

Introducir y validar modelos de proximidad numéricos novedosos en el área de la Quimioinformática usando una metodología apropiada que integre un algoritmo de

búsqueda eficiente, conjuntos de datos farmacológicos estándares, un tipo de representación informativa, una técnica de selección de rasgos automática consistente con el *principio de similitud*, y una métrica de validación apropiada para el *problema de la detección temprana*.

Dicho objetivo general se desglosa en las siguientes ***Etapas u Objetivos Específicos***:

1. 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 estrategias y métricas de validación automática.
2. Introducir y comparar varios modelos novedosos procedentes del área de la Psicología con otros ya reportados usando una metodología de validación apropiada.
3. Establecer una metodología que integre el cálculo y selección de descriptores moleculares consistentes con el principio de similitud, un algoritmo de búsqueda efectivo basado en fusión de datos y una métrica de exactitud que se ajuste al problema del "enriquecimiento temprano".

Así, después de haber realizado el marco teórico se llegó a formular la siguiente ***Hipótesis de Investigación***:

> Es posible identificar modelos de proximidad relacional más potentes que los modelos reportados mediante el desarrollo de una metodología de validación que combina un método de representación consistente con el principio de similitud, un algoritmo de búsqueda efectivo y una métrica de exactitud adecuada para la "detección temprana".

La presente tesis se estructura en tres capítulos. El primero de ellos está dedicado al Marco Teórico donde se brinda información acerca del paradigma tradicional del descubrimiento de fármacos, de las técnicas de cribado virtual como alternativa a dicho proceso y de las técnicas de búsqueda de similitud según los propósitos del presente trabajo. El segundo se dedica a la presentación de los materiales y métodos utilizados a lo largo de la investigación, donde se muestra una descripción teórica de los modelos de proximidad estudiados, las bases de datos para la comparación y validación de dichas

medidas, los descriptores calculados para los entes moleculares y las estrategias de cálculo molecular y selección de rasgos; finalmente se expone el diseño de experimentos más apropiado para la comparación de los modelos de semejanza. En el tercer y último capítulo se describen los análisis de los resultados arribados en las etapas de selección de rasgos, y, la comparación y validación de las medidas de proximidad.

CAPÍTULO 1. REVISIÓN BIBLIOGRÁFICA

Este capítulo estará dedicado al Marco Teórico. Comenzaremos explicando las características generales del proceso tradicional de descubrimiento de fármacos, luego hablaremos de las limitaciones inherentes de estos paradigmas. Posteriormente se introducirá a las técnicas de cribado virtual como alternativa racional a dicho proceso. Después nos detendremos en la técnica virtual de búsqueda de similitud, su tipología y componentes esenciales, la cual constituye la fuente de motivación para este trabajo.

1.1 Generalidades del Proceso de Descubrimiento de Fármacos

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 cierto 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, el tercer paso (validación clínica) debe probar su eficiencia y seguridad a través de una evaluación extensiva en animales y humanos (Drews, 2000, Kubinyi, 1995).

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, lo que se conoce en este ámbito como durabilidad(Drews, 2000). 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* (Chanda and Caldwell, 2003).

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 (Ren and Stammers, 2005). 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 (Drews, 2000). La identificación de candidatos prometedores en esta vasta (casi infinita) cantidad de moléculas depende fuertemente de la pericia bioquímica y tradicionalmente se logra en un proceso iterativo, denominado como 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(Manly *et al.,* 2001).

Durante los ensayos biológicos iniciales del ciclo de descubrimiento de fármacos son identificados los denominados *hits*. A esta fase de generación de hits le sigue la fase de generación de *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 (Jorgensen, 2004). 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.

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 (DiMasi et al., 2003).

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 (DiMasi *et al.,* 2003):

- 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 provistos 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 (*Food and Drugs Administration,* FDA, en inglés) en los Estados Unidos de América, por la Agencia Europea para la Evaluación de Productos Médicos en Europa o para el caso particular de nuestro país por el Centro para el Control Estatal de la Calidad de los Medicamentos (CECMED).

Necesidad de Nuevos Paradigmas

El descubrimiento y desarrollo de nuevos medicamentos es un proceso en extremo complejo y costoso en términos de tiempo y dinero. Durante la década de los 90 el costo promedio del desarrollo de un nuevo fármaco, desde la identificación del blanco hasta su aprobación, estaba alrededor de los 15 años y US$ 900 millones (Drews, 2000).En las dos décadas pasadas, los avances tecnológicos, junto con la progresiva reconsideración del proceso en sí, ha conducido a una gran revolución del proceso de encontrar un nuevo fármaco.

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 fármacos PDDF (*Drug Discovery and Development Process*, DDDP, en inglés) debido al costo de la síntesis y evaluación de nuevas moléculas (Bleicher *et al.,* 2003). Durante esta etapa las esperanzas de resolver el problema del PDDF fueron puestas en el desarrollo de las tecnologías de alto rendimiento (Bajorath, 2002) y la química combinatoria (Lazo and Wipf, 2000), 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 (Bleicher *et al.,* 2003). 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 (Bleicher *et al.*, 2003).

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 (Ekins *et al.,* 2002). 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. 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 (Bleicher *et al.,* 2003).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. Este cambio de paradigma marca un avance significativo hacia su racionalización (DiMasi *et al.,* 2003).

1.2 Métodos Computacionales en el Proceso de Descubrimiento de Fármacos

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" definiendo la 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» (Cruz Monteagudo, 2009). La Quimioinformática implica el uso de las tecnologías informáticas para procesar los datos químicos. Lo que diferencia a los datos de procesamiento químico de la transformación de otros datos es que los datos químicos implica la obligación de trabajar con estructuras químicas. Esta definición general engloba múltiples aspectos, en particular, la representación, almacenamiento, recolección y análisis de la información química en un sistema informático (Hann and Green, 1999).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.

Cribado virtual o "Virtual Screening"

Cribado virtual o *in silico* es el término usado para denotar el análisis computacional de bases de datos de compuestos, dirigido a identificar candidatos que posean la actividad biológica deseada sobre un blanco terapéutico específico. Esto puede verse como una alternativa al desarrollo de ensayos experimentales con la principal ventaja de que pueden ser evaluadas cantidades arbitrarias de moléculas reales o virtuales. Con la identificación de compuestos potencialmente activos, el cribado virtual puede por tanto ayudar a la reducción del número de ensayos experimentales y motivar la obtención o síntesis de nuevas moléculas (Debouck and Goodfellow, 1999). En la práctica, el cribado virtual requiere del conocimiento de la estructura del blanco terapéutico, usualmente obtenido por métodos cristalográficos, o de la actividad, que se mide 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 (Jorgensen, 2004). Si se desconoce la estructura del blanco, los métodos de cribado virtual pueden derivarse de un *pool* de compuestos con actividad conocida obtenidos de ensayos experimentales previos. Estos métodos se conocen como enfoques del cribado virtual basados en ligandos, en oposición al enfoque anterior basado en la estructura del blanco. Un método simple basado en ligandos consiste en el ordenamiento de las moléculas de una base de datos con respecto a su similitud con un conjunto de compuestos activos conocidos, y la selección de los candidatos mejor ordenados como los más prometedores (Xue *et al*., 2004). 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 (Xue *et al.,* 2004).

Cuando se conoce la estructura del blanco, las aproximaciones basadas en la estructura constituyen la vía más racional para el cribado virtual (Klebe, 2000). Por su parte, las aproximaciones basadas en ligandos son de un interés considerable ya que los métodos de acoplamiento son muy costosos desde el punto de vista computacional, y difíciles de

automatizar (Kubinyi, 1995). En segundo lugar, la estructura del blanco es difícil de obtener, siendo en este caso igualmente imposible de aplicar un estudio de acoplamiento.

El enfoque más general 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 *pool* de moléculas previamente evaluadas, integrando así información tanto de la actividad como de la inactividad de los compuestos, Este problema se conoce como la modelación de la relación estructura-actividad (REA) más comúnmente conocido por sus siglas en inglesQSAR, acrónimo de *Quantitative Structure-Activity Relationships*, e involucra métodos de los campos de la estadística y el aprendizaje computarizado (machine learning).

Una de las herramientas más simples del cribado virtual lo constituye la similitud molecular, la cual es ampliamente utilizada en las primeras etapas 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 otros estudios detallados decribado virtual que emplean técnicas más sofisticadas (Flower, 1998). En una búsqueda normal, las moléculas en un repositorio están clasificadas por su similitud a una consulta de una o varias moléculas. Si se emplea una métrica de similitud apropiada, es más probable que las moléculas más similares a la consulta exhiban propiedades físicas, químicas, biológicas similares a la molécula de la consulta (Fligner *et al.,* 2002).La búsqueda de similitud con huellas dactilares moleculares es un enfoque bien establecido en la investigación del cribado virtual, con una historia que se remonta a varias décadas. A partir de la década de los sesenta, uno de los orígenes primarios de la similitud molecular fue el desarrollo de algoritmos para la búsqueda de subestructura (Fligner *et al.,* 2002). Este enfoque comienza con un fragmento molecular como consulta a bases de datos químicas y recupera todos los compuestos que contienen este fragmento. Por lo tanto, la búsqueda de subestructura examina "si/no" la ocurrencia del fragmento en las moléculas, pero no evalúa los diferentes grados de similitud estructural.

Importancia de la Similitud Molecular en el Proceso de Descubrimiento de Fármacos

Los métodos de similitud molecular como cribado de alto rendimiento y la técnica de química combinatoria nos permiten producir y seleccionar los compuestos de una manera racional, y probar sus diferentes actividades biológicas (Oprea and Matter, 2004). Estos

enfoques se aplican en grandes bases de datos de cientos de miles de moléculas (Bayada *et al.,* 1999). El número previsto de dianas farmacológicas potenciales (ver Figura 1.1) anima a los químicos medicinales, para usar la similitud molecular y técnicas de diversidad en el diseño de fármacos (Cruz Monteagudo, 2009). El interés actual en la similitud molecular y en el aumento de la diversidad es esencial para la investigación de nuevos y mejores fármacos potencialmente más seguros y eficientes y con menos costo en comparación con las moléculas existentes. Los laboratorios farmacéuticos han entendido el potencial interés económico para buscar moléculas similares en lugar de nuevos compuestos.

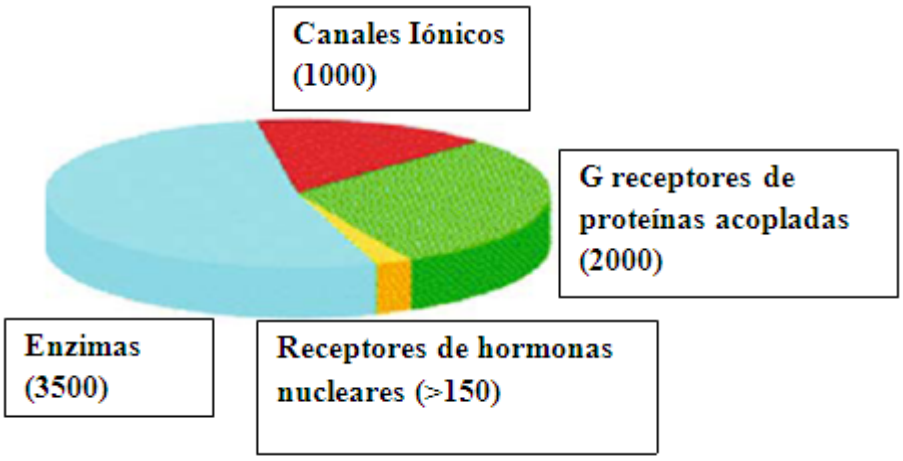


Figura 1.1 Número estimado de dianas farmacológicas potentes que pertenecen a diferentes clases bioquímicas

1.3 Generalidades de la Similitud Molecular

La noción de similitud desempeña un papel importante en la Quimioinformática debido a un paradigma central dentro de esta área conocido como el *principio de similitud*, el cual plantea que moléculas con estructuras similares tienden a exhibir propiedades biológicas similares (Bunke and Shearer, 1998). Este principio justifica los métodos que involucran algoritmos de agrupamiento, particionamiento y ordenamiento, y es por esto, que el acceso a la similitud estructural ha venido despertando una atención considerable en la comunidad quimioinformática.

Las diferentes representaciones y sus correspondientes medidas de similitud pueden llevar a resultados radicalmente diferentes (Cuissart *et al.,* 2002). Cuatro tipos de objetos matemáticos se utilizan normalmente para representar las moléculas, los cuales son:

conjuntos, gráficos, vectores y funciones. Los conjuntos son los objetos más generales y, básicamente, la base de los otros tres. Normalmente, los químicos representan moléculas como "gráficos químicos" (Cuissart *et al.,* 2002), que están estrechamente relacionados con los tipos de gráficos tratados por los matemáticos en el campo de la teoría de grafos (Adamson and Bush, 1973). La mayoría de los grafos químicos describen la naturaleza de los átomos y cómo estos son enlazados. Por lo tanto, se dice a veces que los grafos químicos proporcionan una representación en 2-D (dos dimensiones) de las moléculas. Ellos no suelen contener información sobre las características esenciales en 3-D de las moléculas, aunque los grafos químicos captan parte de esta información (Adamson and Bush, 1975). Con todo, las estructuras tridimensionales son utilizadas ampliamente, sobre todo ahora que numerosos programas informáticos han sido desarrollados para su cálculo y visualización.

Los grafos químicos proporcionan una metáfora potente e intuitiva para la comprensión de muchos aspectos de la química, pero 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 las características de los vectores, donde cada componente corresponde a una función "local" o "global" o característica de una molécula, que por lo general, es representada por uno, de una serie de descriptores que podrían guardar relación con la función elegida. Las características localesincluyen fragmentos moleculares ("subestructuras"), potentes farmacóforos (Willett and Winterman, 1986), varios índices topológicos (Brown and Martin, 1996), y cargas atómicas parciales, entre otras. Las globales incluyen características tales como el peso molecular, logP, la superficie polar, el volumen molecular, etc., (Brown, 1997).

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 3-D, se han vuelto más frecuentes. Las características generalmente se refieren a diversos tipos de campos moleculares, algunos, como la densidad electrónica ("estérica"), otros como los campos potenciales eléctricos (Bayada *et al.,* 1999) y también como campos potenciales lipofílicos (Matter and Potter, 1999). Los campos moleculares son generalmente representados como funciones continuas. Los campos discretos también se han utilizado, aunque menos frecuentemente(Patterson *et al.,* 1996).

Debido a la gran diversidad de descriptores moleculares, numerosos métodos han sido propuestos para cuantificar la similitud molecular, basados, por ejemplo, en índices de conectividad (Martin *et al.,* 1998), pares o tripletes de átomos en las representaciones moleculares 2-D y 3-D (Martin, 2001), matrices de distancias interatómicas y mapeo atómico óptimo, o vectores de autocorrelación (Salim *et al.,* 2003). El enfoque más extendido de cuantificación de la similitud molecular consiste en el conteo del número de características sub-estructurales compartidas. Varios *coeficientes de asociación* han sido definidos para este propósito, entre estos, el *coeficiente de Tanimoto* definido para pares de vectores moleculares (*X*, *Y*) ha emergido como una medida de similitud estándar (Cruz Monteagudo, 2009), su expresión matemática está dada por:

$$T_{XY} = \frac{X^T Y}{(X^T X + Y^T Y - X^T Y)} \qquad (1)$$

Otra forma de cuantificar la similitud molecular desde el punto de vista estadístico es a través de *medidas de distancia*, particularmente útiles en el área del aprendizaje basado en casos (Aha, 1992) y el análisis de conglomerados (Massart and Kaufman, 1983). Aunque existen otras opciones posibles, la *distancia euclidiana* es la más extendida. La distancia euclidiana DE_{XY} de un caso **X**con respecto a un caso **Y**se define como:

$$DE_{XY} = \left[\sum (x_j - y_j)^2\right]^{\frac{1}{2}} \qquad (2)$$

Donde, x_jy y_j representan los respectivos valores del atributo*j* para los casos **X** e **Y**, respectivamente.

Otras medidas de distancia pueden ser más apropiadas en circunstancias especiales. De cualquier modo, aun cuando se han realizado múltiples esfuerzos dirigidos a resolver el problema de la similitud molecular (Horvath and Jeandenans, 2003), acceder de forma eficiente a la similitud molecular continúa siendo un capítulo abierto de la quimioinformática.

Como se pudo apreciar anteriormente, una de las herramientas de cribado virtual es la similitud molecular, que a su vez, dentro de esta herramienta, está la técnica de búsqueda de similitud, la cual se divide en cuatro métodos de búsqueda: búsqueda simple, fusión de

grupo, fusión de similitud y turbo similitud (Willett, 2006b) y que se comentan a continuación.

Técnicas de Búsqueda de Similitud

El principio de la búsqueda de similitud establece que si una molécula bioactiva es conocida, entonces se puede utilizar una estructura de referencia para buscar en una base de datos química las moléculas que son estructuralmente más parecidas a la estructura de referencia, que son a su vez las de mayor probabilidad de presentar la misma actividad, es decir, las que más se asemejen a dicha referencia. La búsqueda de similitud se puede implementar de muchas maneras diferentes, dando lugar a una gran cantidad de estudios comparativos que tratan de determinar qué método de similitud es "mejor" con un criterio cuantitativo de la eficiencia del cribado. Sin embargo, como Sheridan y Kearsley han señalado (Sheridan and Kearsley, 2002b), es muy poco probable que un solo mecanismo de búsqueda pueda actuar con un alto nivel en todas las circunstancias. En cambio, un enfoque más realista es el uso de múltiples métodos de búsqueda y luego combinar los resultados en un rango de salida única, un enfoque que normalmente se conoce como la fusión de datos (Willett, 2006a).

La fusión de datos se utilizó por primera vez en la búsqueda de similitud a finales de los años noventa. Básicamente, existen tres enfoques principales para la aplicación de la fusión de datos (Chen *et al.,* 2010). El primer enfoque es la fusión de similitud, que implica la búsqueda de una estructura de referencia única con múltiples medidas de similitud. Por ejemplo, una búsqueda puede llevarse a cabo mediante las huellas dactilares *Daylight*, las huellas digitales con un diseño molecular limitado MDL (Molecular Design Limited, siglas en inglés) y las huellas dactilares Unity, con la clasificación final de la base de datos que se obtiene mediante la combinación de los valores de similitud (o la clasificación correspondiente) resultantes de cada una de las tres huellas digitales diferentes. El otro enfoque que le sigue es turbo similitud, un algoritmo de fusión de grupos que se aplica cuando sólo se conoce una única estructura de referencia. La búsqueda de turbo similitud se basa en dos conceptos; en primer lugar, que la combinación de los resultados de las búsquedas de similitud con múltiples estructuras de referencia, utilizando métodos de fusión de datos, es probable que sea más eficaz que la búsqueda mediante una única estructura de referencia. En segundo lugar, que los vecinos más cercanos de una estructura

de referencia son más probables de ser igualmente activos, ya que comparten muchas características químicas comunes y que, si suponemos que son activas, entonces podemos utilizar estos vecinos más cercanos para representar las estructuras de referencias múltiples. Sería posible, por tanto, entrenar a los pesos para cada combinación de clase activa y el tamaño del rango con un único activo conocido como punto de partida. Mientras más activos de esta clase y de tamaño similar se conozcan, se pueden utilizar para mejorar aún más el entrenamiento.

El tercer y último enfoque es la fusión de grupos, que consiste en buscar múltiples estructuras de referencia con una sola medida de similitud y se ha encontrado que es más eficaz que la fusión de similitud (Chen *et al.,* 2010). La clasificación final se obtiene mediante la combinación de los valores individuales de similitud, pero este último se obtiene a partir de la utilización de diferentes estructuras de referencia (en lugar de diferentes medidas de similitud como en la fusión de similitud). Muchas reglas de fusión han sido reportadas en la literatura Quimioinformática, y algunos de ellos han demostrado ser eficaces en la operación (Chen *et al.,* 2010). En el mundo ideal, debería ser posible la aplicación de diferentes reglas de fusión para identificar los que serán más efectivos en el posicionamiento de las moléculas activas en la parte superior de los rangos fusionados. Este fue el punto de partida para los estudios de Fligner y coautores (Fligner *et al.,* 2002), pero lamentablemente, se encontró que sólo era posible modelar exitosamente la operación de una regla de fusión dada una cantidad considerable de información en cuanto a las distribuciones de similitud de las moléculas activas e inactivas. Esta información es muy poco probable que esté disponible en la etapa de descubrimiento de líderes sobre proyectos de búsquedas de fármacos, que es cuando los métodos basados en la similitud, son utilizados normalmente.

Es importante tener en cuenta los componentes de una similitud molecular a la hora de utilizar dicha herramienta del cribado virtual, ya que se deben escoger adecuadamente estos componentes para obtener resultados satisfactorios y que estén acordes al problema que se esté tratando.

1.3.1 Componentes de la Búsqueda de Similitud

En Quimioinformática, la similitud molecular y las medidas de la diversidad son complementarias: la similitud molecular proporciona un método simple y popular para el cribado virtual y la utilización de métodos de agrupamiento de bases de datos químicos. Por otra parte, el análisis de la diversidad molecular explora el camino de las moléculas para cubrir un espacio estructural determinado y subyace en muchos enfoques para la selección de compuestos y el diseño de bibliotecas combinatorias. La elección de métricas en un espacio óptimo que representa la diversidad estructural de un compuesto de la población es determinante en la eficiencia del modelo (Snarey *et al.,* 1997). Un espacio adecuado de diversidad/similitud debe permitirnos colocar las moléculas en una posición correcta en relación con las demás. El tiempo de cálculo de las similitudes y diversidades entre los compuestos es igualmente importante, debido a la creciente popularidad de grandes bases de datos (Brown, 1997). La búsqueda de similitud molecular y/o diversidad molecular implican, en general, cuatro componentes principales: el conjunto de datos, la medida de similitud/disimilitud, los descriptores moleculares y el algoritmo de búsqueda (Brown, 1997).

Conjuntos de Datos

La medición del rendimiento de los índices de similitud, descriptores moleculares (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 (Maldonado *et al.,* 2006). 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 (Sheridan and Kearsley, 2002a). Una revisión exhaustiva

acerca de las bases de datos empleadas actualmente en la Quimioinformática, haciendo énfasis en las bases de datos farmacológicas, se puede encontrar en (Jónsdóttir *et al.,* 2005). La tendencia actual de dichos repositorios es pasar el dominio público (Bender, 2010).

Espacio Químico y Representación Molecular

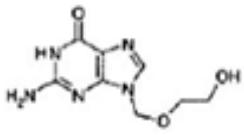
Cercanamente aliado con la noción de similitud molecular es el de *espacio químico*. Los espacios químicos proveen un medio para conceptualizar y visualizar la similitud molecular. El concepto de espacio químico se deriva de la noción de espacio usado en Matemáticas y consiste en un conjunto de moléculas y un conjunto de relaciones asociadas (similitudes, disimilitudes, distancias) entre las moléculas, lo cual le da al espacio una "estructura" (Johnson, 1989).

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, 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 (quizás infinita).La posición absoluta de las moléculas en este espacio se puede calcular solamente si se miden todas las distancias por pares y la dimensionalidad del espacio puede ser conocida (descriptores de pares de átomos, árboles de rasgos, enfoques de Subestructura Máxima Común (Maggiora and Shanmugasundaram, 2004, Agrafiotis *et al.,* 2007, Wegner *et al.,* 2006)

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(Maldonado et al., 2006).

Tabla 1.1 Algunos ejemplos de descriptores moleculares y su clasificación, calculados para estructuras moleculares de una, dos y tres dimensiones

Dimensión	Representación Típica	Descriptores típicos
1D	C8H10N5O3	Peso molecular Conteo de átomos
2D		Conteo de fragmentos Índices topológicos Conectividad
3D		Superficie molecular Volumen molecular Energía de interacción

La presentación anterior está lejos de ser exhaustiva, por lo que para una presentación detallada los lectores interesados pueden referirse a libros de texto que desarrollan este tema con profundidad (Todeschini and Consonni, 2009a). El número de descriptores moleculares propuestos en la literatura hasta el momento es realmente amplio; para ello recientemente se han desarrollado algunos sistemas para el cálculo de grandes conjuntos de descriptores (CODESSA, DRAGON, etc.).No obstante, la selección de descriptores "apropiados" para un problema específico continúa siendo una cuestión abierta.

Selección de Rasgos

La aplicación de modelos a la Quimioinformática requiere condensar la información sobre las moléculas en un conjunto limitado de *rasgos*, lo cual dista de ser una tarea trivial debido a la gran cantidad de descriptores moleculares disponibles en la actualidad (Fligner *et al.,* 2002). En este sentido, a medida que la dimensionalidad de los datos incrementa, muchos tipos de análisis de datos y problemas de clasificación se vuelven significativamente 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" (Janecek *et al.,* 2008). Para propósitos de aglomeración, 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 (Steinbach *et al.,* 2000). 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 (Böcker *et al.,* 2004). En resumen, existe una amplia variedad de DMs 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 (Glen and Adams, 2006b). 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 (Guyon and Elisseeff, 2003b).

Antes de la aparición de la química combinatoria y la generación automática de las moléculas, el problema de la gestión de grandes volúmenes de moléculas en las bases de datos era desconocido para los químicos. Diversos enfoques se han propuesto para la selección y clasificación de las moléculas, tales como la eliminación de características de descriptores que no transmiten una información relevante. Las aplicaciones de minería de datos son múltiples, estos pueden ser: la selección de sub-bases de datos que implican una o más características deseadas, la optimización del agrupamiento de compuestos para obtener una base de datos totalmente diferente, la elección inteligente de los descriptores para un modelo con propiedades particulares, etc. Numerosos métodos se han propuesto continuamente en Quimioinformática, por ejemplo, la técnica paso a paso de los procesos de integración hacia adelante o eliminación hacia atrás (Selwood *et al.,* 1990), y el análisis de componentes principales, también ha sido propuesto el uso de los *k*-vecinos más cercanos (Zheng and Tropsha, 2000).Otros métodos de selección más usados en la modelación REA se encuentran en 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 (Dudek *et al.,* 2006).

En el pasado, algunos enfoques estaban directamente relacionados con las Redes Neuronales Artificiales, como son: división de los pesos (Nath *et al.,* 1997), correlación en cascada (Koivalishyn *et al.,* 1998), mapas de Kohonen (Todeschini et al., 1999), determinación de la relevancia automática (Burden et al., 2000), etc. También han sido presentados en la literatura especializada los Sistemas Artificiales de Colonias de Hormigas y Enjambres (Agrafiotis and Cedeno, 2002). También ha sido evaluada la eficiencia de algunos algoritmos de poda, (Tetko *et al.,* 1996). El método de validación tiene como objetivo evaluar la eficiencia de los enfoques, de los descriptores o métodos elegidos, que es posible que no solamente se utilicen para comparar diferentes herramientas, sino también para encontrar el rango de resultados óptimos para una técnica determinada. Hoy día

todavía se discuten entre los investigadores los métodos para la selección, clasificación y validación de datos, índices o descriptores (Böcker *et al.,* 2004).

Medidas de Similitud/Disimilitud

La disimilitud es generalmente aceptada como el complemento desimilitud (y viceversa), es decir,

$$D(A, B) = 1 - S(A, B) \quad (3)$$

Para que una distancia sea descrita como métrica debe tener las siguientes propiedades:

1. los valores de distancia deben ser cero o positivo, y la distancia de un objeto a sí mismo debe ser cero:

$$D_{A,B} \geq 0, \; D_{A,A} = D_{B,B} = 0$$

2. los valores de distancia deben ser simétricos:

$$D_{A,B} = D_{B,A}$$

3. los valores de distancia deben obedecer a la desigualdad triangular:

$$D_{A,B} \leq D_{A,C} + D_{C,B}$$

4. La distancia entre los objetos no idénticos debe ser mayor que cero:

$$A \neq B \iff D_{A,B} > 0$$

Una distancia que tiene sólo las tres primeras de estas propiedades se llama pseudométrica, y una que no tiene la tercera propiedad es no métrica.

Aunque un gran número de coeficientes de similitud y de distancia se han definido (a menudo redefinidos por diferentes autores), muchos de ellos están estrechamente relacionados entre sí. En algunos casos, el mismo coeficiente se puede obtener por diferentes vías, en otros casos donde los coeficientes son diferentes cuando se calculan para los atributos continuos, se igualan cuando se aplica a los atributos binarios. Ciertos coeficientes se describen como monotónicos entre sí, lo que significa que se puede demostrar analíticamente que siempre se producen clasificaciones de similitud idéntica de los objetos frente a un objetivo determinado, a pesar de que los valores del coeficiente real son diferentes. A pesar de que dos coeficientes pueden no ser completamente monótonos, los valores resultantes de su uso pueden presentar un alto grado de correlación, como lo

demuestra Holliday et al. (Holliday *et al.,* 1995), en una comparación de los coeficientes Coseno y Tanimoto. Algunos pares de coeficientes, por el contrario, presentan correlaciones muy bajas, lo que sugiere que reflejan características muy diferentes de los objetos que se comparan (Ellis *et al.,* 1994a). 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. En este contexto, una serie de símbolos útiles pueden ser definidos. Para los objetos A y B que se caracterizan por vectores **X** y **Y** que contienen *n* valores binarios (tales como huellas digitales) se puede escribir como:

$a = \sum_{i=1}^{n} x_i$, es el número de bits activos en A

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

$c = \sum_{i=1}^{n} x_i y_i$,es el número de bits activos en A y B

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

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

Las cantidades anteriores también se pueden expresar en notación de teoría de conjuntos, si definimos $X_A$como el conjunto de todos los elementos x_i en el vector **X** cuyo valores1 (los bits activos) y $X_B$como el conjunto de todos los elementos y_i en el vector **Y** cuyovalores1, entonces:

$a = |X_A|$, $b = |X_B|$, $c = |X_A \cap X_B|$, $d = n - |X_A \cup X_B|$

y, como corolario de lo anterior, el número debits activos al menos en una de las moléculas está dada por

$a + b - c = |X_A \cup X_B|$

Es habitual definir un "espacio químico" como un espacio con *n* descriptores que definen la posición de una molécula en un espacio químico *n*-dimensional; la diversidad y

la similitud de los compuestos es intuitivamente relacionada con la distancia inter-molecular como medida en el espacio. Los coeficientes de similitud (índice, distancia) son funciones que transforman los pares de representaciones moleculares compatibles en números reales, que por lo general se extiende al intervalo unidad. Dichos coeficientes proporcionan una medida cuantitativa de la sustancia química con un grado de semejanza (Pepperrell and Willett, 1991). 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 en las similitudes globales, la semejanza se mide entre dos objetos enteros (moléculas, proteínas, etc.).

Algoritmos de Búsqueda

En la igualdad ("macheo") exacta y parcial, o global, los algoritmos de búsqueda son ampliamente utilizados en sistemas de información química basados en computadoras con el fin de buscar una subestructura idéntica. Una facilidad menos común es la provisión para el mejor macheo, o vecino más cercano, que son búsquedas en las que la estructura o las estructuras más similares a una estructura de consulta se recuperan, 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 macheo es la base para la clasificación del *k-vecino más cercano* (*k-Nearest Neighbor*, KNN, en inglés) y juega un papel importante en el uso de árboles de expansión y técnicas de clasificación automática (Willett, 1983).

El problema general de encontrar los mejores macheos se define por Friedman *et al.* (Friedman *et al.,* 1977) como: «... dado un fichero de *m* muestras (cada uno de los cuales es descrito por *n* atributos con valores reales) y una medida de similitud/disimilitud, encontrar las *k* muestras más cercanas a la muestra 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 macheo es calcular la distancia entre la consulta y cada uno de las muestras 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 extremadamente grandes.

Existen varias heurísticas que, aunque no se ha reducido la complejidad de la búsqueda por debajo de *O(mn)*, son lo suficientemente potentes para permitir la búsqueda del vecino más próximo en los ficheros con estructura química con un costo computacional razonable. La mayoría de los experimentos sólo tienen en cuenta la recuperación del vecino más cercano para $k = 1$, pero existen procedimientos descritos en la literatura que son generalizados al problema de la búsqueda de los k vecinos más cercanos para el cual$k > 1$.

Un algoritmo eficiente del vecino más cercano será uno que evita el cálculo de la mayoría de las distancias, calculándose solamente las distancias de las pocas muestras que estén cerca de la estructura de la consulta. Existen varios tipos de criterios que se han sugerido para reducir el número de cálculos necesarios, incluyendo la proyección de las muestras d dimensional en un espacio de menor dimensión, de manera tal, que varias muestras puedan ser buscadas, o eliminadas desde una búsqueda, simultáneamente (Bentley *et al.,* 1980). Muchos de los algoritmos citados pueden no ser directamente aplicables a la búsqueda de los mejores macheos en contextos químicos, ya que asumen que los atributos son variables continuas, mientras que las estructuras químicas se caracterizan por fragmentos de descripción binaria. En este sentido, 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 muestras se encuentran en un espacio d-dimensional, donde d es pequeño, por lo general 2 o 3, por lo que los términos multiplicativos en d en la ecuación que describe el número de macheos que pueden pasarse por alto en un sistema de estructura química, d puede ser del orden de 10^2 o 10^3 (el número de bits en la cadena de bits), y los algoritmos son, por tanto, muy poco factibles. Por tanto, el procedimiento *O(nlog n)* debido a Friedman et al. (Friedman *et al.,* 1977) implica una constante de proporcionalidad alrededor de1.6^d, mientras que el método de búsqueda de Bentley *et a1*. (Bentley *et al.,* 1980) implica la inspección de todas las $3^d - 1$ celdas adyacentes a una celda dada en un espacio d-dimensional.

Los algoritmos de búsqueda citados a continuación están basados en representación binaria, tema en el cual se han enfocado mucho los investigadores de esta materia. Smeaton and Van Rijsbergen (Smeaton and Van Rijsbergen, 1981) tienen en cuenta que un archivo invertido puede ser utilizado para aumentar la eficiencia de la búsqueda del macheo de una

consulta en documentos donde tiene al menos un término en común. A partir de aquí ellos describieron experimentos con un procedimiento de límite superior que permite que la búsqueda del mejor macheo se termine antes de que todos los documentos en la lista de los ficheros invertidos correspondientes a la consulta hayan sido inspeccionados. Murtagh (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.

Van Marlen y Van den Hende (Van Marlen and Van Den Hende, 1979) y Rasmussen *et al.* (Rasmussen *et al.,* 1979) han descrito algoritmos de recuperación de los mejores macheos 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, mientras que otros autores han estudiado la búsqueda del mejor macheo en los sistemas de recuperación de información molecular (Chen *et al.,* 2010, Guyon and Elisseeff, 2003a, Willett, 1983).

Baldi *et al.* (Baldi *et al.,* 2008) plantea un algoritmo diferente a los demás descritos en este epígrafe, 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* $n = p$). La forma de obtener el vector $\vec{a}$ es aplicando el operador XOR (*exclusive* OR, siglas en inglés) al vector $\vec{A}$. Esta información adicional puede ser vista como un líder precediendo del vector $\vec{A}$, el cual puede ser usado para derivar los límites útiles en las medidas de similitud. Este enfoque es uno de los más eficientes y efectivos en la recuperación de información molecular con representaciones binarias.

1.4 Consideraciones finales del capítulo

Es evidente que el proceso tradicional de desarrollo y descubrimiento de un nuevo fármaco no es una vía eficiente ya que demanda recursos materiales y de tiempo abundantes. Una alternativa a este paradigma son las herramientas de cribado virtual, siendo una de las técnicas más simples de búsqueda de similitud, que contando solamente con un ordenador potente, un conjunto de datos químicos virtuales, una medida de similitud

y un algoritmo de búsqueda, es capaz de simular las etapas iniciales del cribado experimental de manera muy eficiente, lo cual conduce al descubrimiento de nuevos compuestos líderes en cortos períodos de tiempo, demandando así una inversión de recursos materiales que está al alcance de cualquier país en vías de desarrollo.

CAPÍTULO 2. MATERIALES Y MÉTODOS

En este capítulo se tratan primeramente los aspectos teóricos relacionados con los modelos de proximidad que aparecen en el presente trabajo, donde se destacan los modelos que están siendo aplicados por primera vez en Quimioinformática. Luego aparece la metodología para la comparación de los nuevos modelos con los reportados en la literatura, que incluye la elección de los conjuntos de datos, la representación molecular, el cálculo y selección automática de los descriptores, los experimentos y métricas de validación. Por último se trata la parte computacional del estudio, haciéndose énfasis en el análisis de eficiencia del algoritmo de búsqueda y de la ingeniería de software.

2.1 Similitud Molecular

El concepto de similitud es fundamental para varios aspectos del razonamiento y análisis químico, 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 (generalmente denominado *el problema de la comparación*). El resultado de este ejercicio es un valor, *la medida de similitud*, que caracteriza el grado de coincidencia, de asociación, proximidad, semejanza, alineamiento, 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. 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 (Tversky, 1977), sin embargo, ha argumentado persuasivamente que las comparaciones con ciertas similitudes son inherentemente asimétricas. Aunque su trabajo se orientó hacia la psicología, este tiene aplicabilidad además, en los estudios de similitud molecular (Chen and Brown, 2007). En una revisión relativamente reciente, Willett *et al.* (Willett *et al.,* 1998b) presentó un panorama general de muchas de las medidas de similitud que se usan en la actualidad. Su revisión incluyó una tabla que resume la forma de las diversas medidas en relación con el tipo de representación utilizado y se debe consultar para más detalles.

2.1.1 Medidas de Semejanza Molecular Reportadas

Tradicionalmente, los estudios de similitud molecular se han enfocado al uso de "huellas dactilares" (*fingerprints*, en inglés) que no son más que cadenas binarias que codifican la presencia (o no) de fragmentos moleculares (Geppert and Bajorath, 2010). El uso de las huellas dactilares se debe a la eficiencia con la que pueden generarse, compararse y almacenarse para conjuntos de datos muy grandes. Su adopción tuvo sus orígenes en los mismos comienzos de la Quimioinformática por la época de los sesenta, cuando los recursos computacionales estaban aun severamente limitados y los tamaños de los repositorios de datos aumentaban vertiginosamente en las compañías farmacéuticas*. Este enfoque estaba basado en la lógica de "eficiencia primero, luego efectividad". Sin embargo, los enormes recursos computacionales actuales le permite al investigador adoptar una línea de pensamiento contraria "efectividad primero, luego eficiencia" que se basa en el concepto de la información estadística y su relación con la *fortaleza de las escalas de medición*, i.e., absoluta, razón, aditiva, intervalo, ordinal y nominal. Cada vez que las mediciones hechas en escalas continuas (o discretas) se convierten a escalas ordinales, y estas a su vez, a escalas nominales, el proceso estará acompañado con una pérdida de información estadística y una disminución correspondiente en la potencia de los métodos (*e.g.* ANOVA de dos criterios de Fisher vs Friedman) (Siegel and Castellan, 1988). También, los modelos de proximidad pierden versatilidad *(e.g.,* degeneración en las métricas de Minkovski) y capacidad de resolución de ataduras en los objetos químicos cuando se aplica un proceso de transformación de escala de medición. Teniendo en cuanta lo anterior, la filosofía adoptada en nuestro trabajo fue "efectividad primero, luego eficiencia"**.

Las medidas tomadas como referencia de comparación en el presente estudio son las reportadas en el trabajo de Al Khalifa et al. (Al Khalifa *et al.,* 2009), que a su vez se basaron en la excelente revisión de Ellis *et al. (*Ellis *et al.,* 1994b). El conjunto resultante consiste en 12 medidas de proximidad cuyas fórmulas y clasificación se brindan en la Tabla 2.1.

*Esta explicación no aparece en ningún reporte científico regular, procede de un revisor anónimo de la revista Journal of Chemical Information and Modeling que arbitró un artículo del tutor principal de la tesis.

**Esta fue la respuesta brindada por el tutor principal de la tesis a dichos comentarios.

Tabla 2.1 Medidas de proximidad no binarias como referente de comparación

Medida[a]	Fórmula[b]	Tipo[c]	Ecuación No.[d]
Manhattan Media	$MM_{XY} = \frac{\sum_{j=1}^{n} \lvert x_j - y_j \rvert}{n}$	D	1
Euclidiana Media	$EM_{XY} = \frac{\sqrt{\sum_{j=1}^{n} \lvert x_j - y_j \rvert^2}}{n}$	D	2
Euclidiana Cuadrada Media	$ECM_{XY} = \frac{\sum_{j=1}^{n} \lvert x_j - y_j \rvert^2}{n}$	D	3
Bray/Curtis	$BC_{XY} = \frac{\sum_{j=1}^{n} \lvert x_j - y_j \rvert}{\sum_{j=1}^{n} (\lvert x_j \rvert + \lvert y_j \rvert)}$	D	4
Tan	$T_{XY} = \frac{\sum_{j=1}^{n} x_j y_j}{\sum_{j=1}^{n} x_j^2 + \sum_{j=1}^{n} y_j^2 - \sum_{j=1}^{n} x_j y_j}$	A	5
Dice	$D_{XY} = \frac{2 \sum_{j=1}^{n} x_j y_j}{\sum_{j=1}^{n} x_j^2 + \sum_{j=1}^{n} y_j^2}$	A	6
Fossum	$F_{XY} = \frac{n\left(\sum_{j=1}^{n} x_j y_j - \frac{1}{2}\right)^2}{\sum_{j=1}^{n} x_j^2 \sum_{j=1}^{n} y_j^2}$	A	7
Sokal/Sneath(1)	$SS1_{XY} = \frac{\sum_{j=1}^{n} x_j y_j}{2 \sum_{j=1}^{n} x_j^2 + 2 \sum_{j=1}^{n} y_j^2 - 3 \sum_{j=1}^{n} x_j y_j}$	A	8

Kulczynski(1)	$Kul1_{XY} = \frac{\sum_{j=1}^{j} x_j y_j}{\sum_{j=1}^{j} x_j^2 + \sum_{j=1}^{j} y_j^2 - 2\sum_{j=1}^{j} x_j y_j}$	A	9
Cosine/Ochiai	$Cos_{XY} = \frac{\sum_{j=1}^{j} x_j y_j}{\sqrt{\sum_{j=1}^{j} x_j^2 \sum_{j=1}^{j} y_j^2}}$	A	10
Simpson	$Sim_{XY} = \frac{\sum_{j=1}^{j} \min(x_j, y_j)}{\min\left(\sum_{j=1}^{j} x_j, \sum_{j=1}^{j} y_j\right)}$	A	11
Pearson	$r_{XY} = \frac{\sum_{j=1}^{j} (x_j - \bar{X})(y_j - \bar{Y})}{\sqrt{\sum_{j=1}^{j} (x_j - \bar{X})^2 \sum_{j=1}^{j} (y_j - \bar{Y})^2}}$	C	12

[b] $xj(y_j)$ representa el valor del descriptor de la molécula X (Y) en el atributo j; [c] 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) (Ellis *et al.,* 1994b). [d] Identificador usado a lo largo del trabajo.

2.1.2 Acuerdo Relacional como Medidas de Similitud Molecular Novedosas

Las medidas de similitud molecular propuestas en nuestro trabajo están basadas en la teoría Zegers y ten Berge (Zegers and ten Berge, 1985). Estos autores propusieron una fórmula general para los coeficientes de asociación bivariada correspondientes a las escalas métricas, y Zegers (Zegers, 1986) propuso una versión corregida por aleatoriedad de la fórmula general. Su teoría se basó en la premisa de que la elección de un coeficiente de asociación entre dos variables depende del tipo de escala de las variables, definida por la clase de transformaciones admisibles. Stine (Stine, 1989), extendió la teoría de Zegers-ten Berge a varias escalas adicionales no considerados por Zegers y ten Berge, y cambió el enfoque de "asociación" entre dos variables al de "acuerdo relacional" entre observadores. Stine (Stine, 1989) también demostró la utilidad de los conceptos de “relevancia” (*meaningfulness*) en la evaluación del acuerdo interobservador.

Tipo de Escala, Transformaciones de misibles y Acuerdo Relacional

La Tabla 2.2 recoge cuatro escalas métricas consideradas por Zegersyten Berge (Zegers and ten Berge, 1985), y dos escalas métricas adicionales propuestas por Stine (Stine, 1989), junto con la transformación admisible (de definición) para cada una de las mismas. Por ejemplo, la escala desazón sólo permite la multiplicación por una constante positiva (cambio de unidad), la escala de intervalo permite transformaciones lineales (afines) positivas(cambio de unidad y del origen), la escala aditiva sólo permite la traslación(constante aditiva) y la escala absolutano permite transformación de la escala original.

Nota: Es justo señalar que el nombre de escala "aditiva" fue utilizada por Zegersyten Berge, 1985 (Zegers and ten Berge, 1985) y seguidamente por Stine (Stine, 1989), pero anteriormente fue llamada la escala de "diferencia" por Suppes y Zinnes (Suppes and Zinnes, 1963).

Tabla 2.2 Escalas métricas y sus transformaciones admisibles

Escala[a]	Transformación Admisible[b]
Absoluta	$\varphi' = \varphi$
Aditiva	$\varphi' = \varphi + \alpha$
Razón	$\varphi' = \beta\varphi$
Intervalo	$\varphi' = \beta\varphi + \alpha$
razón-log	$\varphi' = \varphi^{\gamma}$
Intervalo-log	$\varphi' = \beta\varphi^{\gamma}$
ordinal	$\varphi' = f(\varphi)$

[a]Las primeras cuatro escalas están incluidas en el trabajo de Zegers y ten Berge (Zegers and ten Berge, 1985), la quinta aparece en el trabajo de Stine (Stine, 1989); [b]φ y φ' son funciones representantes, α y γ son números reales, β es un número real positivo y f es una función estrictamente monótona

La base para el rol central del tipo de escala en la construcción de un coeficiente de acuerdo inter observador es precisamente la diferencia entre las escalas en cuanto a las transformaciones admisibles, y por tanto, en el número y tipo de factores de escala arbitrarios. Por ejemplo, supongamos que el conjunto de pares ordenados (1,16), (2,32), (3,48) indican las evaluaciones de dos jueces sobre tres objetos. Parecería ser que existe un acuerdo pobre entre los jueces, ya que los puntajes para cada objeto no son iguales. Pero si las evaluaciones se refieren a las masas de los objetos con la evaluación (medición) de primer juez (balanza 1) en libras y el segundo (balanza 2) en onzas, entonces los "jueces" están en perfecto acuerdo en relación a la escala de razón. En otras palabras, según esta teoría, una condición suficiente para el acuerdo perfecto sobre una escala de razón es la *proporcionalidad* de las evaluaciones de los jueces. De forma general, *acuerdo perfecto* significa que las evaluaciones de los jueces están relacionadas por una transformación admisible con respecto al tipo de escala, y el coeficiente de acuerdo debe estimar el grado en que esta relación se cumple.

Estas ideas están plasmadas en el concepto de *acuerdo relacional* de Stine (Stine, 1989) que denota el acuerdo con respecto a las *relaciones empíricamente significativas*. En cualquier aplicación pueden encontrarse a la vez el desacuerdo con y sin significado. Así, en el ejemplo anterior, si las evaluaciones se hacen en una escala de razón, entonces la proporcionalidad en los puntajes de los jueces representa un desacuerdo sin significado (lo que refleja las diferencias en las unidades permitidas por una escala de razón), pero un desacuerdo aditivo representa un desacuerdo con significado (ya que el origen se fija para una escala de razón). La idea es que un coeficiente de acuerdo relacional debe variar con (ser atenuado por) discrepancias con significado, pero debe ser independiente de discrepancias sin significado.

Nota: Las ideas modernas sobre las relaciones significativas se deben a Stevens, (ver por ejemplo (Stevens, 1951)); Stine (Stine, 1989b) proporciona una buena introducción al tema de la inferencia significativa).

Teoría de Zegers-ten Berge

Zegers y ten Berge (Zegers and ten Berge, 1985) propusieron una fórmula general para los coeficientes de asociación entre dos variables para las escalas métricas que figuran en la **Tabla 2.2**. Ellos requirieron que el coeficiente de asociación fuera invariante bajo

transformaciones admisibles de la escala y sensible a las transformaciones no admisibles. La clave para el desarrollo de la fórmula general fue el concepto de *versión uniformadora* de las variables, que cumple las condiciones de *invariancia* bajo transformaciones admisibles y de *sensibilidad* a las transformaciones no admisibles. Una transformación uniformadora es un miembro de la clase de transformaciones admisibles para un tipo de escala dada. Sea $x_i$una evaluación de la variable *i-ésima* antes de la transformación y U_i la versión uniformadora. Entonces, las transformaciones uniformadoras son:

$$U = X \quad \text{(escala absoluta)} \qquad 13$$

$$U = X - \bar{X} \quad \text{(escala aditiva)} \qquad 14$$

$$U = \frac{X}{T_X} \quad \text{(escala de razón)} \qquad 15$$

$$U = \frac{(X - \bar{X})}{S_X} \quad \text{(escala de intervalo)} \qquad 16$$

$$U = X^{\frac{1}{L_X}} \quad \text{(escala razón-log)} \qquad 17$$

$$U = \left(\frac{X}{G_X}\right)^{\frac{1}{N_X}} \quad \text{(escala intervalo-log)} \qquad 18$$

$$U = Rank(X) \quad \text{(escala ordinal)} \qquad 19$$

donde, $\bar{X}$ y S_x son la media aritmética y la desviación estándar de la variable (juez) *X*,

$$T_X = +\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}, \quad L_X = +\sqrt{\frac{\sum_{i=1}^{n} [\ln(x_i)]^2}{n}}, \quad G_X = \sqrt[n]{\prod_{i=1}^{n} x_i}, \quad N_X = \sqrt{\frac{\sum_{i=1}^{n} \left[\ln\left(x_i / G_X\right)\right]^2}{n}},$$

Rank representa la transformación RT-2 de Conover e Iman (Conover and Iman, 1981) y *n* es el número de puntajes para cada objeto *i*.

La transformada aditiva (Ec. 14) centra la variable en cero, la transformación multiplicativa (Ec. 15) reescala la variable para obtener un valor cuadrado medio de uno, la transformación lineal (Ec. 16) es la conocida transformada de "estandarización" Z (con media cero y desviación estándar de uno), la transformada de potencia (Ec. 17) es la

versión análoga de la escala de razón, y la transformada de potencia (Ec. 18) es la versión logarítmica de la escala de intervalo. La idea detrás de las transformaciones uniformadoras es que después de aplicarse tales transformadas apropiadas a X_i, entonces si las versiones uniformadas se igualaran, existe un acuerdo perfecto entre los jueces, y el coeficiente de asociación estima el grado en que las versiones uniformadas concuerdan (en valor numérico). Por ejemplo, si el conjunto de pares ordenados (1,2), (2,4) y (3,6) denota las evaluaciones (proporcionales) de dos jueces para tres objetos en una escala de razón, entonces aplicando la transformación (Ec. 15), $T_1 - 2.16$, $T_2 - 4.32$ y las versiones uniformadas son (0.463, 0.463), (0.926, 0.926) y (1.389, 1,389), lo que indica un acuerdo perfecto en esa escala. Aunque Zegers y ten Berge no exponen sus resultados formalmente en términos de la teoría del significado, está claro que las versiones uniformadas tienen el efecto de asegurar que los coeficientes varíen con el desacuerdo de significado, pero son independientes de desacuerdo sin significado (relativo al tipo de escala).

Sobre la base de una función de diferencia cuadrada media de las versiones uniformadas, Zegers y ten Berge derivaron su fórmula general de los coeficientes de asociación para escalas métricas (g_{XY}) entre dos versiones uniformadas U_X y U_Y:

$$g_{XY} = \frac{2\sum_{j=1}^{n} U_{jX}U_{jY}}{\sum_{j=1}^{n} U_{jX}^2 + \sum_{j=1}^{n} U_{jY}^2} \qquad (20)$$

Zegers (Zegers, 1986) obtuvo luego una fórmula general g'_{ij} para el caso bivariado, corrigiendo la Ec. 5 para el *acuerdo casual* a través de la ecuación:

$$g'_{XY} = \frac{(g_{XY})_{obs} - (g_{XY})_{cas}}{(g_{XY})_{max} - (g_{XY})_{cas}} \qquad (21)$$

donde, $(g_{XY})_{obs}$, $(g_{XY})_{cas}$, $(g_{XY})_{max}$ son los valores observado, esperado y máximo, correspondientemente, del coeficiente g_{XY}.

Luego de consideraciones estadísticas y alguna manipulación algebraica, este autor obtiene finalmente:

$$g'_{XY} = \frac{2\left(\sum_{j=1}^{n} U_{jX}U_{jY} - n^{-1}\sum_{j=1}^{n} U_{jX}\sum_{j=1}^{n} U_{jY}\right)}{\sum_{j=1}^{n} U_{jX}^2 + \sum_{j=1}^{n} U_{jY}^2 - 2n^{-1}\sum_{j=1}^{n} U_{jX}\sum_{j=1}^{n} U_{jY}} \qquad (22)$$

Sustituyendo la transformación uniformadora adecuada en la fórmula para s_{XY} y s'_{XY}, se pueden obtener los coeficientes de asociación para cada una de las escalas mostradas en la Tabla 2.1 derivándose las expresiones siguientes:

Medidas Reportadas en el Trabajo de Zergers y ten Berge

Identidad

No corregido: es igual al coeficiente de Dice (ver Ec. 6, Tabla 2.1)

Corregido

$$s'_{XY} = \frac{2s_{XY}}{s_X^2 + s_Y^2 + (\overline{X} - \overline{Y})^2} \tag{23}$$

Aditividad

No corregido (es el caso especial del "punto de anclaje" de Winer (Winer, 1971))

$$a_{XY} = \frac{2s_{XY}}{s_X^2 + s_Y^2} \tag{24}$$

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

Corregido (no cambia la expresión)

Proporcionalidad

No corregido: es igual al coeficiente Cosine/Ochiai (ver Ec. 10, Tabla 2.1)

Corregido

$$p'_{XY} = \frac{s_{XY}}{T_X + T_Y - \overline{XY}} \tag{25}$$

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. 12, Tabla 2.1)

Corregido (no cambia la expresión)

Medidas Reportadas en el Trabajo de Stine

Log-razón o Log-proporcionalidad

No corregido

$$Lp_{XY} - \frac{2\sum_{j=1}^{n} x_j^{1/L_X} y_j^{1/L_Y}}{\sum_{j-1}^{n} x_j^{2/L_X} + \sum_{j-1}^{n} y_j^{2/L_Y}} \quad (26)$$

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

Corregido

$$^{1}/_{L} \quad (27)$$

Log-intervalo o Log-linealidad

No corregido

$$Lr_{XY} = \frac{2\sum_{j=1}^{n} x_j^{1/N_X} y_j^{1/N_Y}}{\left(G_Y^{1/N_Y} / G_X^{1/N_X}\right)\sum_{j=1}^{n} x_j^{2/N_X} + \left(G_X^{1/N_X} / G_Y^{1/N_Y}\right)\sum_{j=1}^{n} y_j^{2/N_Y}} \quad (28)$$

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

(29)

Ordinal

No corregido

$$r_{XY}^{R} = \frac{2\sum_{i=1}^{n} R(x_i)R(y_i)}{\sum_{i=1}^{n} R(x_i)^2 + \sum_{i=1}^{n} R(y_i)^2}$$

(30)

Corregido (es el mismo que la ρ de Spearman)

$$r_{XY}^{Rc} = 1 - \frac{6\sum_{i=1}^{n}[R(x_i) - R(y_i)]^2}{n(n^2 - 1)} \qquad (31)$$

Finalmente, en el presente trabajo se proponen doce medidas de similitud novedosas en Quimioinformática, y nueve basadas en los conceptos de *asociación generalizada* y *acuerdo relacional* (Ec. 23-31). En nuestro estudio establecemos un nexo entre dichos conceptos y la similitud molecular. Por analogía, consideramos que la naturaleza de la escala métrica (aunque no se conozca explícitamente *a priori*) estará dada por la naturaleza del desarrollo teórico-experimental de los descriptores; que los "jueces" representan en nuestro caso "moléculas" y que los "objetos" representan el vector de descriptores de representación; y que entre las moléculas existen relaciones de similitud absoluta, aditiva, de proporción, de intervalo, etc., que estarán reflejadas para cada conjunto de datos por el coeficiente más efectivo en dicha recuperación.

2.2 Conjuntos de Datos para Validación

La medición de la efectividad de los índices de similitud, descriptores moleculares, e incluso enfoques de validación, es estrictamente dependiente de las moléculas/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 de esteroides, el conjunto de datos 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 conjunto de datos estándar para la comparación de medidas de similitud y descriptores moleculares, probablemente por la imposibilidad de encontrar un único conjunto de moléculas que reagrupe todas las necesidades de cribado de la Quimioinformática moderna (Maldonado *et al.,* 2006). Por este motivo, como se ha comentado antes, 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 (Sheridan and Kearsley, 2002a). Una revisión exhaustiva acerca de las bases de datos empleadas actualmente en la Quimioinformática, haciendo énfasis en las bases de datos farmacológicas se puede encontrar en (Jónsdóttir *et al.,* 2005). La tendencia actual de dichos repositorios es pasar al dominio público (Bender, 2010).

Para nuestro estudio, se seleccionaron nueve conjuntos de datos procedentes de la Química Medicinal. Los primeros ocho repositorios fueron usados originalmente por Sutherland *et al.* (Sutherland *et al.,* 2004) en un estudio QSAR comparativo entre modelos de Cuadrados Mínimos Parciales y Redes Neuronales con propagación hacia atrás, basados en descriptores moleculares 2-D y hasta 4-D. Los mismos fueron extraídos de otros estudios QSAR precedentes y de estudios farmacológicos experimentales, garantizando en todas las colecciones, una alta representatividad de las clases químicas y una alta diversidad molecular de los elementos respectivos. La descripción general de estos conjuntos se muestra en la Tabla 2.3.

Tabla 2.3 Conjuntos de datos de la Química Medicinal empleados en la presente investigación

CD[a]	Diana farmacológica	Número de compuestos	Variable farmacocinética	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 (Sutherland *et al.,* 2004). 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, pKi = - logKi, donde Ki representa la constante de inhibición del fármaco, también se usa como una medida de la potencia del fármaco.

Cabe señalar que estos conjuntos de datos también fueron empleados más recientemente por otros investigadores en estudios QSAR que emprendieron la comparación de multiclasificadores basados en modelos de Árboles y Reglas de decisión, Máquinas de Soporte Vectorial y Redes Neuronales(Bruce *et al.,* 2007, Johansson et al., 2009, Sönströd et al., 2009, Culp *et al.,* 2010).

El noveno caso de estudio estuvo basado en el programa de cribado antiviral frente al SIDA del Instituto Nacional del Cáncer (NCI's AIDS *antiviral screen*, en inglés). El NCI es la agencia principal del gobierno federal de los Estados Unidos para la investigación y capacitación del cáncer, que se ha convertido en la organización más prestigiosa de investigación sobre el cáncer en el mundo. El Instituto Nacional del Cáncer forma parte de los Institutos Nacionales de la Salud (*National Institutes of Health*, NIH, en inglés) los cuales son una dependencia del gobierno federal de los Estados Unidos.

Estos experimentos fueron desarrollados como un esfuerzo para descubrir nuevos compuestos capaces de inhibir el virus VIH. El desarrollo y mantenimiento actual del cribado es responsabilidad de la Rama de Tecnologías de Cribados (*Screening Technologies Branch*, STB, en inglés,) que pertenece al Programa de Terapéuticas en Desarrollo (*Developmental Therapeutics Program*, DTP, en inglés). El cribado utilizó un ensayo de Formazán soluble para medir la protección de las células CEM (línea de células de Linfocitos T) humanas de la infección del VIH-1 (Weislow et al., 1989). Como resultado, los compuestos capaces de brindar al menos 50 % de protección a las células CEM fueron reevaluados. Los compuestos que brindaron al menos un 50 % de protección en la reevaluación fueron listados como "moderadamente activos" (CM). Los compuestos que de forma reproducible brindaron 100 % de protección fueron listados como "activos confirmados" (CA). Los compuestos que no satisfacieron estos criterios anteriores fueron listados como "inactivos confirmados" (CI). En la página web http://dtp.nci.nih.gov/docs/aids/aids_data.html aparecen disponibles gratuitamente los resultados del cribado y los datos estructurales químicos de compuestos que no están protegidos por un acuerdo de confidencialidad. La información experimental más reciente (Mayo 2004), se puede descargar como un fichero ASCII delimitado por comas que contiene los resultados del cribado para 43850 compuestos; aparece también para descargar

el fichero ASCII de datos estructurales en forma de fichero compacto, con 260071 compuestos químicos en formato MDL SD.

El interés en usar los datos del NCI's AIDS *Antiviral Screen* (NAAS) en el presente estudio quimioinformático es coherente con los estudios de modelización realizados en el Laboratorio de Bioinformática enfocado a las mutaciones en el genoma viral del VIH-1 responsables de la resistencia a los fármacos, específicamente en los genes que codifican las dianas moleculares de la terapia de fármacos, las enzimas reverso transcriptasa y proteasa. Adicionalmente, este conjunto de datos presenta características especiales arrojadas por un estudio comparativo relativamente reciente entre la Base de Datos Abierta del NCI (*Open NCI Database*) con varias de las bases de datos comerciales y medicinales internacionales más reconocidas como *Available Chemicals Directory* (ACD-MDL), *Sigma-Aldrich Catalog* (SIGALCAT),*WorldDrug Index* (WDI), *Cambridge CrystallographicDatabase* (CCD), mostró que la misma no solo es atractiva por ser gratuita, sino que además posee un grado de solapamiento relativamente bajo con las demás bases de datos, y más importante, que posee el mayor número de estructuras únicas (al menos 200000) de entre todas las colecciones estudiadas (Voigt et al., 2001).

2.3 Representación Molecular y Obtención de las Matrices de Datos

El procedimiento seguido para la representación abstracta y obtención de los datos numéricos de los ocho primeros conjuntos de datos (ver Tabla 2.3) fue el mismo empleado en un estudio anterior por el autor y tutor principal (O. M. R-B) de esta tesis (Rivera Borroto *et al.,* 2011a). De modo concreto los pasos secuenciales fueron:

Los conjuntos de datos fueron manipulados y editados con la utilidad *JChemfor Excel* ChemAxon, 2010); cada uno de los mismos fue reoptimizado con el software generador de estructuras 3D CORINA (Sadowski et al., 1994); con el objetivo de "estandarizar" las bases de datos, los parámetros más relevantes del software fijados en este proceso fueron: wh, escribir átomos de hidrógenos adicionados; rs, eliminar pequeños fragmentos desconectados; neu, neutralizar cargas formales. Los ficheros de salida fueron cargados en el software para el cálculo de descriptores moleculares DRAGON (Talete srl, 2007); en esta etapa se calcularon todas las familias de descriptores moleculares disponibles (un total de 3224 descriptores), y entonces los rasgos binarios fueron eliminados. Los ficheros resultantes fueron cargados en el software de minería de datos Weka (Hall *et al.,* 2009), los

cuales fueron sujetos a un tratamiento que incluyó prefiltrado, reescalado, y selección de rasgos; en esta etapa, los atributos nominales (los no disponibles) fueron eliminados con el filtro *RemoveType*, los atributos que no varían o varían demasiado fueron eliminados con el filtro *RemoveUseless* manteniendo los parámetros de la herramienta por defecto; los atributos numéricos resultantes, así como el atributo de clase fueron estandarizados para que tuvieran media nula y desviación estándar unitaria con el filtro *Standardize*. La selección de rasgos fue llevada a cabo usando el filtro *Atribute Selection*; dentro de las opciones presentes, el evaluador *CfsSubsetEval* fue elegido con el resto de los parámetros por defecto; este evaluador aplica una selección de rasgos basada en correlación de modo que el subconjunto de rasgos estén altamente correlacionados con la clase mientras que posean una baja intercorrelación entre ellos (Hall, 1998).

Más específicamente, la medida de heurística que emplea este evaluador es:

$$G_{ext,S} = \frac{k * \overline{r_{ext,i}}}{\left(k + k * \overline{r_{i,j}}\right)} \qquad (32)$$

donde: $G_{ext,S}$ es la medida de correlación global de la variable externa *ext* con el subconjunto de descriptores seleccionados *S*, $\overline{r_{ext,i}}$ es el promedio de la correlación de las combinaciones de la variable externa *ext* con cada uno de los *k* descriptores *i* del subconjunto *S*, y, $\overline{r_{i,j}}$ es el promedio de la correlación de las combinaciones de a dos entre los *k* descriptores del subconjunto *S*.

A partir de la relación anterior (ver Ec.1) se pueden derivar algunas características:

- Mientras más altas sean las correlaciones por pares entre descriptores y la variable externa ($\overline{r_{ext,i}}$), más alta será la medida de correlación global ($G_{ext,S}$).
- Mientras más bajas sean las correlaciones por pares entre los descriptores ($\overline{r_{i,j}}$), más alta será la medida de correlación global ($G_{ext,S}$).
- A medida que el número de componentes en *S* incrementa (asumiendo que los descriptores adicionales son los mismos que los componentes originales en término de su inter correlación con otros componentes y la variable externa), la medida de correlación global incrementa.

La medida de correlación por pares (*r*) que emplea la Ec. 1 es el *Coeficiente de Incertidumbre Simétrico* (*CIS*) para variables nominales (Press *et al.,* 1988), cuya expresión matemática está dada por:

$$r = CIS = 2 * \left[\frac{GI}{H(Y) + H(X)}\right] \quad (33)$$

con:

$$GI = H(Y) - H(Y|X) \quad (34)$$

En las expresiones anteriores, *GI* representa la ganancia de información de Quinlan cuya expresión viene dada por la Ec. 34 y refleja la cantidad de información adicional acerca de la variable *Y* que no es capaz de explicar la variable de inducción *X* , o simplemente, la cantidad de información ganada acerca de *Y* luego de observar *X* (Quinlan, 1986). Para propósitos de selección de rasgos, *Y* representa la variable nominal externa (clasificación de cribado) y *X* los rasgos a seleccionar (descriptores moleculares); por otra parte *H(Y)* y *H(Y|X),* representan la entropía de *Y* y la entropía de *Y* luego de observar *X* o *entropía condicional*, respectivamente, cuyas expresiones matemáticas pueden encontrarse en cualquier libro estándar de estadística. Cuando los rasgos de partida son continuos (descriptores continuos), el algoritmo aplica una etapa de pre procesamiento para convertir rasgos continuos a nominales usando el método de discretización de Fayyad e Irani (Fayyad and Irani, 1993).

Es justo señalar que la ganancia de información de Quinlan (*GI*) había sido usada en un estudio comparativo de métodos de búsqueda de similitud para la selección de rasgos binarios, lográndose resultados superiores en las medidas de la calidad de recuperación de información química (Bender *et al.,* 2004a). Esta medida tiene la propiedad de ser simétrica en las variables, lo cual es una característica deseable en medidas de correlación por pares, pero desafortunadamente estás es gada a favor de los rasgos con mayor número de valores. Además, las correlaciones en la Ec. 33 deben ser normalizadas para asegurar que son comparables y tienen el mismo efecto. En este sentido, el *CIS* (ver Ec. 33) compensa los sesgos de *GI* hacia atributos con más valores y normaliza su valor en el rango [0,1] (Hall, 1998). En nuestro criterio, este evaluador también es importante para la búsqueda de similitud porque garantiza que las moléculas recuperadas al principio de la lista (mas

similares a la consulta) tengan una alta probabilidad de poseer también las mismas propiedades farmacológicas que la molécula de referencia, esto se debe a que solamente los descriptores linealmente relacionados con la clase externa cumplen con el *principio de similitud o vecindad*(Nikolova and Jaworska, 2003a). Adicionalmente, este evaluador ha sido utilizado por otros investigadores sobre los mismos conjuntos de datos con resultados relativamente buenos en la exactitud de los clasificadores comparados (Johansson *et al.,* 2009 , Sönströd *et al.,* 2009)

2.4 Métricas de Rendimiento

Existe un debate en curso en la literatura sobre "puntajes de mérito" adecuados (o indicadores de rendimiento) para evaluar los ensayos de cribado virtual retrospectivos. Unamétrica popular es el"factor de enriquecimiento", que es intuitivo y sencillo de interpretar. Un problema asociado con el cálculo de los factores de enriquecimiento simple es la dependencia de un valor de corte elegido, por lo general el 1 o5 % de la base de datos para cribado. Nicholls (Nicholls, 2008) aboga firmemente por el uso de medidas estándares, incluyendo la curva de la Característica en Operación del Receptor (*Receiver Operating Characteristic*, ROC, en inglés) y el área bajo la curva AUC[ROC] (Witten and Frank, 2005), que se aplican habitualmente en otros campos que emplean el análisis estadístico, minería de datos, olas técnicas de aprendizaje automático. Sin embargo, Truchon y Bayly (Truchon and 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 (*Boltzmann-Enhanced Discrimination* of ROC, BEDROC, en inglés), que utiliza una ponderación exponencial para asignar mayor peso a la detección temprana. Esta

medida es esencialmente una versión normalizada de la medida Mejora Inicial Robusta (*Robust Initial Enhancement*, RIE, en inglés) (Sheridan et al., 2001). Del mismo modo, se ha sugerido el escalado semilogarítmico de la ROC, pROC (Clark and Webster-Clark, 2008). Sin embargo, Nicholls (Nicholls, 2008) también presenta evidencias de una fuerte correlación entre el AUC[ROC]yAUC[BEDROC], lo que sugiere aAUC[ROC] como una medida suficiente para evaluar la eficiencia de cribado virtual. Este mismo autor recomienda que 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].

Curva ROC Concentrada

Basados en la idea de Nicholls (Nicholls, 2008), aunque no lo manifiestan explícitamente ni lo citan, Swamidass *et al.* (Swamidass *et al.,* 2010) proponen la curva ROC Concentrada (*Concentrated* ROC, CROC, en inglés) 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. 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 que 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 (*Concentrated Acumulation Curve*, CAC, en inglés), pROC y ROC.

Teniendo en cuenta el análisis anterior, en nuestro trabajo decidimos usar la métrica de validación AUC[CROC] con una transformación de magnificación exponencial del eje x (*fpr*) de la ROC dada por:

$$h(x) = \frac{1 - e^{-\alpha x}}{1 - e^{-\alpha}} \quad (35)$$

donde, α es el factor de magnificación, que en nuestro caso toma el valor $\alpha = 20$,que corresponde aproximadamente a un 8 % de enriquecimiento (Truchon and Bayly, 2007).

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* correspondientes a los rangos de las instancias positivas como:

$$AUC[CROC] = \frac{\sum_{i=1}^{n}[1 - h(fpr_i)]}{n} \quad (36)$$

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 fórmula:

$$AUC[CROC]_{rand} = \frac{1}{\alpha} - \frac{e^{\alpha}}{1 - e^{-\alpha}} \quad (37)$$

Para nuestro estudio, donde $\alpha = 20$ la métrica del clasificador aleatorio toma el valor $AUC[CROC]_{aleat} = 0.2809$.

2.5 Diseño Experimental y Análisis Estadístico

La serie de experimentos fue planificada similarmente al trabajo de Swamidass *et al.* (Swamidass *et al.,* 2010). En esta ocasión se utilizó el conjunto de datos NAAS descrito anteriormente. El conjunto de datos consiste en 42 131 compuestos, donde los compuestos "activos confirmados" (CA) y "moderadamente activos" (CM) se han agrupado en la categoría "**1**" (activos), y los "inactivos confirmados" (CI) se han agrupado en la categoría "**0**" (inactivos), resultando en un conjunto de datos 1 556 activos y 40 575 inactivos. Las métricas de rendimiento fueron calculadas empleando experimentos de validación cruzada de 10 pliegues (10-*fold cross validation*, 10-CV, en inglés) estándar (Efron and Tibshirani, 1993), empleando solamente subconjunto de activos. El clasificador empleado para comparar las medidas de similitud fue MAX-SIM (Hert *et al.,* 2004a, Hert *et al.,* 2005). Básicamente, el algoritmo MAX-SIM (máxima similitud) es uno de los métodos más simples para el cribado virtual masivo 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 $\{x_1, x_2, \ldots, x_q\}$, el puntaje asignado a una molécula del conjunto de datos x_m viene dado por:

$$z(x_m) = max\left[_1^q \{S(x_m, x_i)\}\right] \quad (38)$$

Donde, $S(x_m, x_i)$ es la similitud de la molécula del conjunto de datos x_m a la referencia x_i de la multi consulta, *S* es la función de similitud y para este estudio específico se decidió comparar las funciones de similitud dadas por las ecuaciones Ec. 1-12 y 23-31. Como se indicó anteriormente, para evaluar la calidad de la recuperación temprana, en este estudio se usó la métrica AUC[CROC], 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* con la corrección de Welch (Sawilowsky, 2002).

2.6 Diseño e Implementación Computacionales

En Quimioinformática se suele trabajar con conjuntos de datos muy grandes, por lo tanto el algoritmo de fuerza bruta, cuando se quiere encontrar los *k* vecinos más cercanos es muy costoso y no es recomendable aplicarlo a la recuperación de información molecular. Varios algoritmos de búsqueda de similitud se han presentado en la literatura los cuales son eficientes para encontrar los compuestos líderes (ver **Capítulo 1**, subsección ***Algoritmos de Búsqueda***). Estos algoritmos son eficientes para un valor *k* relativamente pequeño fijado por el usuario; no obstante para $k = N$, siendo *N* el tamaño del conjunto de datos, se evaluarían todos los compuestos del conjunto de datos al igual que en el algoritmo de fuerza bruta; el autor de esta tesis opina que incluso resultaría más costoso, ya que el algoritmo KNN planteado en la literatura requiere la construcción de un árbol *k*-d, que es un árbol binario generalizado, donde los nodos terminales representan los posibles *k* vecinos más cercanos y la unión de todos estos nodos conforman el conjunto de datos completo. Es evidente que cuando $k = N$, el árbol va a contener un solo nodo, que es a su vez un terminal y se hacen los mismos cálculos que en fuerza bruta. El tiempo de ejecución del algoritmo de fuerza bruta va a ser prácticamente el mismo que el de Friedman, pero la complejidad espacial si será mucho menor (Willett, 1983).

El interés de nuestro trabajo en ordenar los conjuntos de datos completos radica en accederal comportamiento global de las medidas de (dis)similitud con respecto a su selectividad por los compuestos de interés o activos, lo cual será cuantificado a través de las curvas CROC.

Por todo lo explicado anteriormente se decidió implementar el algoritmo de fuerza bruta paralelizado (es decir, que el conjunto de datos se dividió en 8 partes, cada parte va a tener *N*/8 elementos aproximadamente, de forma tal que se puedan utilizar 8 hilos, donde cada hilo se ocuparía de hacer los cálculos pertinentes de la parte del conjunto de datos que le corresponda). El pseudocódigo de este algoritmo es el siguiente:

Step 1: load(File dataset) (size N)

Step 2: build(File query) and div(File query in 10 fold) (size M)

Step 3: for i=1 until cMSelected (# of selected measures)

For each element of dataset

distance $_i$ (element, fold)

clasifMAXSIM(element, fold)

Step 4: quicksort(array of compound), End

En el Paso 1 se carga el conjunto de datos con *N* elementos, en el Paso 2 se construye el fichero consulta conformado por los compuestos activos de la base de datos que se cargó en el paso 1 y se divide en 10 pliegues para aplicar validación cruzada. El Paso 3 se encarga de calcular cada distancia seleccionada para cada elemento del conjunto de datos con cada elemento de cada pliegue de consulta, aplicándose adicionalmente el clasificador MAX-SIM, y guardándose los resultados en un arreglo para que finalmente en el Paso 4 se ordene y a partir del mismo imprimir los resultados.

Análisis de la Complejidad Temporal

A simple vista este algoritmo puede tener una complejidad temporal de $O(N^3)$, pero cuando se analiza detalladamente resulta no ser así. Esto se debe a que el Paso 1 tiene una complejidad de *O(N),* el paso 2 tiene una complejidad*O(M),* en el paso3 la complejidad sería *O(c10(N*M*f + N*M))*donde c es la cantidad de medidas seleccionadas, 10 es la

cantidad de pliegues, *N*M*f* se obtiene de calcular la distancia de cada elemento del conjunto de datos con cada elemento de la consulta, siendo *f* una constante de la fórmula correspondiente a la medida seleccionada, *M*N* es de aplicar el clasificador MAX-SIM de cada elemento a la consulta y el paso 4 es un*O(N*logN)*. Desde un primer momento la complejidad temporal es de *O(N + M +c10(N*M*f + N*M) + N*logN),* pero al aplicar propiedades de complejidad temporal tenemos que *O(N + M +c10(N*M*f + N*M) + N*logN)=O(max(N,M +c10(N*M*f + N*M) + N*logN))=O(M +c10(N*M*f + N*M) + N*logN).* Es evidente que aplicando esta propiedad sucesivamente llegamos a la conclusión que la complejidad temporal es igual a*O(N*M*f).*

Descripción de la Aplicación

El algoritmo anteriormente descrito se implementó en el lenguaje de programación Java, para ello nos apoyamos en la construcción de varias clases, las cuales daremos una breve descripción de cada una:

Class Auxiliar.java: Aquí están implementados todos los métodos auxiliares, necesarios para realizar cualquier operación, que es a su vez, utilizada en las diferentes clases.

Class Group.java: Esta clase se encarga de aplicar la estrategia fusión de grupo mediante el método groupFusion pasándole como parámetros principales el nombre del fichero de datos y el nombre de la medida que se quiere aplicar.

Class Distance.java: Permite calcular las diferentes medidas propuestas en nuestro trabajo, la cual está conformada por 21 métodos que responden a la fórmula de su medida correspondiente.

Class Compound.java: Recoge los datos principales de un compuesto como es el nombre, la distancia con respecto a la estructura de referencia que se está consultando en un momento determinado y se guarda además la clase a la cual pertenece.

Class Fuse10View.java: Es la clase principal, la cual se encarga de recoger, para poder obtener los resultados finales, todos los datos pertinentes, como son: la ubicación del fichero de datos que se va a analizar, el camino donde se van a guardar los resultados y las medidas a aplicar. Para más detalles ver el diagrama de clases.

Diagrama de Clases

El diagrama de clases fue realizado en el software *Visual Paradign Suite for UML Enterprise Edition v3.0*, nombre en inglés. A partir de aquí podemos ver las diferentes clases conformadas para la realización de nuestra aplicación, el mismo está dado por la Figura 2.1.

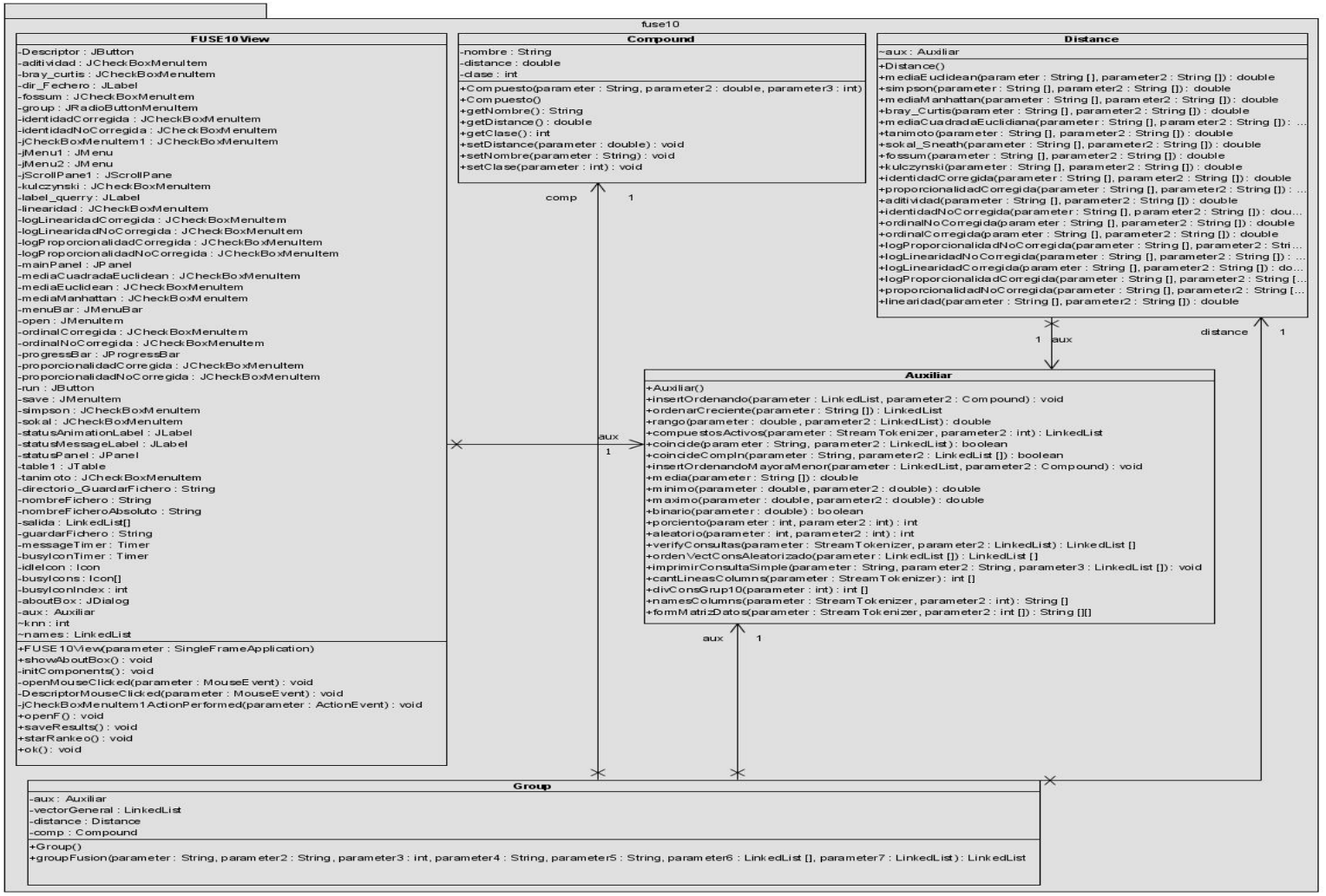


Figura 2.1 Diagrama de Clases UML

Interfaz Visual

A continuación se presenta una breve ayuda de cómo utilizar nuestra aplicación. Primeramente debe ejecutarse el fichero .jar llamado “FUSE 1.0.jar” que es el encargado de recoger todos los datos para iniciar los cálculos convenientes para obtener los resultados. La Figura 2.2 muestra nuestra aplicación.

El software FUSE v1.0 (FUsion SEarching o Búsqueda por Fusión, en español) se compone de un conjunto de herramientas de cribado virtual para aplicaciones quimioinformáticas, con el fin de realizar búsqueda de similitud en conjuntos de datos químicos. Incluye además múltiples criterios de búsqueda, como es la fusión de datos, que está estrechamente relacionado con el nombre del software. La idea general de esta aplicación consiste en recuperar las moléculas del conjunto de datos que son más similares a una estructura de consulta definida internamente, utilizando una definición cuantitativa de la similitud estructural intermolecular. El fundamento de estos procedimientos es que las moléculas similares es muy probable que compartan características similares, por lo que juegan un rol importante en el diseño de fármacos.

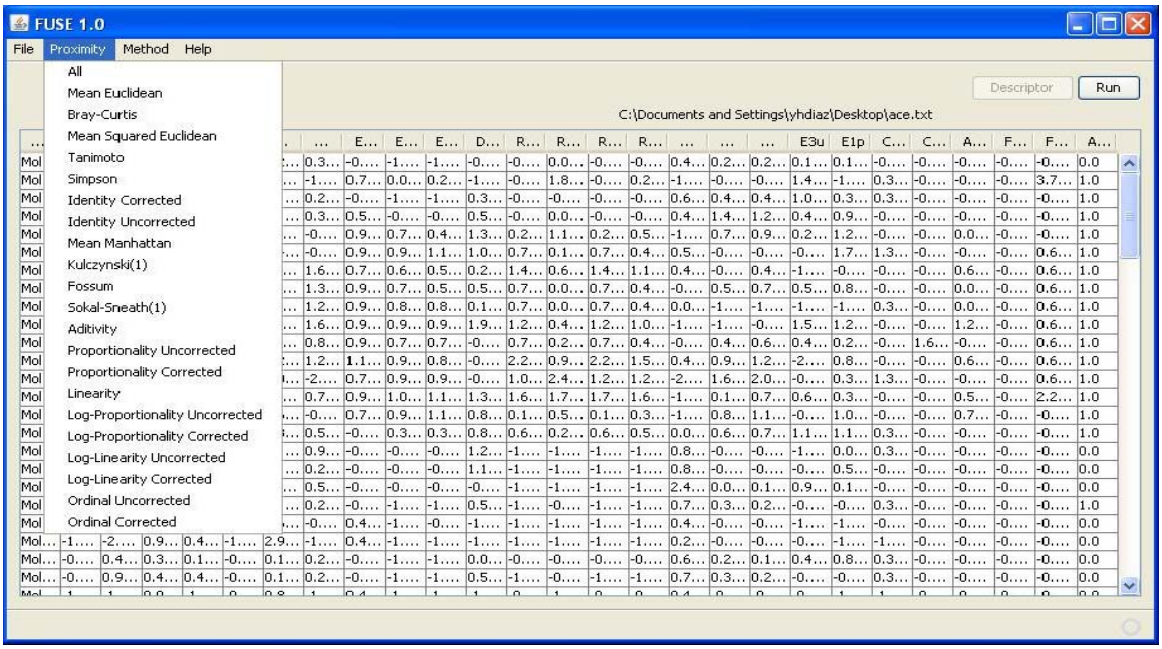


Figura 2-2 Interfaz Visual de la aplicación FUSE v1.0

El menú File, además de cargar el fichero de datos (.txt) que se va a analizar, también le permite al investigador guardar los resultados en un fichero de texto especificado. En el menú Proximity aparecen las medidas de proximidad implementadas en nuestro trabajo y da la opción al investigador de escoger la medida o las medidas que se quieren comparar. En el menú Method aparecen las distintas estrategias de búsqueda de similitud, como búsqueda simple, fusión de grupo, turbo similitud y fusión de similitud, pero en nuestro caso, solo se implementó la fusión de grupos. Por último, el botón Run es el que se encarga de comenzar todas las operaciones para llegar a los resultados finales. El fichero que se obtiene como respuesta tiene como nombre la medida que se utilizó junto con el método y el clasificador utilizado donde se imprimen, entre otros resultados, la media de las áreas AUC[CROC] y su desviación estándar.

2.7 Conclusiones Parciales del Capítulo

Se introdujeron nuevas medidas de similitud basadas en el concepto de acuerdo relacional junto con otras medidas de semejanza molecular reportadas en la literatura. Se seleccionaron nueve 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.

CAPÍTULO 3. RESULTA1DOS Y DISCUSIÓN

En este capítulo presentamos los resultados obtenidos en los experimentos de validación cruzada destinados para la evaluación de la efectividad de los modelos de proximidad. Seguidamente se procede al análisis de significación de los modelos en su conjunto y por pares, empleando técnicas estadísticas no paramétricas y paramétricas. Finalmente, se realiza una aplicación de las mejores medidas de semejanza en la "recuperación temprana" de uno de los conjuntos de datos estudiados, y se muestra una versión gráfica de los resultados para el análisis visual correspondiente.

3.1 Análisis de Dimensionalidad de los Datos

Conjuntos de Datos de la Química Medicinal

En la Tabla 3.1.1 se muestran los descriptores linealmente relevantes seleccionados para los ocho conjuntos de datos de Sutherland *et al.* (Sutherland *et al.,* 2004). Los porcentajes de reducción de dimensionalidad en la fase de eliminación de los rasgos binarios (selección del Weka) correspondiente a cada conjunto de datos fueron: 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 %), THR 56,36 % (98,93 %), la media geométrica de estos valores siendo 55,01 % (98,74 %), respectivamente. Estos resultados indican que aproximadamente 55 % de los descriptores moleculares del DRAGON son binarios y, por tanto, descartados en este estudio; además que las etapas subsecuentes de limpieza y selección en el Weka resultaron en un 98,74 % de eliminación de rasgos no relevantes.

Tabla 3.1.1 Lista de descriptores moleculares seleccionados por la técnica de selección de rasgos del Weka CfsSubsetEval

Datos	Descriptores	Familia[a]	Dim[b]	Datos	Descriptores	Familia[a]	Dim[b]
ACE	MAXDP	topological descriptors	2D	COX-2	Lop	topological descriptors	2D
	PW4	topological descriptors	2D		D/Dr09	topological descriptors	2D
	Lop	topological descriptors	2D		X1A	connectivity indices	2D
	BIC5	information indices	2D		G(N..O)	geometrical descriptors	3D

	ATS4m	2D autocorrelations	2D
	MATS8m	2D autocorrelations	2D
	MATS3p	2D autocorrelations	2D
	EEig03d	edge adjacency indices	2D
	EEig11d	edge adjacency indices	2D
	EEig12d	edge adjacency indices	2D
	DISPp	geometrical descriptors	3D
	RDF035u	RDF descriptors	3D
	RDF035m	RDF descriptors	3D
	RDF035e	RDF descriptors	3D
	RDF035p	RDF descriptors	3D
	Mor23m	3D-MoRSE descriptors	3D
	Mor26v	3D-MoRSE descriptors	3D
	Mor26p	3D-MoRSE descriptors	3D
	E3u	WHIM descriptors	3D
	E1p	WHIM descriptors	3D
	C-006	atom-centred fragments	1D
	C-026	atom-centred fragments	1D
	ALOGP2	molecular properties	Others
	F03[O-O]	2D frequency fingerprints	2D
	F06[O-O]	2D frequency fingerprints	2D
AchE	D/Dr07	topological descriptors	2D
	RDF060m	RDF descriptors	3D
	RDF060v	RDF descriptors	3D
	Mor12u	3D-MoRSE descriptors	3D
	Mor30m	3D-MoRSE descriptors	3D
	Mor08v	3D-MoRSE descriptors	3D
	Mor30v	3D-MoRSE descriptors	3D
	Mor12e	3D-MoRSE descriptors	3D
	E3u	WHIM descriptors	3D
	P1v	WHIM descriptors	3D
	E1e	WHIM descriptors	3D
	R6u+	GETAWAY descriptors	3D
	R3m+	GETAWAY descriptors	3D
	H-049	atom-centred fragments	1D
	O-058	atom-centred fragments	1D
	F03[N-O]	2D frequency fingerprints	2D
	F05[N-N]	2D frequency fingerprints	2D
	F07[N-F]	2D frequency fingerprints	2D
DHFR	nR05	constitutional descriptors	0D
	D/Dr10	topological descriptors	2D
	GATS7m	2D autocorrelations	2D
	GATS6p	2D autocorrelations	2D
	BELm2	Burden eigenvalues	2D

IC4	information indices	2D
SIC5	information indices	2D
BIC5	information indices	2D
MATS4m	2D autocorrelations	2D
MATS4p	2D autocorrelations	2D
GATS4m	2D autocorrelations	2D
GATS6e	2D autocorrelations	2D
GATS5p	2D autocorrelations	2D
JGI10	topological charge indices	2D
RDF045u	RDF descriptors	3D
RDF090u	RDF descriptors	3D
RDF155u	RDF descriptors	3D
RDF090m	RDF descriptors	3D
RDF090e	RDF descriptors	3D
RDF155e	RDF descriptors	3D
Mor22m	3D-MoRSE descriptors	3D
Mor11e	3D-MoRSE descriptors	3D
Mor32e	3D-MoRSE descriptors	3D
E3u	WHIM descriptors	3D
G2m	WHIM descriptors	3D
G3v	WHIM descriptors	3D
G1e	WHIM descriptors	3D

	BELe1	Burden eigenvalues	2D
	RCI	geometrical descriptors	3D
	Mor10u	3D-MoRSE descriptors	3D
	Mor03m	3D-MoRSE descriptors	3D
	Mor04m	3D-MoRSE descriptors	3D
	Mor09e	3D-MoRSE descriptors	3D
	R5u	GETAWAY descriptors	3D
	C-033	atom-centred fragments	1D
	O-057	atom-centred fragments	1D
	F04[C-N]	2D frequency fingerprints	2D
	F04[N-O]	2D frequency fingerprints	2D
GBP	X5A	connectivity indices	2D
	BIC1	information indices	2D
	MATS8v	2D autocorrelations	2D
	MATS7e	2D autocorrelations	2D
	Mor13m	3D-MoRSE descriptors	3D
	R5m+	GETAWAY descriptors	3D
	C-006	atom-centred fragments	1D
	H-046	atom-centred fragments	1D
	F02[N-O]	2D frequency fingerprints	2D
	F07[O-O]	2D frequency fingerprints	2D
THERM	X5v	connectivity indices	2D

	E3p	WHIM descriptors	3D
	R6e+	GETAWAY descriptors	3D
	nR=Cs	functional group counts	1D
	nArCONR2	functional group counts	1D
	H-053	atom-centred fragments	1D
	O-058	atom-centred fragments	1D
BZR	TI2	topological descriptors	2D
	Vindex	information indices	2D
	ATS7e	2D autocorrelations	2D
	J3D	geometrical descriptors	3D
	HOMA	geometrical descriptors	3D
	RDF020u	RDF descriptors	3D
	RDF030m	RDF descriptors	3D
	RDF055m	RDF descriptors	3D
	RDF020p	RDF descriptors	3D
	RDF030p	RDF descriptors	3D
	Mor09u	3D-MoRSE descriptors	3D
	Mor04v	3D-MoRSE descriptors	3D
	G3p	WHIM descriptors	3D
	H7m	GETAWAY descriptors	3D
	R6u	GETAWAY descriptors	3D
	R3m	GETAWAY descriptors	3D
	IC1	information indices	2D
	GATS5m	2D autocorrelations	2D
	GATS7p	2D autocorrelations	2D
	RDF065m	RDF descriptors	3D
	Mor17m	3D-MoRSE descriptors	3D
	Mor31m	3D-MoRSE descriptors	3D
	Mor16e	3D-MoRSE descriptors	3D
	Du	WHIM descriptors	3D
	R5p	GETAWAY descriptors	3D
	nCt	functional group counts	1D
	nROH	functional group counts	1D
	F01[O-S]	2D frequency fingerprints	2D
	F03[C-N]	2D frequency fingerprints	2D
	F09[C-N]	2D frequency fingerprints	2D
THR	TI2	topological descriptors	2D
	MATS5v	2D autocorrelations	2D
	EEig02x	edge adjacency indices	2D
	JGI7	topological charge indices	2D
	P2u	WHIM descriptors	3D
	E2p	WHIM descriptors	3D
	E3s	WHIM descriptors	3D
	H8m	GETAWAY descriptors	3D

C-005	atom-centred fragments	1D	HATS6v	GETAWAY descriptors	3D
H-047	atom-centred fragments	1D	nCq	functional group counts	1D
N-072	atom-centred fragments	1D	nHDon	functional group counts	1D
Hy	molecular properties	Others	H-053	atom-centred fragments	1D
F01[O-S]	2D frequency fingerprints	2D	Hy	molecular properties	Others
F07[N-F]	2D frequency fingerprints	2D	F08[C-S]	2D frequency fingerprints	2D
RDF055m	RDF descriptors	3D	F08[N-O]	2D frequency fingerprints	2D

[a]Clasificación acorde a la familia de descriptores (o bloques, según el software DRAGON), [b]Clasificación acorde a la dimensionalidad o complejidad de la representación molecular.

Este resultado es importante pues en la bibliografía especializada en problemas de clasificación, se ha reportado que a medida que la dimensionalidad aumenta, los datos se vuelven cada vez más dispersos en el espacio que ocupan lo cual trae grandes problemas para ambos, el aprendizaje supervisado y no supervisado, este fenómeno se ha referido antes como *la maldición de la dimensionalidad*(Janecek *et al.,* 2008).Además, un número grande de descriptores en la representación pueden contener rasgos irrelevantes o débilmente irrelevantes, y se conoce que afectan negativamente la exactitud de los algoritmos de predicción (John *et al.,* 1994); el caso extremo de este fenómeno se ilustra en *el teorema del patito feo* de Watanabe; básicamente, 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) (Watanabe, 1969). Por otra parte, desde el punto de vista químico, estos valores significativos sugieren un grado alto de especificidad en las relaciones rasgos moleculares-actividad farmacológica.

Con el objetivo de estudiar el comportamiento estadístico de los rasgos seleccionados, la distribución empírica de frecuencia (*def*) de la variable categórica "familia" de cada conjunto de datos fue comparada con la *def* para los conjuntos de datos tomados como un todo, *i.e*. el *conjunto de fusión*, y también a la *def* a priori del DRAGON. Para este fin, se realizó una prueba de bondad de ajuste χ^2 (ver Tabla 3.1.2).

Tabla 3.1.2 Significación de la prueba de bondad de ajuste χ^2 entre las distribuciones empíricas de frecuencia de los descriptores moleculares

Objetivo[a]	Fusión	ACE	AchE	BZR	COX-2	DHFR	GBP	THERM	THR
Fusión	1	0,4300	0,5770	0,7245	0,8601	0,0081**	0,7834	0,8080	0,2837
DRAGON	~ 0***	0,0011**	~ 0***	0,0022**	0,0027**	0,0031**	0,0152*	0,1840	0.0045**

[a] Distribuciones objetivos para comparación, *Pruebas estadísticas significativas ($p < 0.05$), **Pruebas estadísticas altamente significativas ($p < 0.01$),***Pruebas estadísticas extremadamente significativas ($p < 0,001$)

A partir de estos puntajes de probabilidad se puede inferir que existe una similitud significativa entre cada *def* de cada repositorio y la *def* del repositorio de fusión en cuanto a la variable "familia", excepto para el conjunto DHFR. En otras palabras, con respecto al análisis de las familias de descriptores es equivalente tratar estos conjuntos ya sea por separado o como un todo, o *contexto*. Esta tendencia se puede visualizar en los **Gráficos 3.1.1** y **3.1.2** para las variables "familia" y "dimensionalidad", respectivamente, considerando el conjunto de fusión como referencia. Por otra parte, resulta evidente que existe una disimilitud significativa entre cada *def* de cada conjunto y la *def* de los descriptores en el DRAGON en cuanto a la variable "familia", excepto para el conjunto THERM, lo cual sugiere que los rasgos seleccionados representan una particularidad específica de la modelización quimioinformática y no están sesgados por la distribución *a priori* que brinda este software. Además, es interesante notar que la información química codificada por descriptores 2-D y 3-D está estrechamente relacionada con las actividades farmacológicas pues estos contribuyen al 84 % de los rasgos seleccionados.

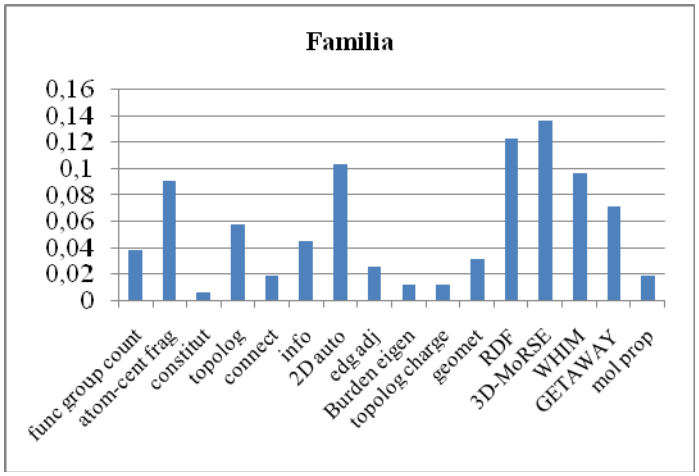


Gráfico 3.1.1. Contribución relativa de los descriptoresal conjunto de fusión acorde al criterio "familia"

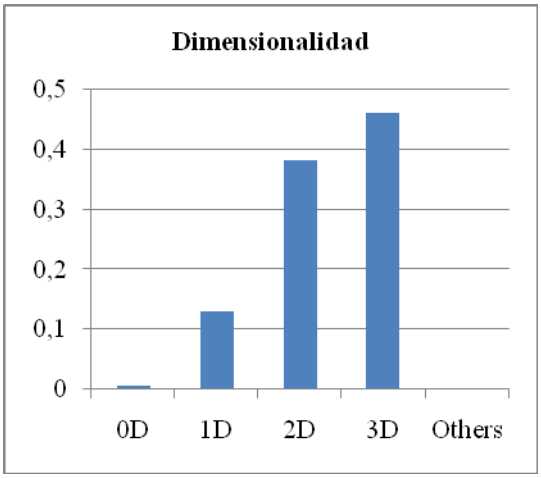


Gráfico 3.1.2. Contribución relativa de los descriptoresal conjunto de fusión acorde al criterio "dimensionalidad"

Un análisis más refinado de comparación de proporciones, confirmó que las contribuciones de estos dos grupos son significativamente mayores (p ~ 0 en ambos casos) que la probabilidad de hipótesis nula (nul-p = 1/6). Un análisis similar pero considerando los bloques de descriptores mostró (ver Tabla 3.1.3) que las contribuciones de las familias 2-D autocorrelacionadas, descriptores RDF (3-D), descriptores 3-D-Morse y descriptores WHIM (3-D) son significativamente mayores que la probabilidad de hipótesis nula (nul-p = 1/6)lo cual es consistente con el resultado previo. En términos de la Química-Teórica ambos resultados sugieren que la configuración espacial de los átomos, así como la topología molecular juegan un rol dominante en las interacciones ligando-receptor, lo cual también está de acuerdo con regularidades similares reportadas por otros autores.

Tabla 3.1.3. Prueba binomial para homogeneidad en las familias de descriptores

Familia	Sig[a]	Familia	Sig[a]	Familia	Sig[a]
functional group counts	0,8499	2D autocorrelations	**0,0112***	3D-MoRSE descriptors	**~ 0*** **
atom-centred fragments	0,0528	edge adjacency indices	0,9554	WHIM descriptors	**0,0256***
constitutional descriptors	0,9965	Burden eigenvalues	0,9909	GETAWAY descriptors	0,2666
topological descriptors	0,5161	topological charge indices	0,9909	molecular properties	0,9789
connectivity indices	0,9789	geometrical descriptors	0,9143		
information indices	0,7593	RDF descriptors	**0,0005*** **		

[a]Significación estadística, hipótesis alternativa de trabajo: p > 0.0625 (1/16). Se usó la aproximación normal con la corrección clásica X + 0.5 para la variable aleatoria.*Pruebas estadísticas significativas (p < 0.05), **Pruebas estadísticas altamente significativas (p < 0.01), ***Pruebas estadísticas extremadamente significativas (p < 0.001).

Conjuntos de Datos NAAS

Se acuerdo con los resultados obtenidos anteriormente se decidió no calcular el total de 3 224 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 NAAS, que también pertenece a la química medicinal o terapéutica. Este tipo de proceder es consistente con la teoría del *meta aprendizaje* (Biggs, 1985). Exceptuando este punto, las demás etapas de prefiltrado, reescalado, y selección de rasgos se llevaron a cabo de manera igual a como se hizo con los ocho conjuntos de Sutherland *et al*. (ver subsección **2.3 “Representación Molecular y Obtención de las Matrices de Datos”, en el**

Capítulo 2 "Materiales y Métodos"). De este modo, 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 3.1.4. 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 3.1.4 Descriptores seleccionados con el evaluador *CfsSubsetEval* del Weka para el conjunto NAAS

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

3.2 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 cuantificada a través de la exactitud de la clasificación, expresada a través de los valores medios y desviaciones estándares de las AUC[CROC] obtenidas del experimento de validación cruzada de diez pliegues. Los resultados obtenidos se muestran en las Tablas 3.1.5-3.1.7.

Tabla 3.1.5 Exactitud de los modelos de proximidaden 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 3.1.6 Exactitud de los modelos de proximidaden 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 3.1.7 Exctitud de los modelos de proximidaden 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 y se corresponden con las brindadas en el Cap 2, Sec. 2.1.1; [**]Cada casilla presenta el resumen de los resultados obtenidos en el experimento de validación cruzada de diez pliegues para cada modelo de proximidad en cada conjunto de datos. El formato presentado en la tabla es de la forma: (“media”, “desviación estándar”) donde, la primera componente representa la media de las AUC[CROC] obtenidas en los diez pliegues y la segunda componente la desviación estándar de dichos valores.

Con el objetivo de detectar diferencias globales entre los modelos de proximidad en la efectividad de la recuperación temprana de activos, se aplicó un contraste de Friedman, el cual arrojó diferencias extremadamente significativas entre los mismos ($p \sim 0$). A partir de estos resultados se obtuvo un ordenamiento jerárquico tentativo (existen ataduras) de la potencia de dichos modelos de acorde al rango promedio asignado por la prueba (ver Tabla 3.1.8).

Tabla 3.1.8 Potencia relativa de los modelos de similitud en la recuperación temprana de activos según Friedman

Orden[a]	Modelo	Rango Medio	Orden	Modelo	Rango Medio
1**[t]	e^R	18,39	12	*BC*	10,67
2**[t]	e^{Rc}	18,39	13	*F*	9,78
3*	*Cos*	15,22	14**	*Lp*	8,11
4**	p^c	14,33	15	EM^t	6,83
5*	*r*	14,11	16	ECM^t	6,83
6**	*a*	14	17**	Lp^c	6,78
7**	e^c	13,67	18	*Sim*	6,11
8[t]	*Tan*	13,61	19	*MM*	6
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]Mientras más alto es el rango medio, más potente es el modelo; *modelos de acuerdo relacional usados genéricamente como "medidas de similitud" por Al Khalifa *et al.* (Al Khalifa *et al.,* 2009), **modelos de acuerdo relacional aportados por el presente trabajo, [t] modelos atados por el "Rango Medio".

Los resultados mostrados en la Tabla 3.1.8 muestran que los modelos de proximidad basados en el acuerdo relacional se comportan relativamente superiores a otras medidas de similitud de propósito no específico, evidencia que resulta más concisa cuantitativamente, si se tiene en cuenta que el rango promedio para los modelos de acuerdo relacional es aproximadamente 12,00, mientras que el rango promedio para los modelos que no son de acuerdo relacional es 9,67. Si solamente se tiene en cuenta el criterio de novedad de las medidas propuestas en este trabajo, entonces los modelos propuestos aportan un rango

promedio de 11,22 mientras que los modelos ya reportados aportan un rango promedio de 10.83. Además, resulta alentador que el 55,56 % de los modelos propuestos en el trabajo están incluidos entre los 10 modelos más potentes ("top-10") inspeccionados.

Del conocimiento estadístico se sabe que la *no aceptación* de la hipótesis nula en la prueba de Friedman no implica que cada variable representa una población por sí misma, sino que sugiere la presencia de clústeres con una alta homogeneidad intra grupos y una baja homogeneidad inter grupos. Basado en esta idea se decidió aplicar una prueba de Wilcoxon por pares con el objetivo de detectar y agrupar modelos con similitudes poblacionales estadísticamente significativas. En esta ocasión se construyó una matriz de disimilitud *21 x 21* (21 es la cantidad de modelos estudiados) con entradas binarias, de modo que a pares de modelos con diferencias estadísticamente significativas según Wilcoxon les fue asignada una diferencia de "1", mientras que pares de modelos que no presentan diferencias significativas según este estadístico les fue asignado el valor "0". Esta matriz fue sometida a un análisis de conglomerados tipo Ward implementado en el paquete de análisis de datos ecológicos SYN-TAX (Podani, 2001). El dendograma obtenido se muestra en la Figura 3.1.

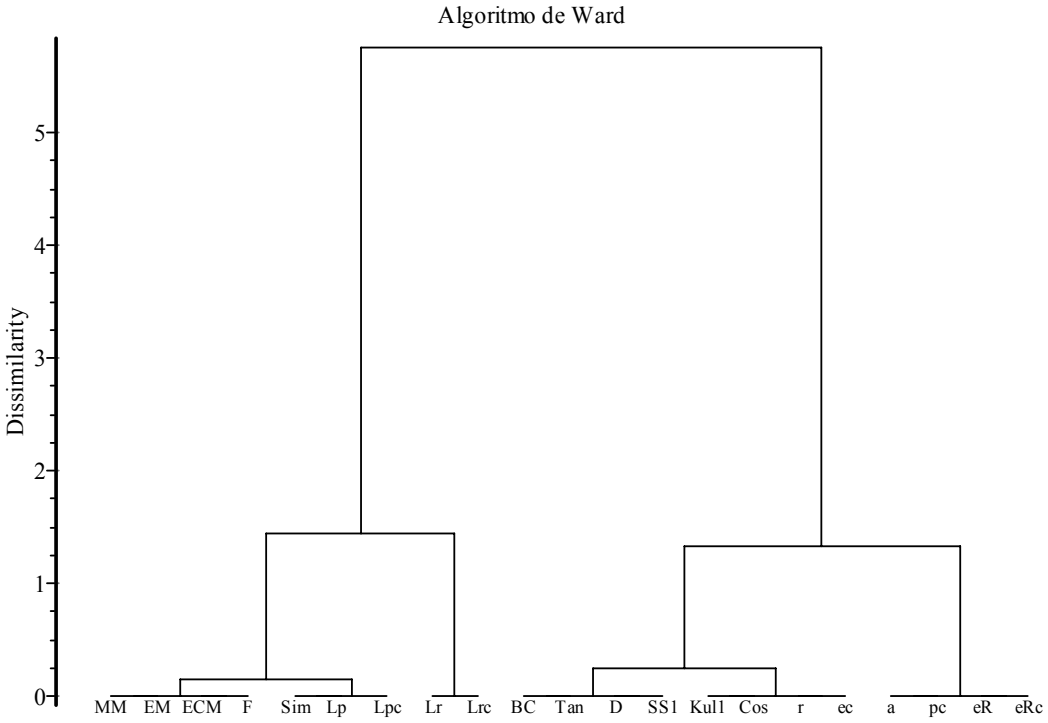


Figura 3.1 Dendograma obtenido del algoritmo de Ward para el agrupamiento de modelos homogéneos según la prueba de Wilcoxon.

El árbol de relaciones jerárquicas mostrado en la Figura 3.1, sugiere a simple vista una estructuración de los modelos en seis grupos homogéneos (distancia 0), ellos son: $G_1 = \{MM, EM, ECM, F\}$, $G_2 = \{Sim, Lp, Lpc\}$, $G_3 = \{Lr, Lrc\}$, $G_4 = \{BC, Tan, D, SS1\}$, $G_5 = \{Kul1, Cos, r, ec\}$, G en una forma familiar alo analizado con los valores del rango medio de la prueba de Friedman ya que estos clústeres se corresponden en buena lid con una partición secuencial practicada a la lista ordenada de la Tabla 3.1.8. En este sentido, en el dendograma se observa claramente la formación de dos grandes grupos, según la prueba de Mojena (Mojena, 1977)del cambio máximo entre niveles para determinar el número óptimo de clústeres; donde, uno de los grupos consiste en la fusión (unión) de los grupos G_1, G_2 y G_3, y, el otro grupo grande consiste en la fusión (unión) de los grupos G_4, G_5 y G_6; adicionalmente, estos resultados también son consistentes, en gran extensión con los anteriores, por cuanto sugiere la separación de los modelos en un grupo de 12 mejores modelos y otro de 9 peores modelos.

Teniendo en cuenta los resultados obtenidos anteriormente, un próximo paso en una investigación quimioinformática consistiría en desambiguar los mejores modelos en un

conjunto representativo de los problemas reales en dicha área, para ello se aplicó la prueba de comparación por pares *t*-student con la corrección de Welch (ver **Capítulo 2: Materiales y Métodos** para más detalles) empleando los datos de valores medios y desviaciones estándar las **Tablas 3.1.5-3.1.8** obtenidos para el repositorio NAAS. La técnica empleada fue similar que para la prueba de Wilcoxon, *i.e.* en una matriz binaria se registraron las diferencias significativas con "1" y no significativas con "0", la matriz de disimilitud fue sometida a un análisis de agrupamiento jerárquico de Ward.

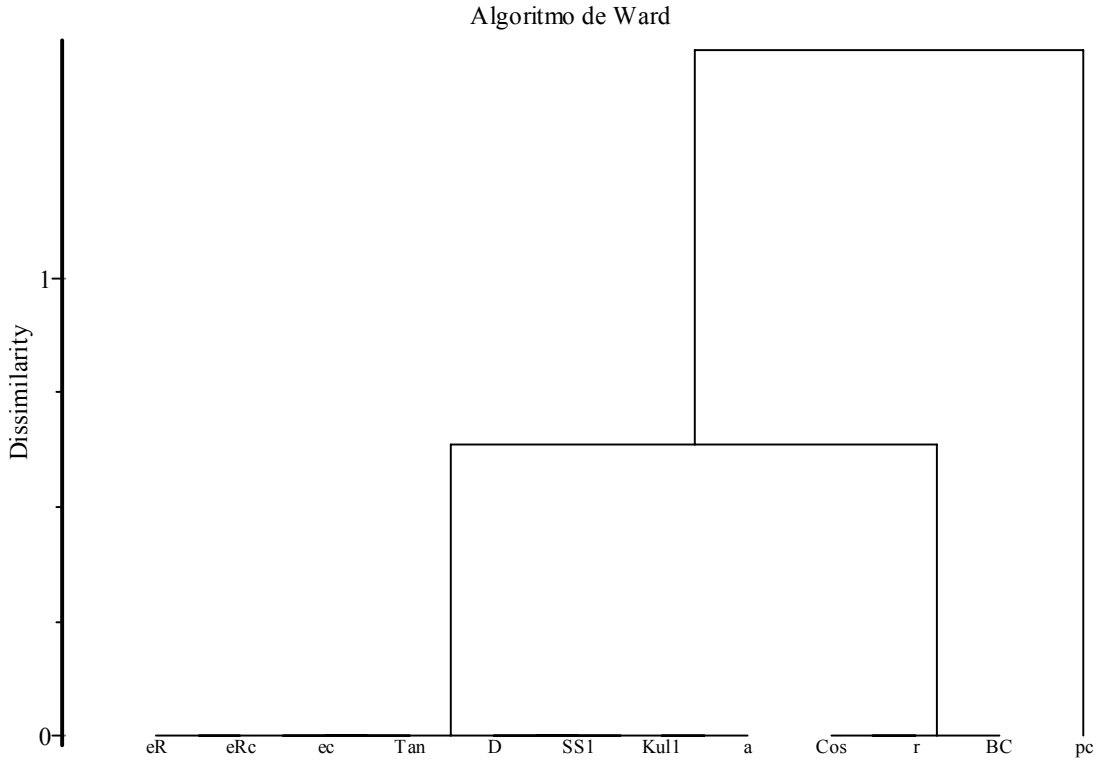


Figura 3.2 Dendograma obtenido del algoritmo de Ward para el agrupamiento de modelos homogéneos según la prueba *t*-student con corrección de Welch

El dendograma 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. Estos resultados son consistentes con los anteriores pues muestran que en este grupo de modelos equivalentes en potencia, cualquiera de ellos puede comportarse relativamente mejor que el resto para una problemática dada; no obstante, los modelos e^R y e^{Rc} se

mantuvieron como los mejores del grupo. Por último, para propósitos de comprobación visual de los estos resultados, en el Gráfico 3.2 se muestran las curvas de calidad CROC para el grupo de modelos de proximidad anteriores.

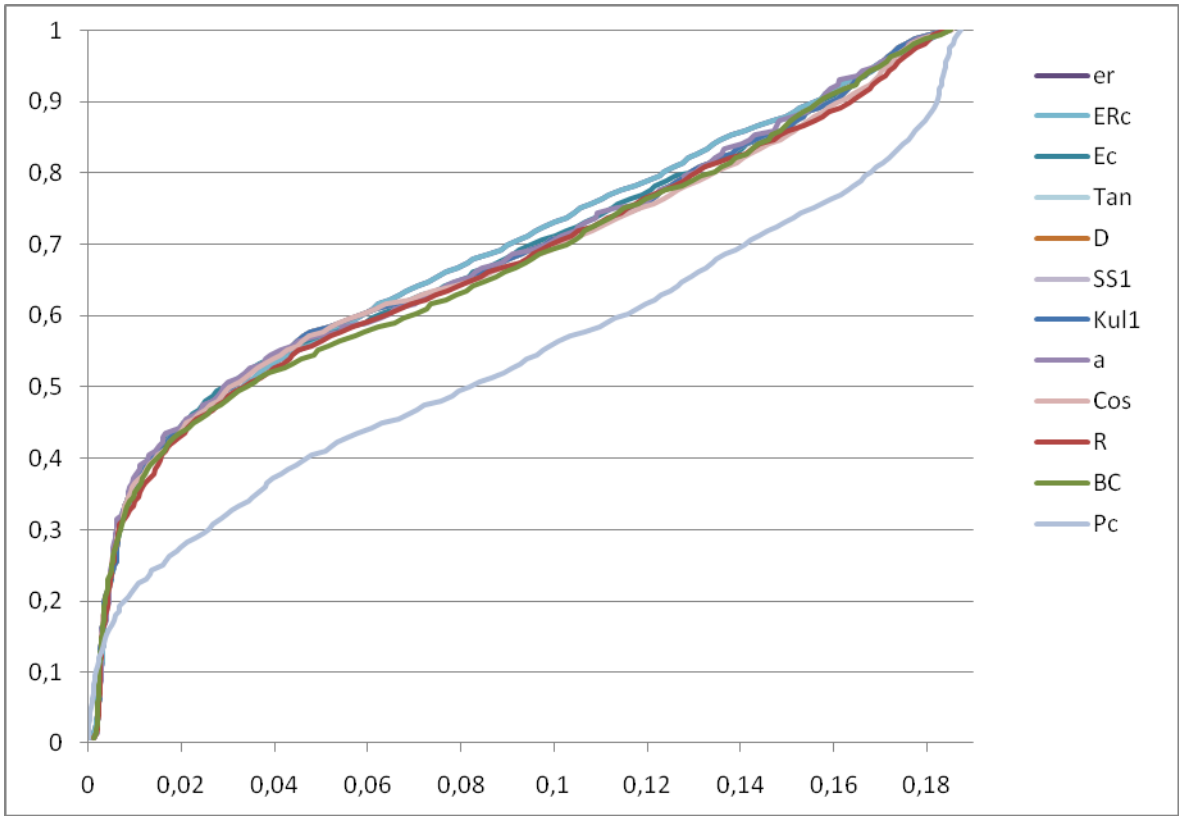


Gráfico 3.2 Secciones de curvas CROC para los mejores modelos de proximidad mostrando "el reconocimiento temprano" de los activos en el conjunto de datos NAAS.

3.3 Conclusiones Parciales

1. Las familias de descriptores 2-D y 3-D atraparon mayoritariamente la información química directamente relacionada al contexto farmacológico definido por los ocho conjuntos de datos de Sutherland. Específicamente, las familias 2D autocorrelacionadas (2-D), descriptores RDF (3-D), descriptores 3D-Morse (3-D) y descriptores WHIM (3-D) atraparon la información química implicada en estos fenómenos de interacción ligando-receptor.
2. Los resultados de la prueba de Friedman mostraron que de forma general, los modelos de proximidad basados en el acuerdo relacional se comportan relativamente superiores a otras medidas de similitud que no están basadas en esta teoría.

3. Más de la mitad de los modelos propuestos como novedosos en el trabajo están incluidos entre los 10 modelos más potentes analizados, estos son: e^R, e^{Rc}, p^c, a y e^c
4. La aplicación de los doce mejores modelos a la recuperación temprana del conjunto NAAS derivó en valores de exactitud promedio AUC[CROC] en el rango 0.913-0.942, evidenciando la utilidad potencial de estas medidas en problemas quimioinformáticos reales.

CONCLUSIONES

1. Se implementaron herramientas para el cribado virtual de conjuntos quimio-bio-informáticos que consisten en 21 medidas de similitud para datos numéricos acopladas a un algoritmo de búsqueda rápida, acopladas, a su vez, a varias técnicas de fusión de datos que comprende fusión de grupos (MAX-SIM, SUM-IR-MAX), fusión de similitud y turbo similitud. 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. Los resultados del análisis del comportamiento de los descriptores moleculares en los conjuntos de datos de Suhterland mostró que unas pocas familias de descriptores 2-D y 3-D pueden atrapar la información relacionada con las interacciones ligando-receptor, estas fueron: autocorrelacionadas (2-D), descriptores RDF (3-D), descriptores 3D-Morse (3-D) y descriptores WHIM (3-D).

3. Los modelos de proximidad basados en el acuerdo relacional se comportan relativamente superiores a otros modelos no definidos a partir de esta teoría en el reconocimiento temprano de compuestos líderes. Más de la mitad de los modelos propuestos como novedosos en el trabajo están incluidos entre los 10 modelos más potentes, los cuales resultaron ser: e^R, e^{Rc}, p^c, a y e^c.

RECOMENDACIONES

1. Introducir nuevas medidas de proximidad para el cribado virtual basado en búsqueda de similitud.

2. 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

1. Adamson, G. W. & Bush, J. A. (1973): «A method for the automatic classification of chemical structures», *Inf. Stor. Retriev,* (9)**:** 561-568.
2. Adamson, G. W. & Bush, J. A. (1975): «A comparison of the performance of some similarity and dissimilarity measures in the automatic classification of chemical structures», *J. Chem. Inf. Comput. Sci.,* (15)**:** 55-58.
3. Agrafiotis, D. K., Bandyopadhyay, D., WEGNER, J. K. & VAN VLIJMEN, H. (2007): «Recent Advances in Chemoinformatics», *J. Chem. Inf. Model. (*47)**:** 1279-1293.
4. Agrafiotis, D. K. & Cedeno, W. (2002): «Feature selection for structureactivity correlation using binary particle swarms». *J. Med. Chem.,* (45): 1098-1107.
5. Aha, D. W. (1992). «Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms». *International Journal of Man-Machine Studies,* (36): 267-287.
6. Al Khalifa, A., Haranczyk, M. & Holliday, J. (2009). «Comparison of Nonbinary Similarity Coefficients for Similarity Searching, Clustering and Compound Selection». *J. Chem. Inf. Model.,* (49)**:** 1193-1201.
7. Bajorath, J. (2002): «Integration of virtual and high-throughput screening». *Nat. Rev. Drug.Discov.,* (1**):** 882-894.
8. Baldi, P., Hirschberg, D. S. & Nasr, R. J. (2008): «Speeding Up Chemical Database Searches Using a Proximity Filter Based on the Logical Exclusive OR», *J. Chem. Inf. Model.,* (48)**:** 1367-1378.
9. Bayada, D. M., Hamersma, H. & Van Geerestein, V. J. (1999): «Molecular diversity and representativity in chemical databases», *J. Chem. Inf. Comput. Sci.,* (39**)** 1-10.
10. Bender, A. (2010): «Compound bioactivities go public.» *Nature Chemical Biology,* (6): 309.
11. Bender, A., Jenkins, J. L., Scheiber, J., Sukuru, S. C. K., Glick, M. & Davies, J. W. (2009): «How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space», *J. Chem. Inf. Model. ,* (49)**:** 108-119.
12. Bender, A., Mussa, H. Y. & Glen, R. C. (2004a): «Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier». *J. Chem. Inf. Comput. Sci. ,* (44**):** 170-178.

13. Bender, A., Mussa, H. Y., Glen, R. C. & Reiling, S. (2004b): «Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D):â€‰ ¿? Evaluation of Performance», *J. Chem. Inf. Comput. Sci. ,* (44)**:** 1708-1718.
14. Bentley, J. L., Weide, B. W. &Yao, A. C. (1980): Optimal Expected Time Algorithms for Closest Point Problems». *ACM Tram. Math. Sofr.,* (6)**:** 563-580.
15. Biggs, J. B. (1985). «The role of meta-learning in study process», *Brit. J. Educ. Psychol.,* (55)**:** 185-212.
16. Bleicher, K., Bohm, H., Muller, K. & Alanine, A. (2003): «Hit and lead generation: beyond high-throughput screening», *Nat. Rev. Drug. Discov.,* (2)**:** 369-378.
17. Böcker, A., Schneider, G. & Teckentrup, A. (2004): «Status of HTS Data Mining Approaches», *QSAR Comb. Sci.,* (23): 207-213.
18. Brown, R. D. (1997): «Descriptors for diversity analysis», *Perspect. Drug Disc. Design,* (7)**:** 31-49.
19. Brown, R. D. & Martin, Y. C. (1996): «Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection», *J. Chem. Inf. Comput. Sci. ,* (36)**:** 572-584.
20. Bruce, C. L., Melville, J. L., Pickett, S. D. & Hirst, J. D. (2007): «Contemporary QSAR Classifiers Compared», *J. Chem. Inf. Model. ,* (47)**:** 219-227.
21. Bunke, H. & Shearer, K. (1998): «A graph distance metric based on the maximal common subgraph», *Pattern Recog. Lett.,* (19)**:** 255-259.
22. Burden, F. D., Ford, M. G., Whitley, D. C. & Winkler, D. A. (2000): «Use of automatic relevance determination in QSAR studies using Bayesian neural networks», *J. Chem. Inf. Comput. Sci.,* (40): 1423-1430.
23. Carhart, R. E., Smith, D. H. & Venkataraghavan, R. (1985): «Atom pairs as molecular features in structure-activity studies: definition and applications», *J. Chem. Inf. Comput. Sci. ,* (25)**:** 64-73.
24. Chanda, S. & Caldwell, J. (2003): «Fulfilling the promise: drug discovery in the postgenomic era», *Drug Discov. Today,* (8)**:** 168-174.
25. Chemaxon 2010. JChem for Excel. 5.3.8 (166)
26. Chen, B., Mueller, C. & Willett, P. (2010): «Combination Rules for Group Fusion in Similarity-Based Virtual Screening». *Mol. Inf.,* (29)**:** 533-541.

27. Chen, X. & Brown, F. K. (2007): «Asymmetry of Chemical Similarity», *ChemMedChem,* (2**):** 180 - 182.
28. Chen, X. & Reynolds, C. H. (2002): «Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients», *J. Chem. Inf. Comput. Sci. ,* (42**):** 1407-1414.
29. Clark, R. & Webster-Clark, D. (2008): «Managing Bias in ROC Curves». *J. Comput.-Aided Mol. Des. ,* (22)**:** 141-146.
30. Conover, W. J. & Iman, R. L. (1981): «Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics». *The American Statistician,* (35)**:** 124-129.
31. Cruz Monteagudo, M. (2009): *Métodos de Correlación Estructura-Actividad Multiobjetivos Aplicados al Desarrollo Racional de Fármacos.* Dr., Universidad Central "Marta Abreu" de Las Villas.
32. Cuissart, B., Touffet, F., Cremilleux, B., Bureau, R. & Rault, S. (2002): «The maximum common substructure as a molecular depiction in a supervised classification context: experiments in quantitative structure/biodegradability relationships», *J. Chem. Inf. Comput. Sci.,* (42**):** 1043-1052.
33. Culp, M., Johnson, K. & Michailidis, G. (2010): «The EnsembleBridge Algorithm: A New Modeling Tool for Drug Discovery Problems», *J. Chem. Inf. Model,* (50**):** 309-316.
34. Debouck, C. & Goodfellow, P. (1999): DNA microarrays in drug discovery and development. *Nat. Genet.,* 21**,** 48-50.
35. Dimasi, J., Hansen, R. & Grabowski, H. (2003): The price of innovation: new estimates of drug development costs. *J. Health Econ.,* (22**):** 151-185.
36. Drews, J. (2000): «Drug discovery: a historical perspective» *Science,* (287)**:** 1960-1964.
37. Dudek, A. Z., Arodz, T. & Gálvez, J. (2006): «Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review», *Comb. Chem. High Throughput Screen.,* (9**):** 1-16.

38. Edgar, S. J., Holliday, J. D. & Willett, P. (2000): «Effectiveness of retrieval in similarity searches of chemical databases: A review of performance measures», *Journal of Molecular Graphics and Modelling* (18**):** 343-357.
39. Efron, B. & Tibshirani, R. J. (1993): *An introduction to the Bootstrap,* New York, USA, Chapman & Hall.
40. Ekins, S., Boulanger, B., Swaan, P. & Hupcey, M. (2002): Towards a new age of virtual ADME/TOX and multidimensional drug discovery. *J. Comput. Aided Mol. Des.,* (16**):** 381-401.
41. Ellis, D., Furner-Hines, J. & Willett, P. (1994a): Measuring the Degree of Similarity between Objects in Text-Retrieval Systems. *Perspect. Inf. Manage.,* (3**):** 128-149.
42. Ellis, D., Furner-Hines, J. & Willett, P. (1994b): «Measuring the degree of similarity between objects in text retrieval systems», *Perspect. Inf. Manag.,* (3**):** 128-149.
43. Fayyad, U. M. & Irani, K. B.: «Year. Multi-interval discretisation of continuousvalued attributes for classification learning», *In:* KAUFMANN, M., ed. Proceedings of the Thirteenth International Join Conference on Artificial Intelligence, 1993.
44. Fligner, M., Verducci, J. & Blower, P. (2002): «A Modification of the Jaccard/Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings», *Technometrics,* (44**):** 110-119.
45. Flower, D. (1998): «On the Properties of Bit String-Based Measures ofChemical Similarity», *J. Chem. Inf. Comp. Sci.,* (38**):** 379-386.
46. Friedman, J. H., Bentlev, J. L. & Finkel, R. A. (1977): «An Algorithm for Finding Best Matches in-Logarithmic Expected Time», *k M Trans. Marh. Sofr.,* (3): 209-226.
47. Friedman, M. (1937): «The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance». *Journal of the American Statistical Association,* (32**):** 675-701.
48. Geppert, H. & Bajorath, J. (2010): «Advances in 2D fingerprint similarity searching», *Expert Opin. Drug Discov.,* (5**):** 529-542.
49. Glen, R. & Adams, S. (2006a.): «Similarity Metrics and Descriptor Spaces – Which Combinations to Choose?» *QSAR & Combinatorial Science,* (25)**:** 1133-1142.
50. Glen, R. C. & Adams, S. E. (2006b). «Similarity Metrics and Descriptor Spaces – Which Combinations to Choose?» *QSAR Comb. Sci. ,* (25)**:** 1133 - 1142.

51. Guyon, I. & Elisseeff, A. (2003°): «An Introduction to Variable and Feature Selection». *Journal of Machine Learning Research,* (3): 1157-1182.
52. Guyon, I. & Elisseeff, A. (2003b). «An Introduction to Variable and Feature Selection». *Journal of Machine Learning Research,* (3): 1157-1182.
53. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009): «The WEKA Data Mining Software: An Update», *SIGKDD Explorations,* 3-6-4 (11).
54. Hall, M. A. (1998): *Correlation-based Feature Subset Selection for Machine Learning.* PhD, The University of Waikato.
55. Hann, M. & Green, R. (1999): «Chemoinformatics--a new name for an old problem?», *CurrOpin Chem Biol,* (3**):** 379-383.
56. Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E. & Schuffenhauer, A. (2004a): «Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures». *J. Chem. Inf. Comput. Sci. ,* (44**):** 1177-1185.
57. Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E. & Schuffenhauer, A. (2005): «Enhancing the Effectiveness of Similarity-Based Virtual Screening Using Nearest-Neighbour Information», *J. Med. Chem.,* (48**),** 7049-7054.
58. Hert, J. R. M., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E. & Schuffenhauer, A. (2004b): «Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures», *J. Chem. Inf. Comput. Sci.,* (44**):** 1177-1185.
59. Holliday, J. D., Hu, C.-Y. & Willett, P. (2002): «Grouping of Coefficients for the Calculation of Inter-Molecular Similarity and Dissimilarity using 2D Fragment Bit-Strings» *Combinatorial Chemistry & High Throughput Screening,* (5**):** 155-166.
60. Holliday, J. D., Ranade, S. S. & Willett, P. (1995): «A Fast Algorithm for Selecting Sets of Dissimilar Molecules from Large Chemical Databases». *Quant. Struct.-Act. Relat.,* (14**):** 501-506.
61. Horvath, D. & Jeandenans, C. (2003): «Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces-a novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles», *J Chem Inf Comput Sci,* (43)**:** 680-690.

62. Janecek, A., Gansterer, W., Demel, M. & Ecker, G.: «Year. On the Relationship Between Feature Selection and Classification Accuracy». *In:* SAEYS, Y., LIU, H., INZA, I., WEHENKEL, L. & VAN DE PEER, Y., eds. FSDM 2008 : Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery, September 15 2008 Antwerp, Belgium. JMLR: Workshop and Conference Proceedings, 90-105.
63. Johansson, U., Löfström, T. & Norinder, U.: «Year. Evaluating Ensembles on QSAR Classification». *In:* JOHANSSON, R., VAN LAERE, J. & MELLIN, J., eds. 3rd Skövde Workshop on Information Fusion Topics, 2009. Univeristy of Skövde, 49-54.
64. Johansson, U., Löfström, T. & Norinder, U. «Year. Evaluating Ensembles on QSAR Classification». *In:* JOHANSSON, R., VAN LAERE, J. & MELLIN, J., eds. 3rd Skövde Workshop on Information Fusion Topics 2009 (SWIFT 2009), Oct 12-13 2009 Skövde, Sweden. Univeristy of Skövde, 49-54.
65. John, G. H., Kohavi, R. & Pfleger, K.: «Year. Irrelevant features and the subset selection problem». *In:* COHEN, W. W. & HIRSH, H., eds. ICML 1994: Proceedings of the Eleventh International Conference on Machine Learning July 10-13 1994 Rutgers University, New Brunswick, NJ, USA. Morgan Kaufman, 121-129.
66. Johnson, M. A. (1989): «A review and examination of mathematical spaces underlying molecular similarity analysis». *J. Math. Chem. ,* (3**):** 117-145.
67. Jónsdóttir, S. Ó., JØRGENSEN, F. S. & BRUNAK, S. (2005): «Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates». *Bioinformatics,* (21): 2 145-2 160.
68. Jorgensen, W. (2004). The many roles of computation in drug discovery. *Science,* (303)**:** 1813-1818.
69. Klebe, G. (2 000): «Recent developments in structure-based drug design». *J Mol Med,* (78)**:** 269-281.
70. Koivalishyn, V., Tetko, V. I., Luik, A. I., Kholodovych, V. V., Villa, A. E. P. & Livingstone, D. J. (1998): «Neural networks studies. Variable selection in the cascade-correlation learning architecture». *J. Chem. Inf. Comput. Sci.,* (38): 651-659.
71. Kruskal, W. H. & Wallis, W. A. (1952): «Use of Ranks in One-Criterion Variance Analysis», *Journal of the American Statistical Association,* (47**):** 583-621.

72. Kubinyi, H. (1995): «Strategies and recent technologies in drug discovery». *Pharmazie,* (50)**:** 647-662.
73. Kubinyi, H. (1995): «Strategies and recent technologies in drug discovery», *Pharmazie,* (50**):** 647-662.
74. Lazo, J. & WIPF, P. (2000): «Combinatorial chemistry and contemporary pharmacology». *J Pharmacol Exp Ther,* (293**):** 705-709.
75. Maggiora, G. M. & SHANMUGASUNDARAM, V. (2004): «Molecular Similarity Measures». *In:* BAJORATH, J. (ed.) *Chemoinformatics.* Humana Press.
76. Maldonado, A. G., DOUCET, J. P., PETITJEAN, M. & FAN, B.-T. (2006): «Molecular similarity and diversity in chemoinformatics: From theory to applications». *Molecular Diversity,* (10**):** 39-79.
77. Manly, C., Louise-May, S. & Hammer, J. (2001): «The impact of informatics and computational chemistry on synthesis and screening», *Drug Discov Today,* (6**):** 1101-1110.
78. Martin, Y. C. (2001): «Molecular Diversity: How we measure it? Has it lived up to its promise?» *Il Farmaco,* (56**):** 137-139.
79. Martin, Y. C., Bures, M. G. & Brown, R. D. (1998): «Validated descriptors for diversity measurements and optimization». *Pharm. Pharmacol. Commun., (*4): 147-152.
80. Massart, D. L. & Kaufman, D. L. (1983): *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis,* New York, Wiley.
81. Matter, H. & Potter, T. (1999): «Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets». *J. Chem. Inf. Comput. Sci.,* (39)**:** 1211-1225.
82. Mojena, R. (1977): «Hierarchical grouping methods and stopping rules: an evaluation», *The Computer Journal,* (20**):** 359-363.
83. Murtagh, F. (1982): «A Very Fast, Exact Nearest Neighbour Algorithm for Use in Information Retrieval», *Znf. Technol.: Res. Deu.,* I**,** 275-283.
84. Nath, R., Rajagopalan, B. & Ryker, R. (1997): «Determining the saliency of input variables in neural networks classifiers», *Comput. Ops. Res.,* (24**):** 767-773.

85. Nicholls, A. (2008): «What Do We Know and When Do We Know It?» *J. Comput.-Aided Mol. Des.,* (22**):** 239-255.
86. Nikolova, N. & Jaworska, J. (2003a): «Approaches to Measure Chemical Similarity - a Review», *QSAR Comb. Sci.,* (22): 1006-1026.
87. Nikolova, N. & Jaworska, J. (2003b): «Approaches to Measure Chemical Similarity – a Review», *QSAR & Combinatorial Science,* (22**):** 1006-1026.
88. Oprea, T. & Matter, H. (2004): «Integrating virtual screening in lead discovery», *Curr. Opin. Chem. Biol.,* 8.
89. Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D. & Weinberger, L. E. (1996): «Neighbourhood behaviour: a useful concept for validation of "molecular diversity" descriptors», *J. Med. Chem.,* (39**):** 3049-3059.
90. Pepperrell, C. A. & Willett, P. (1991): «Techniques for the calculation of the three-dimensional structural similarity using inter-atomic distances», *J. Comput.-Aided Mol. Design,* (5**):** 455-474.
91. Podani, J. 2001. Syn-Tax (2000): «Computer Programs for Data Analysis in Ecology and Systematics». User's Manual *Scientia Publishing, Budapest*.
92. Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1988). *Numerical Recipes in C,* Cambridge, UK, CambridgeUniversity Press.
93. Quinlan, R. R. (1986): «Induction of decision trees», *Machine Learning,* (1**):** 81-106.
94. Rasmussen, G. T., Isenhour, T. L. & Marshall, J. C. (1979): «Mass Spectral Library Searches Using Ion Series Data Compression». *J. Chem. Znf. Compur. Sci.,* (19**):** 98-104.
95. Ren, J. & Stammers, D. (2005): «HIV reverse transcriptase structures: designing new inhibitors and understanding mechanisms of drug resistance». *Trends*
96. *Referencia incompleta Pharmacol Sci,* (26**):** 4-7.
97. Rivera Borroto, O. M., Hernández Díaz, Y., García De La Vega, J. M., Grau Ábalo, R. D. C. & Marrero Ponce, Y. (2011ª): «Novel similarity measures for the effective and efficient retrieval of pharmacological datasets». *Afinidad,* 68.
98. Rivera Borroto, O. M., Hernández Díaz, Y., García De La Vega, J. M., Marrero Ponce, Y., Grau Ábalo, R. D. C., Rabassa Gutiérrez, M. & Rodríguez Abed, A.: Cribado

virtual en conjuntos de datos farmacológicos utilizando medidas de similitud para la recuperación efectiva y eficiente de los mismos, *In:* SILVA AYÇAGUER, L. C., ed. VIII Congreso Internacional de Informática en Salud y el II Congreso Internacional "Moodle Salud", 7-11 Febrero 2011b La Habana, Cuba. CITMATEL, SLD264.

99. Sadowski, J., Gasteiger, J. & Klebe, G. (1994): Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.,* (34**):** 1000-1008.
100. Salim, N., Holliday, J. & Willett, P. (2003): «Combination of fingerprint-based similarity coefficients using data fusion», *J Chem Inf Comput Sci,* (43**):** 435-442.
101. Sawilowsky, S. S. (2002): «Fermat, Schubert, Einstein, and Behrens-Fisher: The Probable Difference Between Two Means When σ_1^2 ≠ σ_2^2». *Journal of Modern Applied Statistical Methods,* (1)**:** 461-472.
102. 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. (1990): «Structure-activity relationships of antifilarial antimycin analogues, a multivariate pattern recognition study». *J. Med. Chem.,* (33**):** 136-142.
103. Sheridan, R. P. & Kearsley, S. K. (2002a): «Why do we need so many chemical similarity search methods?» *Drug Discov. Today,* 7.
104. Sheridan, R. P. & Kearsley, S. K. (2002b): Why do we need so many chemical similarity search methods? *Drug Discov. Today,* (7**):** 903-911.
105. Sheridan, R. P. & Kearsley, S. K. (2002c): Why do we need so many chemical similarity search methods?, *Drug Discov. Today* (7)**:** 903-911.
106. Sheridan, R. P., Singh, S. B., Fluder, E. M. & Kearsley, S. K. (2001), Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches. *J. Chem. Inf. Model.,* 41**,** 1395-1406.
107. Siegel, S. & Castellan, N. J. (1988): *Nonparametric statistics for the behavioral sciences,* New York, USA, McGraw-Hill.
108. Smeaton, A. F. & Van Rijsbergen, C. J. (1981): The Nearest Neighbour in Information Retrieval. An Algorithm Using Upperbounds. *ACM SZGZR Forum,* (16**):** 83-87.

109. SNAREY, M., TERRET, N. K., WILLETT, P. &WILTON, D. J. (1997): Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graph. Model,* (15**):** 372-385.

110. Sönströd, C., Johansson, U. & Norinder, U. Year. Generating Comprehensible QSAR Models. *In:* Johansson, R., Van Laere, J. & Mellin, J., eds. 3rd Skövde Workshop on Information Fusion Topics 2009, 2009 Skövde, Sweden. University of Skövde.

111. Sönströd, C., Johansson, U. & Norinder, U. Year. Generating Comprehensible QSAR Models. *In:* JOHANSSON, R., VAN LAERE, J. & MELLIN, J., eds. 3rd Skövde Workshop on Information Fusion Topics 2009 (SWIFT 2009), Oct 12-13 2009 Skövde, Sweden. University of Skövde, 44-48.

112. Steinbach, M., Ertöz, L. & Kumar, V. (2000): The Challenges of Clustering High Dimensional Data. *In:* WILLE, L. T. (ed.) *New directions in statistical physics: econophysics, bioinformatics, and pattern recognition.*Berlin/New York: Springer-Verlag.

113. Stevens, S. S. (1951): Mathematics, measurement, and psychophysics. *In:* STEVENS, S. S. (ed.) *Handbook of experimental psychology.*New York: Wiley.

114. Stine, W. W. (1989): Interobserver relational agreement. *Psychological Bulletin,* (106**):** 341-347.

115. Stine, W. W. (1989b): Meaningful inference: The role of measurement in statistics. *Psychological Bulletin,* (105**):** 147-155.

116. Suppes, P. & Zinnes, J. L. (1963): Basic measurement theory. *In:* LUCE, R. D., BUSH, R. R. & GALANTER, E. (eds.) *Handbook of mathematical psychology.*New York: Wiley.

117. Sutherland, J. J., O'brien, L. A. & Weaver, D. F. (2004). A Comparison of Methods for Modeling Quantitative Structure−Activity Relationships. *Journal of Medicinal Chemistry,* (47**):** 5541-5554.

118. Swamidass, S. J., Azencott, C.-A., Daily, K. & Baldi, P. (2010):. A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. *Bioinformatics,* (26**):** 1348-1356.

119. Talete Srl 2007. Dragon for Windows 5.5 ed. Milano, Italy.

120. Tetko, I. V., Villa, A. E. & Livingstone, D. J. (1996): Neural network studies. Variable selection. *J. Chem. Inf. Comput. Sci.,* (36): 794-803.

121. Todeschini, R. & ConsoNNI, V. (2009°): *Molecular Descriptors for Chemoinformatics,* Weinheim, Germany, WILEY-VHC.

122. Todeschini, R. & Consonni, V. (2009b): *Molecular Descriptors for Chemoinformatics,* Weinheim, Germany, WILEY-VCH

123. Todeschini, R., Galvagni, D., Vilchez, J. L., Del Olmo, M. & Navas, N. (1999): 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.,* (18): 93-98.

124. Truchon, J. & Bayly, C. I. (2007): Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.,* (47)**:** 488-508.

125. TverskY, A. (1977): Features of Similarity. *Psychological Review,* (84**):** 327-352.

126. Van Marlen, G. & Van Den Hende, J. H. (1979): Search Strategy and Data Compression for a Retrieval System with Binary-Coded Mass Spectra. *Anal. Chim. Acra,* (112**):** 143-150.

127. Voigt, J. H., Bienfait, B., Wang, S. & Nicklaus, M. C.(2001): Comparison of the NCI Open Database with Seven Large Chemical Structural Databases. *J. Chem. Inf. Comput. Sci. ,* (41**):** 702-712.

128. Watanabe, S. (1969): *Knowing and guessing: A quantitative study of inference and information,* New York, John Wiley & Sons Inc.

129. Wegner, J. K., Fröhlich, H., Mielenz, H. M. & Zell, A. (2006): Data and Graph Mining in Chemical Space for ADME and Activity Data Sets. *QSAR Comb. Sci.,* (25**):** 205-220.

130. Weislow, O. S., Kiser, R., Fine, D. L., Bader, J., Shoemaker, R. H. & Boyd, M. R. (1989): New Soluble-Formazan Assay for HIV-1 Cytopathic Effects: Application to High-Flux Screening of Synthetic and Natural Products for AIDS-Antiviral Activity. *Journal of the National Cancer Institute,* (81**):** 577-586.

131. Whittle, M., Willett, P., Klaffke, W. & Van Noort, P. (2003): Evaluation of Similarity Measures for Searching the Dictionary of Natural Products Database. *J. Chem. Inf. Comput. Sci. ,* (43**):** 449-457.

132. Willett, P. (1983): Some heuristics for nearest-neighbor searching in chemical structure files. *J. Chem. Inf. Comput. Sci. ,* (23**):** 22-25.

133. Willett, P. (2005): Searching Techniques for Databases of Two- and Three-Dimensional Chemical Structures. *Journal of Medicinal Chemistry,* (48**):** 4183-4199.

134. Willett, P. (2006a): Data fusion in ligand-based virtual screening. *QSAR Comb. Sci.,* (25**):** 1143-1152.

135. Willett, P. (2006b): Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today,* (11**):** 1046-1053.

136. Willett, P., BARNARD, J. M. &DOWNS, G. M. (1998a): Chemical similarity searching. *J. Chem. Inf. Comput. Sci. ,* (38**):** 983-996.

137. Willett, P., Barnard, J. M. &Downs, G. M. (1998b): Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.,* (38**):** 983-996.

138. Willett, P. & Winterman, V. (1986): A comparison of some measures for the determination of inter-molecular structural similarity. *Quant. Struct.-Activ. Relat.,* (5**):** 18-25.

139. Willett, P., Winterman, V. & Bawden, D. J. (1986): Implementation of nearest neighbour searching in an online chemical structure search system. *Chem. Inf. Comput. Sci. ,* (26**):** 36-41.

140. Winer, B. J. (1971): *Statistical principles in experimental design,* New York, McGraw-Hill.

141. Witten, I. H. & Frank, E. (2005): *Data Mining - Practical Machine Learning Tools and Techniques,* San Francisco, CA, Morgan Kaufmann.

142. Xue, L., Stahura, F. & Bajorath, J. (2004): Similarity search profiling reveals effects of fingerprint scaling in virtual screening. *J Chem Inf Comput Sci,* (44**):** 2032-2039.

143. Zegers, F. E. (1986): A family of chance-corrected association coefficients for metric scales. *Psychometrika,* (51**):** 559-562.

144. Zegers, F. E. & Ten Berge, J. M. F. (1985): A family of association coefficients for metric scales. *Psychometrika,* (50**):** 17-24.

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