Universidad de las Ciencias Informáticas
Facultad 2

Algoritmo basado en similitud de propiedades moleculares para obtener moléculas similares para una actividad biológica.

Trabajo de diploma para optar por el título de Ingeniero en Ciencias Informáticas

Autor: Jorge Nuñez Labadié

Tutores: MSC. Aurelio Collado Antelo
DrC. Ramón Carrasco Velar

La Habana, 2018

Declaración de autoría

Declaramos ser autores de la presente tesis y reconocemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

_____

Jorge Nuñez Labadié

Autor

_____

Dr. Ramón Carrasco Velar

Tutor

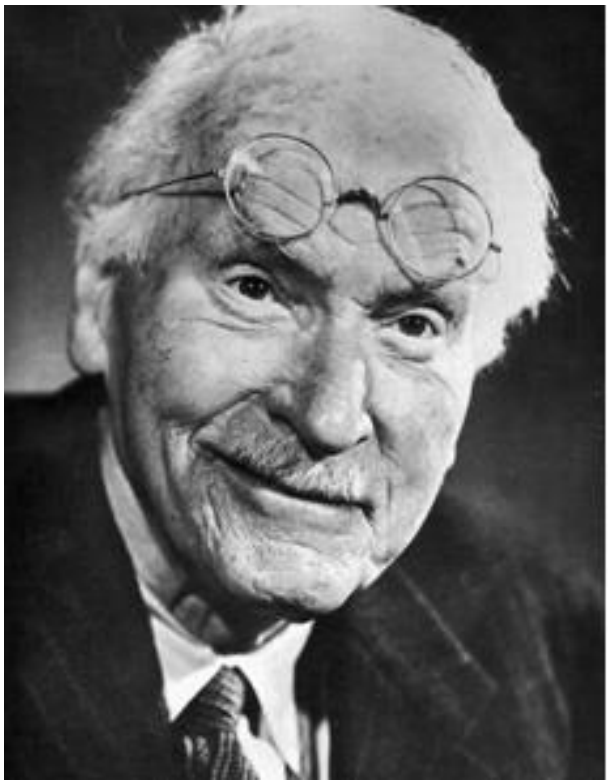
_____

MSc. Aurelio Antelo Collado

Tutor

FRASE:

El encuentro de dos personalidades es como el contacto de dos sustancias químicas: si hay alguna reacción, ambas se transforman.



Carl Gustav Jung

AGRADECIMIENTO:

Le agradezco a todas las personas que hicieron posible este trabajo, a mi tutor Aurelio que más que tutor ha sido un amigo y a su esposa Mayte. También me gustaría agradecerles a todas mis amistades a Rey, José Ramón, Cofiño y Yobi. A todas las personas que estuvieron arriba de mi a partir del día que decidí ser estudiante de esta Universidad, a mi segunda madre Ada Belkis y Yaniselis. Agradecerle a mi novia Yilena que ha estado durante estos dos últimos años apoyándome. A mi familia que han estado siempre pendientes de mí: mis tias Machi, Matuka y Tata, mis tíos José, Leandro y Yimi. También a todas aquellas personas que ahora no forman parte físicamente de mi familia, pero siempre van a formar parte de mi corazón, a mi papá y abuela Mercedes y no por ultimo deja de ser la menos importante agradecerle a mi madre que siempre ha estado arriba de mí y apoyándome en todo.

RESUMEN:

En el presente trabajo se propone un algoritmo capaz de calcular la similitud molecular entre compuestos moleculares utilizando diferentes propiedades química-física, este debe ser capaz de buscar moléculas similares en una colección de grafos moleculares utilizando descriptores híbridos ponderados por propiedades químico-físicas, y debe permitir distinguir diferencias estructurales entre los grafos moleculares. Se realiza la fragmentación del grafo químico utilizando una forma de grafo reducido que emplea agrupaciones de átomos que se denomina centros descriptores. Para la descripción de las estructuras moleculares se emplearon los índices de Estado Refractotopográfico, Electrotopográfico y Lipotopográfico para átomos. Se utilizó el concepto de Propiedad Máxima Común como criterio de identificación de moléculas con fragmentos que poseen valores similares de la propiedad descrita por los índices. Se implementó el método para los coeficientes de similitud de Tanimoto y Dice basados en el algoritmo de Propiedad Máxima Común utilizando los descriptores para átomos, validándose el mismo en el ensayo AID-941. Estos resultados se consideran la base para realizar minería de grafos en bases de datos con un gran número de estructuras moleculares lo que permitirá obtener patrones de similitud en diferentes colecciones de compuestos químicos.

Palabras claves: Índices híbridos, índices topográficos, grafos ponderados, similitud molecular, coeficientes de similitud, propiedad máxima común.

ÍNDICE

ÍNDICE DE FIGURAS.

ÍNDICE DE TABLAS.

INTRODUCCIÓN:

Hoy día no cabe duda que el arma más efectiva y la más frecuentemente usada para enfrentar con éxito casi todas las enfermedades son los medicamentos, los cuales constituyen un recurso médico y terapéutico. Hay datos que afirman que más del 75% de las patologías se pueden curar o evitar con fármacos (1), por eso se calcula que 3 de cada 4 personas con más de 75 años consume al menos 1 fármaco de prescripción, mientras que el 36% consume 4 o más (2). Además, hay muchas enfermedades que antes presentaban una elevada mortalidad a corto plazo, convirtiéndose en enfermedades crónicas, que, gracias al uso preventivo y curativo de los medicamentos existentes, se logra que el promedio de vida de las personas llegue a edades cada vez más altas mejorándose de esta forma la calidad de vida de la población.

A pesar de estos avances obtenidos con un uso de los medicamentos existente en la actualidad se estima que aproximadamente 3 mil millones de personas corren el riesgo de contraer enfermedades infecciosas, siendo un problema mayor en poblaciones con malas condiciones de vida y donde los tratamientos son inadecuados o inaccesibles (1). Reportándose que más del 70% de estas infecciones son resistentes al menos a 1 de los antibióticos que más frecuentemente se emplean para tratarlas (3), ya que ciertos microorganismos han ganado resistencia a los medicamentos antinfecciosos actuales.

La aparición de nuevas enfermedades infecciosas como el SIDA que han invadido amplias poblaciones en todo el mundo, el regreso de enfermedades mortales como la tuberculosis y la persistencia de la malaria, han aumentado su morbilidad y mortalidad debido al incremento de migraciones y viajes de la población mundial, así como al fenómeno de la resistencia a los antibióticos .Por otro lado las imperfecciones de los medicamentos existentes, lo que incluye los efectos secundarios, y la falta de medicamentos preventivos y curativos para las principales enfermedades señalan la importancia de la innovación en nuevos medicamentos en el futuro (1).

Pero el descubrimiento de un nuevo medicamento pasa por un proceso largo y complejo que demora de 10 a 15 años, requiere de nuevos descubrimientos en química y en biología y la realización de largos y costosos ensayos clínicos. El mismo es desarrollado por la industria farmacéutica, la cual debe destinar más de 1000 millones de dólares en proyectos I + D de un nuevo fármaco (4), representando esta cifra alrededor del 2% del producto interno bruto (PIB) de cada país. Pero a pesar del monto destinado por la industria farmacéutica, se ha observado por ejemplo, una progresiva disminución de la proporción relativa de fármacos aprobados anualmente por la Administración de Alimentos y Medicamentos (FDA)

con respecto a todos los que inician el proceso de desarrollo clínico, es decir, que se autorizaron únicamente 22 moléculas frente a las 53 que se aprobaron antes del año 2000, representando el 5% aproximadamente de nuevas moléculas con potencial terapéutico que llegan a iniciar la fase de ensayos preclínicos (5).

Lo más doloroso de todo esto, es que sólo una de cada 10 mil ensayos pasa la fase de desarrollo, una de cada 100 mil superan los ensayos clínicos y solo se logra 3 de cada 10 nuevos medicamentos registrarlo para recuperar su inversión inicial. Por lo que el diseño racional de fármacos, constituye una herramienta casi indispensable en el desarrollo actual de nuevos medicamentos, contribuyendo a un aumento de las posibilidades de éxito y a un decrecimiento de los costos. (6)

Frente a esta problemática, la industria farmacéutica se apoya cada vez más en los adelantos de la ciencias asociadas a la misma como: la química, la física, la biología, la bioinformática, entre otras y el desarrollo de nuevas tecnologías como: la genómica y la biología computacional en la identificación y validación de las dianas biológicas, el diseño y cribado virtual de potenciales candidatos moleculares, por citar algunas; que permiten desarrollar medicamentos más eficaces y seguros de una manera más eficiente, revolucionando y mejorando los procesos de producción de fármacos.

En los últimos años, la industria farmacéutica ha reorientado sus investigaciones hacia aquellos métodos que permitan el diseño computacional de nuevos compuestos. La efectividad de estos métodos depende en gran medida de los descriptores atómicos y moleculares seleccionados para caracterizar la estructura química, con el fin de predecir el comportamiento de estas para identificar los compuestos líderes entre el conjunto de compuestos.

Para identificar en química medicinal los compuestos o moléculas líderes entre un conjunto de compuestos, se utiliza el concepto de **similitud molecular**, por **Johnson y Maggiora** al principio de los 90 donde plantea: Moléculas estructuralmente similares tienden a exhibir propiedades biológicas similares. Aunque este concepto es intuitivo y está soportado por muchas observaciones, los químicos también han demostrado que pequeños cambios químico-estructurales en una molécula pueden modificar sus propiedades (7). Por tanto, se desprende que el reconocimiento de un patrón en un grupo de moléculas mediante el empleo de técnicas de análisis estadísticos o de algoritmos de inteligencia artificial aplicadas al conjunto de datos, dependerá en buena medida de la exactitud del análisis de los patrones de similitud.

En la actualidad se han desarrollado algoritmos de similitud molecular muy eficientes que permiten encontrar subgrafos en una colección de grafos. Entre los más conocidos se encuentran: los

desarrollados para encontrar subgrafos en una colección de grafos (8, 9, 10. 11), los de búsqueda de subgrafos maximales o cerrados (12, 13, 14), los que permiten obtener subgrafos estructuralmente diferentes en colecciones de grafos (15, 16, 17, 18) y por últimos los que utilizan el concepto de superficie máxima común (19, 20, 21, 22). Pero todos presentan las siguientes limitantes: incumplen en la práctica con el principio de similitud planteado por Johnson y Maggiora, el conjunto de subgrafos obtenidos no muestran información químico-físicas, no logran encontrar el subgrafo dentro del grafo químico, responsable de la actividad biológica y por último no identifican o descubren nuevas estructuras que puedan servir como compuestos líderes.

Por otra parte, en el 2004 Carrasco et al. (23), definieron un índice atómico novedoso para estudios de estructura-actividad QSAR (*Quantitative structure-activuty relationship*), denominado índice de estado refractotopológico, el cual permite diferenciar los subgrafos por los valores de la propiedad asociada al índice, lo que abre una nueva línea de investigación en el campo de la similitud molecular.

A partir de las carencias anteriormente descritas, surge el siguiente **problema científico**: ¿Cómo identificar el conjunto de moléculas similares en una colección de grafos moleculares, para una actividad biológica determinada utilizando propiedades químico-físicas?

Por lo que se tendrá como **objeto de estudio** la similitud molecular, enmarcado en el **campo de acción** búsqueda de moléculas similares en colecciones de grafos moleculares.

Para dar solución al **problema** se traza como **objetivo general**: Desarrollar un algoritmo de búsqueda de moléculas similares en una colección de grafos moleculares utilizando propiedades químico-físicas que permita diferencias estructurales entre ellas.

Para dar cumplimiento al objetivo general se realizaron los siguientes objetivos específicos:

1. Construir el marco teórico referencial de la investigación relacionado con la similitud molecular.
2. Diseñar un algoritmo para encontrar moléculas similares en una colección de grafos, utilizando la medida de similitud definida.
3. Desarrollar una herramienta computacional basada en el algoritmo propuestos.
4. Validar la propuesta, en ensayos utilizados en publicaciones referenciadas.

Para el desarrollo del presente trabajo se utilizaron los siguientes **métodos científicos de investigación**:

Teóricos:

- **Analítico-Sintético**: se emplea para buscar información acerca del problema propuesto y para extraer los elementos que están relacionado con el objeto de estudio.

Empíricos:

- **Consulta de las fuentes de información**: se emplean en la selección de la información importante y en la elaboración del marco teórico.
- **Consulta de especialistas**: para que las personas calificadas en el tema evalúen los resultados obtenidos con el algoritmo de similitud propuesto.
- **Pruebas**: se utilizan para comprobar si el algoritmo de similitud propuesto obtiene resultados aceptables.

El aporte principal de la presente investigación es, el diseño e implementación de un algoritmo de similitud molecular basado en el concepto de Propiedad Máxima Común (MCP), el cual utiliza descriptores híbridos ponderado por las propiedades químico-físicas (electrotopográfico, refractotopográfico y lipotopográfico), permitiendo encontrar moléculas similares en una colección de grafos.

Este documento está compuesto por un resumen, introducción, 3 capítulos que constituyen el cuerpo fundamental del documento, conclusiones generales, recomendaciones, bibliografía y referencias bibliográficas. Los capítulos son:

Capítulo 1: Fundamentación Teórica. En este capítulo se presenta un resumen de la investigación realizada sobre la similitud molecular en estructuras químicas. Se aborda en el desarrollo del mismo, el diseño y obtención de fármacos, así como las diferentes formas de reducción del grafo químico, los diferentes índices topográficos para átomos, la similitud molecular basada en descriptores y por último las funciones o coeficientes de similitud utilizados para obtener el cálculo de similitud molecular. Se señalan las tendencias actuales y el estado del arte a tener en cuenta.

Capítulo 2: Materiales y Métodos. En este capítulo se muestra la descripción de los métodos, procedimientos y algoritmos empleados, así como la justificación de su empleo. Se describen también los aspectos fundamentales tenidos en cuenta para la implementación del algoritmo propuesto.

Capítulo 3: Resultados y Discusión. En este capítulo se presentan y analizan los resultados de la investigación y las pruebas realizadas, se realiza una evaluación con otros algoritmos reportados en la literatura y por último se analizan los resultados obtenidos al aplicar el algoritmo propuesto a la colección de grafos del ensayo ADI-941.

CAPITULO 1. FUNDAMENTACIÓN TEORICA.

En este capítulo se verá un estudio general de las moléculas y sus aplicaciones para la obtención de fármacos. Se definen el concepto de un grafo químico y sus diversas formas de reducción. Además, se expondrá los principios de la similitud molecular y sus métodos la cual se abordará más en la similitud molecular basada en descriptores. Donde se exponen los conceptos de los índices topológicos y topográficos los cuales constituyen una herramienta utilizada en la química medicinal para establecer una relación entre estructura y propiedad, definiéndose los índices de estado Electrotopográfico, Refractotopográfico y Lipotopográfico. Finalmente se presenta una recopilación de las funciones de similitud / distancia.

1.1.- Diseño y obtención de fármacos.

En el diseño y síntesis de nuevos medicamentos, la predicción de la actividad biológica de compuestos orgánicos posee un papel fundamental. Para el desarrollo de la industria farmacéutica se hace indispensable el uso de métodos y herramientas que hagan cada vez más eficiente la búsqueda de nuevos fármacos para el tratamiento de las enfermedades. Presenta cuatro fases que van desde la fase de descubrimiento hasta la fase regulatoria, pasando por las fases preclínica y clínica.

En la fase de descubrimiento se trata de generar nuevas moléculas utilizando diferentes ensayos de bioactividad, abordando cuatro etapas: identificación de la diana terapéutica, validación de la diana, identificación del compuesto líder y validación del compuesto líder. Donde se parte de comprender cómo funcionan e influyen las dianas en una enfermedad específica para posteriormente definir la relación entre la diana seleccionada y la enfermedad de interés, luego se selecciona los compuestos que tienen potencial para tratar la enfermedad y por último obtener información para validar los compuestos y seleccionar los de mayor potencial.

Para identificar los compuestos líderes existen cuatro estrategias de diseño, las cuales parten de conocer la estructura en 3D de la proteína en estudio y la estructura de los ligandos presentes en los ensayos de bioactividad: 1) sin estructura 3D de proteína y sin ligando, para ello se utilizan las técnicas de química combinatoria, HTS y ensayos virtuales; 2) con estructura 3D de proteína y sin ligando, se emplean las técnicas de diseño de novo y flexibilidad de proteínas; 3) con estructura 3D de proteína y con ligandos, se aplica la técnica de diseño basado en estructura y por ultimo 4) sin estructura 3D de proteína y con ligandos, en esta estrategia se utilizan las técnicas de farmacóforos, la similitud y los estudios QSAR. Siendo esta última la más empleada por los químicos para el descubrimiento de nuevos compuestos líderes.

Aunque el número de compuestos conocidos sobrepasa los 26 millones, y un gran número de estos está disponible en diferentes bases de datos químicas, muchos de ellos no han encontrado todavía aplicaciones farmacológicas o de otro tipo, lo cual es consecuencia de la diferencia que existe entre la velocidad a la cual las nuevas moléculas son obtenidas y el número de ellas que pueden ser evaluadas en ensayos farmacológicos, toxicológicos y farmacocinéticos (24). El descubrimiento de los fármacos está sustentado en las propiedades de los compuestos que lo conforman y una característica determinante para los efectos deseados es la relación que existe entre la estructura química y la actividad biológica de los mismos. Uno de los propósitos más ambiciosos de la química moderna es encontrar esta relación entre la estructura molecular de productos orgánicos y la función biológica que cumplen.

1.2.- Reducción del grafo químico.

Los ligandos son moléculas pequeñas, cuya estructura química se puede asociar a un grafo matemático conexo y no dirigido, en el que los nodos son los átomos y las aristas son los enlaces entre estos. A partir del mismo se pueden definir diversos descriptores que relacionen las propiedades químico-físicas de las moléculas y la estructura tridimensional de las mismas (25).

La topología es una rama muy importante de las matemáticas que estudia aquellas propiedades de los objetos geométricos que tienen que ver con la "proximidad" o la "posición relativa" entre puntos (26). Es por ello que los estudios de estructura-actividad utilizan la teoría de grafos, basada en las propiedades topológicas de las moléculas, ya que con ella es posible expresar los vínculos que existen entre todos los átomos de la molécula. A partir de estas representaciones surgen los índices topológicos y topográficos como descriptores moleculares y atómicos (27).

Una representación más abstracta de las estructuras químicas se logra con los grafos reducidos. En esta forma de reducción, cada vértice representa un grupo de átomos conectados, y la arista que une dos de estos vértices. Un vértice en un grafo reducido puede representar un sistema de anillos, anillos aromáticos, anillos alifáticos o grupos funcionales (28). Teniendo en cuenta lo anterior, la reducción de grafos consiste en obtener un grafo de menor tamaño (menos aristas y/o vértices) con las características principales o relevantes del grafo original, de forma tal que se puedan realizar análisis sobre el grafo reducido y llegar a conclusiones sobre el grafo original (29). En la actualidad existen diversos sistemas para transformar una molécula en un grafo reducido, las cuales se muestran en la *Figura 1*. Donde en el nivel 1 los vértices en el grafo reducido corresponden a sistemas de anillos (R) y componentes acíclicos conectados (Ac), en el nivel 2 los vértices en el grafo reducido corresponden a átomos de carbono conectados (C) y están enlazados a heteroátomos (H), en el nivel 3 los vértices en el grafo

reducido corresponden a anillos aromáticos (Ar), anillos alifáticos (R) y grupos funcionales (F), en el nivel 4 los vértices en el grafo reducido corresponden a anillos aromáticos (Ar), grupos funcionales (F) y grupos conectados (L) y por último el nivel 5 los vértices en el grafo reducido corresponden a centros descriptores (CD) (anillos, cluster3, cluster4, heteroátomos, metilo, metileno, metino) (30), que se muestran en la *Figura 2*.

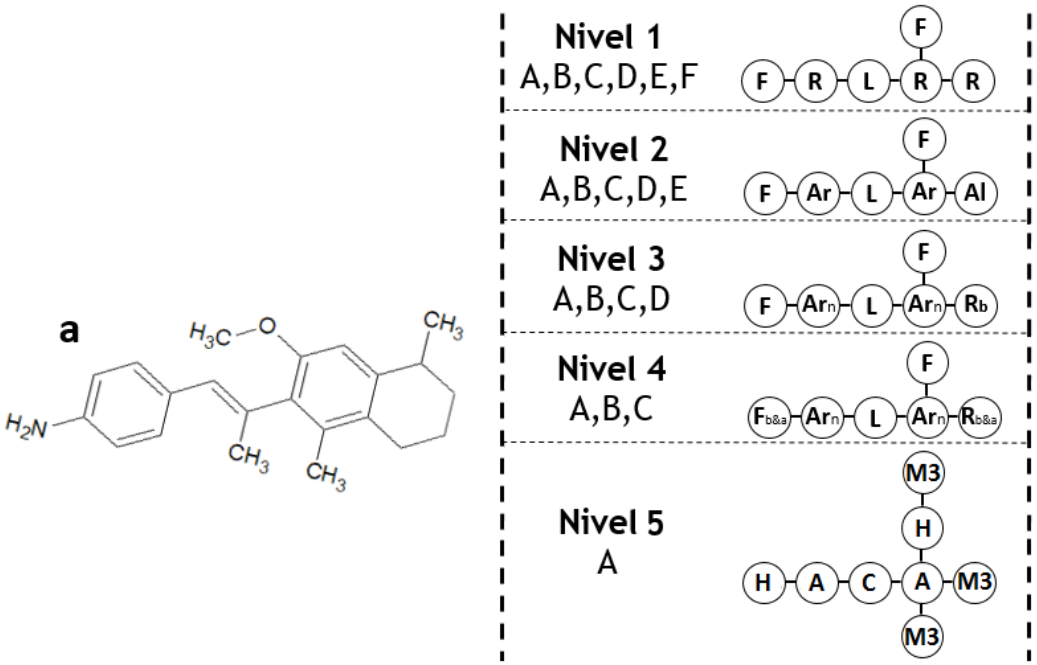


Figura 1. Niveles de reducción de un grafo químico.

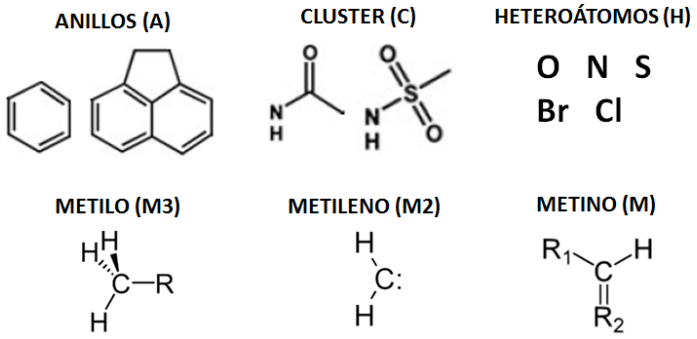


Figura 2. Centros descriptores (CD) utilizados en la fragmentación del grafo químico.

1.3.- Similitud molecular.

A principios de los 90 cuando se popularizó el análisis de similitud molecular, surgiendo el principio de propiedad de similitud (SPP), que establecía que los compuestos similares deberían tener propiedades

similares, siendo la propiedad más estudiada la actividad biológica. Aunque este principio fundamental suena bastante simple, es muy difícil de capturar metodológicamente, porque los compuestos que podrían no considerarse similares a menudo comparten actividad similar u otros valores de propiedad. Por el contrario, los compuestos que probablemente se considerarían muy similares podrían no serlo, lo que demuestra que pequeñas modificaciones químicas que conducen a cambios significativos en la actividad biológica, representa una limitación importante del enfoque SPP.

Es evidente que la similitud molecular es uno de los conceptos más explorados y explotados en la informática química y también es un tema central en la química médica (31,32). La misma es relevante para reconocer y organizar todos los componentes del entorno físico, así como muchos otros aspectos de la vida. Sin embargo, incluso en el mundo molecular más restringido, la similitud puede tener diferentes significados o interpretaciones dependiendo de nuestra perspectiva individual. Por lo tanto, si el objetivo final es describir formalmente la similitud de manera consistente a pesar de sus limitaciones intrínsecas, es de vital importancia distinguir primero entre diferentes criterios de similitud y conceptos.

Aunque los términos similitud química y molecular a menudo se utilizan como sinónimos, esto puede no ser del todo exacto. La similitud química se basa principalmente en las características fisicoquímicas de los compuestos (por ejemplo, solubilidad, punto de ebullición, log P, peso molecular, densidades electrónicas, momentos dipolares, etc.) mientras que la similitud molecular se centra principalmente en las características estructurales (por ejemplo, subestructuras compartidas, anillo sistemas, topologías, etc.) de compuestos y su representación. Las características fisicoquímicas y las características estructurales generalmente se explican por diferentes tipos de descriptores. Tales descriptores generalmente se definen como funciones matemáticas o modelos de propiedades químicas o estructura molecular. Para la evaluación de similitud química, también se puede considerar la información de reacción y diferentes grupos funcionales. En el trabajo actual, la atención se centra más en la similitud molecular que química.

La similitud se puede evaluar sobre la base de representaciones moleculares 2D y 3D. Los métodos de similitud 2D se basan en información deducida de gráficos moleculares. Las comparaciones gráficas directas (33) y los cálculos de similitud gráfica son computacionalmente exigentes y no se aplican ampliamente en el análisis de similitud molecular en la actualidad. Por el contrario, los descriptores moleculares que capturan información gráfica como el fragmento (34) o las huellas dactilares del entorno atómico (35) son muy populares. Las huellas dactilares generalmente se definen como la cadena de bits (34) o las representaciones del conjunto de funciones (35) de estructura y propiedades moleculares. Dichas representaciones moleculares se pueden comparar de manera eficiente computacionalmente,

permitiendo así cálculos de similitud a gran escala. Debido a que los compuestos son intrínsecamente tridimensionales y sus conformaciones moleculares tienen generalmente un contenido de información más alto que sus gráficos moleculares correspondientes, se puede anticipar que la similitud 3D, que implica la comparación de conformaciones moleculares y propiedades asociadas, (36,37) debería preferirse generalmente a la similitud 2D. Sin embargo, este no es el caso por dos razones principales. En primer lugar, los químicos se entrenan sobre la base de gráficos moleculares (es decir, representaciones estructurales 2D) y, en general, se sienten más cómodos basando sus consideraciones en los gráficos que en las estructuras tridimensionales de los compuestos. Los gráficos moleculares típicamente usados por los químicos a menudo también contienen información conformacional y estereoquímica. En segundo lugar, dadas las incertidumbres asociadas con la identificación de conformaciones biológicamente activas en grandes conjuntos conformacionales de compuestos de prueba, los enfoques 2D son típicamente más robustos, a pesar de su relativa simplicidad, y a menudo producen resultados superiores en análisis SAR y predicción de actividad. (38,39) Muchos métodos actuales de similitud utilizar preferencialmente representaciones moleculares 2D; la mayoría, sin embargo, no contienen información estereoquímica, lo que limita su capacidad para tratar adecuadamente los compuestos enantioméricos. Como tales compuestos tienen una conectividad atómica idéntica, sus valores de similitud serán la unidad si se usan representaciones moleculares estereoinsensibles. Además, como se analizará a continuación en detalle, los cálculos de similitud sobre la base de representaciones moleculares 2D tienen una serie de otras limitaciones intrínsecas.

Otro concepto de similitud que requiere consideración es la similitud biológica de los compuestos, que se aparta del marco conceptual del SPP. En cambio, los descriptores de propiedades estructurales o fisicoquímicas usuales son reemplazados por las actividades de los compuestos contra un panel de blancos de referencia, generalmente proteínas, que proporcionan "firmas biológicas" (40, 41) análogas a las representaciones basadas en estructura o propiedad extensamente discutidas aquí. En este caso, los perfiles de actividad correspondientes a las firmas biológicas de los compuestos se comparan usando una función de similitud apropiada como una medida de similitud por pares, independientemente de las características estructurales de los compuestos. Por lo tanto, en este caso, la similitud biológica se evalúa en el espacio objetivo en lugar de espacio químico.

Para el análisis SAR y los programas de química médica, la similitud biológica es generalmente más difícil de implementar que las representaciones basadas en estructuras o propiedades porque los valores de actividad específicos pueden no estar disponibles para los compuestos de interés. Además de su uso como medidas de similitud molecular, las firmas biológicas también pueden proporcionar una medida aproximada de promiscuidad compuesta. (42) Por ejemplo, al sumar los

valores individuales en una firma biológica binaria (activo = 1 o inactivo = 0) se obtiene el número de objetivos contra que el compuesto asociado exhibe actividad.

Un criterio muy importante para el análisis de similitud es distinguir entre vistas de similitud global y local. Por ejemplo, la comparación de modelos de farmacóforos en el diseño de fármacos se centra únicamente en átomos, grupos o funcionalidades seleccionados que son conocidos o que se cree que son responsables de la actividad. Esto representa una visión local de la similitud, en contraste con la visión más general que se encuentra típicamente en la informática química, donde los compuestos se consideran en su totalidad. En el último caso, la propiedad calculada o los descriptores estructurales típicamente usados para calcular las similitudes moleculares generalmente se derivan de información estructural asociada con compuestos enteros. Por ejemplo, si traducimos la información estructural de un compuesto en una huella digital de fragmento, se obtiene una representación molecular global. Esta visión completa de la similitud es característica de la perspectiva de los quimioinformáticos.

Además de los puntos de vista locales y globales, sin embargo, también se debe prestar especial atención a la perspectiva de un químico farmacéutico en este contexto. Considere, por ejemplo, el conjunto de inhibidores bien conocidos de la ciclooxigenasa (COX) comparados en la *Figura 3*. Todos estos inhibidores son fármacos aprobados excepto el lumiracoxib, que perdió su aprobación en los Estados Unidos en 2007. Si aplicamos una vista de compuesto completo, los compuestos como los enantiómeros de ibuprofeno, ibuprofeno y paracetamol, o diclofenaco y lumiracoxib, aparecen visiblemente similares. Sin embargo, desde el punto de vista de la química médica, esta evaluación puede no estar generalmente de acuerdo ya que pequeñas diferencias químicas pueden conducir a cambios importantes en los perfiles de especificidad (por ejemplo, diclofenaco frente a lumiracoxib) o pueden sintetizarse o derivatizarse en diferentes compuestos formas (por ejemplo, ibuprofeno vs paracetamol). Por lo tanto, la visión de semejanza de un químico medicinal podría volver a ser de naturaleza más local y / o tener en cuenta la información de reacción química directamente.

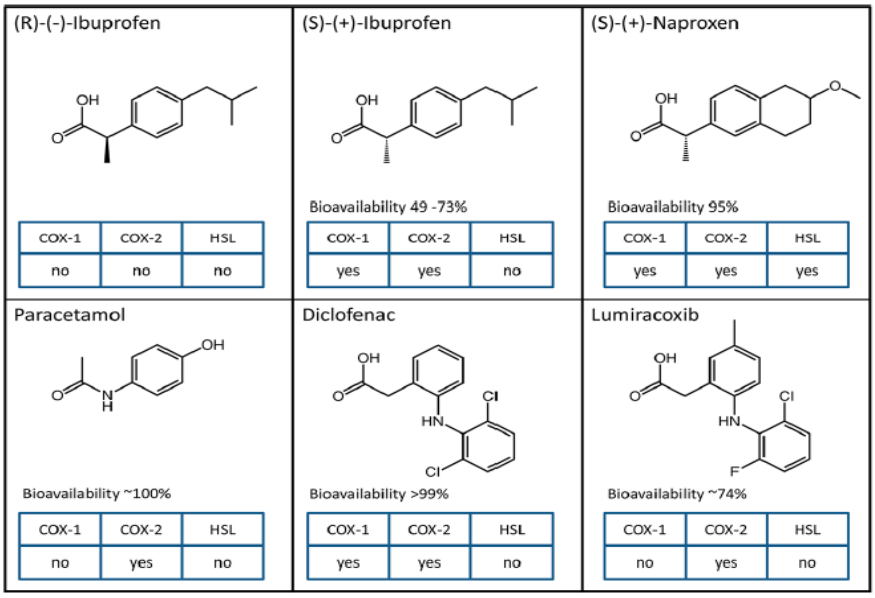


Figura 3.Relaciones de similitud complejas. Se comparan los inhibidores de la ciclooxigenasa (COX) y sus perfiles de actividad. HSL significa hormonesensible lipasa.

Tales ejemplos ilustran que las consideraciones de los criterios químicos y funcionales podrían alterar fácilmente la percepción del parecido molecular global. Claramente, tales consideraciones de similitud caen en una zona gris, ya que están influenciadas por criterios subjetivos, así como por la experiencia del investigador y, por lo tanto, no existe una forma generalmente aceptada de juzgar tales relaciones de similitud.

Para cuantificar el valor de similitud molecular se pueden emplear diferentes enfoques, los cuales se pueden dividir en dos categorías principales: los cálculos de similitud basados en descriptores y la evaluación de la similitud basada en subestructura (31), en los epígrafes siguientes profundizaremos en estas categorías.

1.4.- Similitud molecular basada en descriptores

Los cálculos de similitud basados en descriptores cuantifican la similitud a través de la comparación de las representaciones de descriptores moleculares, en particular las huellas dactilares de compuestos (43,44). Estas representaciones pueden ser codificadas por diferentes tipos de descriptores moleculares obtenidos a través de la estructura topológica o topográfica de los compuestos químicos, los cuales se pueden clasificar en descriptores topológicos o topográficos.

Se conocen como descriptores, los cuantificadores matemáticos que relacionan la estructura molecular y las propiedades físico-químicas de los compuestos a partir de parámetros estructurales simples, lo que posibilita interpretar las propiedades moleculares y describir el comportamiento de las sustancias. Los

descriptores son utilizados para caracterizar la estructura química de un compuesto y la calidad de los mismos condiciona el éxito de los modelos matemáticos que describan los fenómenos biológicos. (45)

En el campo de los descriptores se conjugan diferentes disciplinas como el álgebra, la teoría de grafos, la teoría de la información, la química computacional, las teorías de la reactividad química y la química-física, jugando un importante papel, además, la programación y el software y hardware empleados para su obtención. (46)

En la actualidad existen una gran cantidad de descriptores, cuyo número sobrepasa los miles y están distribuidos según el tipo. Poseen varios enfoques y principalmente han sido empleados en el diseño de fármacos y en estudios de relación estructura-propiedad. Entre los tipos de descriptores más conocidos se pueden citar los constitucionales, éstos son derivados simplemente de la fórmula química, por ejemplo el peso molecular o el número de átomos de nitrógeno de una estructura; los geométricos que tienen en cuenta el análisis de superficie, cálculo de ángulos y distancia entre grupos, los lipofílicos que miden la tendencia de un compuesto determinado, a formar enlaces hidrofóbicos; los electrónicos que tienen en cuenta la carga eléctrica; los estéricos que se obtienen a partir de la forma del sustituyente y el volumen molar; los topológicos que indican una caracterización matemática de una molécula, donde los sitios ocupados por átomos son reemplazados por los vértices y las conexiones entre ellos por aristas conformando el grafo químico, en este tipo de descriptores existe cierta pérdida de información pues se representa un objeto tridimensional por un número simple, los topográficos que son semejantes a los topológicos pero tienen en cuenta la estructura en tres dimensiones del compuesto y los híbridos que son en los que se combinan aspectos estructurales de los compuestos con propiedades físico-químicas particionadas sobre grupos de átomos. Estos descriptores se dividen también en dos grandes clasificaciones, atómicos y moleculares en dependencia del tipo de estructura que describen.

Todos estos descriptores pueden ser aplicables a moléculas, átomos y fragmentos en dependencia de la investigación que se esté desarrollando y las preferencias o conocimientos de los especialistas. Con su empleo se genera una lista de valores numéricos que permite caracterizar a las moléculas. El empleo de los estudios QSAR en investigaciones apoyadas en la teoría de grafos ha sido utilizado para la obtención de modelos aditivos y de regresión para la predicción de propiedades químicas, físicas y biológicas de forma efectiva (47).

1.5.- Índices topográficos para átomos.

Los índices topológicos y topográficos constituyen una herramienta ampliamente utilizada en la química medicinal para el establecimiento de relaciones entre la estructura y la propiedad. Existen numerosos programas que los calculan, de los cuales, el más popular es el Dragon (48). Sin embargo, existe otro

tipo de índices, denominados híbridos, que poseen contenido de información topográfica y de propiedades químico-físicas los cuales se denominan híbridos por esta razón. Los descriptores topológicos y topográficos clásicos no poseen otra forma de representación gráfica que la típica del grafo. Sin embargo, los descriptores híbridos, poseen la capacidad de brindar información dual (estructural y de propiedad). Otra diferencia sustancial entre estos es que los híbridos están basados en la matriz de conectividad del grafo completo y no del grafo desprovisto de hidrógeno, como es usual en los índices topológicos. Inicialmente se reportaron dos índices de naturaleza híbrida, el de Partición de la Refractividad Molecular y el Índice del Estado Refractotopológico para Átomos (49). Este último ha demostrado su aplicabilidad en estudios de relación estructura-actividad. A partir de este último se han definido nuevos índices híbridos que lo complementan, estos son el Índice de Estado Refractotopográfico para Átomos y los Índices Lipotopológico y Lipotopográfico para Átomos. Otro índice, que a pesar de no ser de naturaleza hibrida, también está incluido en esta investigación es el Índice de Estado Electrotopográfico para Átomos (50). Estos descriptores topográficos son los que se utilizarán en la búsqueda de similitudes entre moléculas y fragmentos moleculares, siendo de vital importancia para el desarrollo de esta investigación.

1.5.1.- Índice del Estado Electrotopográfico.

El Índice de Estado Electrotopográfico para átomos (50) (Sstate, Si), se basa en el efecto electrónico de cada átomo sobre los otros átomos en la molécula. El mismo se calcula por la expresión $Si=Ii+\Delta Ii$, donde Si es el valor del índice para el átomo i, Ii es un valor intrínseco asociado al átomo i y ΔI_i expresa el efecto perturbativo de los restantes átomos j en la molécula sobre el átomo i. El valor intrínseco I_i de cada átomo se calcula por la ecuación:

$$I_i=[(2/N)_2\delta_v+1]/\delta$$

donde N es el número cuántico principal del átomo i, δ_v es el número de electrones de valencia en el esqueleto molecular (Z_v-h) y δ es el número de electrones σ en el esqueleto (σ - h). Para cada átomo en el esqueleto molecular, Z_v es el número de electrones de valencia, σ es el número de electrones en orbitales σ y h es el número de hidrógenos enlazados a éste. El efecto perturbativo sobre el átomo i producido por los restantes átomos pesados presentes en la molécula se calcula según la ecuación:

$$\Delta I_i=\Sigma(I_i-I_j)/r_{i,j2}$$

donde $r_{i,j2}$ es la distancia euclideana entre los átomos i y j tomada de la matriz de distancias, correspondiente a la configuración de mínimo energético calculada por algún método semiempírico. El Sstate permite considerar información sobre la estructura tridimensional de los compuestos, al considerarse como distancia entre los átomos, no la topológica, sino la que se obtiene de la optimización de geometría.

1.5.2.- Índice de Estado Lipotopográfico

El Índice de Estado Lipotopográfico para átomos (50) (Lstate3D, Λ_{3D}) representa la solubilidad en grasas de la molécula, y se define por la ecuación:

$$\Lambda_{3D} = AL_i + \Delta AL_{ij}$$

donde AL_i es el valor intrínseco de solubilidad en grasas del átomo i y ΔAL_{ij} representa el término de perturbación definido por la ecuación:

$$\Delta AL_{ij} = \Sigma (AL_i + AL_j) / r_{i,j2N j=1}$$

donde se suman todos los vértices j adyacentes en el grafo químico, AL_i y AL_j son los valores intrínsecos de solubilidad en grasas de los átomos i y j respectivamente, y $r_{i,j2}$ es la distancia euclidiana entre los átomos i y j, calculado a partir de la estructura optimizada con algún método semiempírico.

1.5.3.- Índice de Estado Refractotopográfico

El Índice de Estado Refractotopográfico para Átomos (50) (R-state3D, $\Re_{3D}$), se basa en la influencia de las fuerzas de dispersión de cada átomo sobre cada uno de los restantes en la molécula, modificado por la topología molecular. El mismo para un átomo i se define por la ecuación:

$$\Re_{3D} = AR_i + \Delta AR_i$$

donde AR_i es el valor de refractividad intrínseco del átomo i y ΔAR_i es un término de perturbación definido por la ecuación:

$$\Delta AR_{ij} = \Sigma (AR_i + AR_j) / r_{i,j2N j=1}$$

donde se suman todos los vértices j adyacentes en el grafo, AR_i y AR_j son los valores intrínsecos de la refractividad de los átomos i y j respectivamente, y $r_{i,j2}$ es la distancia euclidiana entre los átomos i y j, calculado a partir de la estructura optimizada con algún método semiempírico.

1.6.- Función de similitud.

Como resultado del esfuerzo por cuantificar la asociación o similitud en varios campos de la ciencia, se han creado una gran variedad de medidas. Se encuentran actualmente en la literatura diversos índices para medir la similitud / diferencia entre individuos, objetos o unidades experimentales, de forma tal, que cuantifican el grado de asociación o semejanza entre cada par de elementos, algunos de los cuales también se pueden emplear para comparar variables. Muchas de estas medidas surgieron a partir de mejoras realizadas a otras de las funciones ya existentes y fueron adaptadas a las particularidades de los entornos para el que fueron diseñadas (51).

Los valores que se obtienen de los coeficientes de similitud varían entre cero (0) y uno (1), siendo el valor 1 el de máxima similitud y el valor 0 el de mínima, mientras que la distancia se puede calcular como un complemento de la similitud. Una distancia alta entre individuos nos indica que son muy diferentes y una baja que son muy similares; los indicadores de similitud actúan de manera contraria: conforme aumente su valor, aumentará la similitud entre los individuos.

Las medidas de proximidad, similitud o semejanza miden el grado de parecido entre dos objetos de forma que, cuanto mayor es su valor, mayor es el grado de semejanza existente entre los objetos. Por otra parte, las medidas de diferencia, desemejanza o distancia miden la distancia entre dos objetos de forma que, cuanto mayor sea su valor, más diferentes son los objetos. En la literatura existen multitud de medidas de semejanza y de distancia dependiendo del tipo de variables y datos considerados, las cuales se muestran en la *Tabla 1*.

Tabla 1. Funciones de similitud y de distancia.

No	Nombre	Fórmula	Intervalo
1	Sørensen	$\frac{\sum \lvert a_i - b_i \rvert}{\sum (a_i + b_i)}$	[1,0]
2	Tanimoto	$\frac{\sum a_i + \sum b_i - 2\sum \min(a_i,\ b_i)}{\sum a_i + \sum b_i - \sum \min(a_i,\ b_i)}$	[1,0]
3	Soergel	$\frac{\sum \lvert a_i - b_i \rvert}{\sum \max(a_i, b_i)}$	[1,0]
4	Czekanowski	$\frac{2 * \sum \min(a_i, b_i)}{\sum (a_i + b_i)}$	[0,1]
5	Jaccard	$\frac{\sum a_i * b_i}{\sum a_i^2 + \sum b_i^2 - \sum a_i * b_i}$	[0,1]
6	Ruzicka	$\frac{\sum \min(a_i, b_i)}{\sum \max(a_i, b_i)}$	[0,1]
7	Dice-Sorensen	$\frac{2 * \sum a_i * b_i}{\sum a_i^2 + \sum b_i^2}$	[0,1]
8	Dice	$\frac{\sum (x_{1j} x_{2j})}{(1/2)\left(\sum x_{1j}^2 + \sum x_{2j}^2\right)}$	[0,1]

1.7.- Algoritmos de obtención de grafos similares.

En la actualidad se han desarrollado algoritmos muy eficientes que permiten encontrar subgrafos en una colección de grafos, entre los más conocidos se encuentran: gSpan(8), Gaston(9), gRed(10) y GraphSig(11). Estos algoritmos permiten encontrar el conjunto completo de subgrafos frecuentes en una colección de grafos, que puede ser muy grande para colecciones moderadas de grafos, además identifican la ocurrencia de un subgrafo candidato resolviendo el problema de isomorfismo de grafos o subgrafos.

Con el objetivo de encontrar un subconjunto significativo de subgrafos, se desarrollaron los algoritmos: ClaseGraph(12), Spin(13) y Margin(14). Estos algoritmos se concentran en la búsqueda de subgrafos maximales o cerrados, pero a pesar de la reducción de la cantidad de subgrafos encontrados sigue siendo trabajoso su análisis por un experto.

Existen otros algoritmos que permiten obtener subgrafos estructuralmente diferentes en colecciones de grafos, entre ellos se encuentran: gApprox(15), APGM(16), GraMi(17) y AGraP(18). Estos algoritmos encuentran un subconjunto significativo, es decir, una menor cantidad de subgrafos que, de alguna manera, retienen la información del conjunto completo de subgrafos.

Por último, se han desarrollado algoritmos que utilizan el concepto de Superficie Máxima Común (MCS), en los que se encuentran: McGregor(19), SMSD(20), MultiMCS(21) y FMCS(22). Estos algoritmos encuentran la parte de la estructura química que es similar estructuralmente, reduciendo considerablemente la cantidad de subgrafos obtenidos en una colección de grafos.

1.8.- Conclusiones del capítulo.

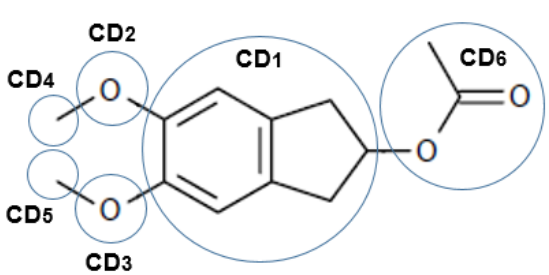
En este capítulo se presentaron las tendencias actuales de la similitud molecular, áreas en los que los investigadores han centrado la atención, diseñando diversos métodos para la búsqueda de moléculas que posean una correlación entre la estructura y la actividad biológica. Además, se expusieron aspectos fundamentales de los grafos químicos y las diferentes estrategias de reducción de los mismos. Para finalizar se mostraron los términos asociados a los descriptores moleculares y los índices híbridos ponderados por propiedades químico-físicas que se proponen en la bibliografía, las funciones de similitud (distancia) para el cálculo de semejanzas entre estructuras reportadas y por último los algoritmos reportados en la literatura utilizados para la obtención de grafos o subgrafos similares.

CAPITULO 2. MATERIALES Y MÉTODOS

En este capítulo se expondrá el procedimiento de fragmentación de un grafo químico, los descriptores y las funciones de similitud utilizadas, así como estas fueron modificadas basada en la Propiedad Máxima Común(MCP). Igualmente se explican los algoritmos del cálculo de similitud molecular mediante MCP, la búsqueda de fragmentos similares por MCP, la reducción de un grafo molecular, la normalización de los descriptores moleculares y el cálculo de similitud. Definiendo el lenguaje de programación y el entorno de desarrollo utilizado, definiendo las librerías y Dataset utilizados, para en el final del capítulo se presenta imágenes de la aplicación utilizando los algoritmos propuestos en este capítulo.

2.1.- Fragmentación del grafo químico.

A partir del estudio de las diferentes formas de reducción del grafo químico reportadas en la literatura, se seleccionó la que representa los vértices en el grafo reducido como centros descriptores (CD) (anillos, cluster3, cluster4, heteroátomos, metilo, metileno, metino) por ser la que mayor información químico-física ofrece siendo esto uno de los objetivos de la presente investigación. Mediante la misma se define un fragmento molecular de orden n a la combinación de n CD relacionados entre sí por la distancia euclidiana entre sus respectivos centros de masas (CM). La cantidad de fragmentos de orden k se pueden obtener utilizando la ecuación presente en la *Figura 4*.



Cantidad de fragmentos de orden n.

$$\mathbf{CFrag_k = C_k^n = \frac{n!}{(n-k)! * k!}}$$

donde:

n: cantidad de centros descriptores.
k: orden del fragmento.

Figura 4. Fragmentación de una molécula en fragmentos de orden k.

2.2.- Descriptores utilizados.

La descripción de las estructuras moleculares es una de las tareas fundamentales en el pre-procesamiento del grafo molecular para realizar los cálculos de similitud molecular. En el presente trabajo se utilizó para la descripción de los diferentes compuestos, índices híbridos ponderados por propiedades químico-físicas tales como: electrónicas, estéricas y lipofílicas. Por lo que, se emplearon los índices para átomos: Refractotopográfico (S_{3D}), Electrotopográfico ($\mathcal{R}_{3D}$) y Lipotopográfico (Λ_{3D}), los cuales abordan las propiedades mencionadas anteriormente.

2.3.- Función de similitud utilizada.

A partir del estudio de las diferentes funciones de similitud presente en la bibliografía, se seleccionaron aquellas que mejor se adaptan al cálculo de similitud molecular, basados en vectores de valores. En la *Figura 5* se presentan los coeficientes de similitud, que serán empleados en esta investigación.

Coeficiente Dice $$S_{\text{Dice}} = \frac{\sum(x_{1j}x_{2j})}{(1/2)\left(\sum x_{1j}^2 + \sum x_{2j}^2\right)}$$

Coeficiente Tanimoto $$S_{\text{Tan}} = \frac{\sum x_{1j}x_{2j}}{\sum x_{1j}^2 + \sum x_{2j}^2 - \sum x_{1j}x_{2j}}$$

donde:

X1: vector de propiedades químico-físicas del fragmento 1.
X1: vector de propiedades químico-físicas del fragmento 2.

Figura 5. Funciones de similitud empleadas.

2.4.- Propiedad Máxima Común.

Antes de definir el concepto de Propiedad Máxima Común (MCP) es necesario tener presente algunos aspectos de la teoría de grafos. Un grafo molecular G = (V,A) consiste de un conjunto de vértices V(G), (átomos en una molécula) y un conjunto de aristas A(G), (enlaces en una molécula).Un grafo molecular G_m consiste de un conjunto de vértices V(G_m) y un conjunto de aristas A(G_m),los vértices en G_m están conectados por una arista si existe una arista (v_i, v_j) ∈ A(G_m) que conecta los vértices v_i y v_j en G_m de forma tal que v_i, v_j ∈ V(G_m).

Un clique (ω) en un grafo molecular G_m puede ser definido como un subconjunto de vértices de tal manera que cada par de vértices esté conectado por una arista en el grafo. Un subgrafo $G'_m \subseteq G_m$ es completo si u, v ∈ A para todos los u, $v \in G(V')$. Se dice que es máximo si no es un subgrafo de un subgrafo mayor en G_m. Un clique máximo $\omega(G_m)$ en un grafo molecular, es el fragmento de un grafo que no es un subgrafo de un clique mayor en G_m, por lo tanto este subgrafo se conoce como el Subgrafo Máximo Común (MCS) (52).

Un grafo reducido se define como el conjunto de centros descriptores (CDs), obtenidos a partir del método de reducción propuesto en la sección 1.2, la representación tridimensional de las moléculas permite obtener el centro de masa de los CDs como un punto (x, y, z) y calcular la distancia entre estos. Como se definió anteriormente (sección 2.1), un fragmento molecular, es un conjunto de CDs obtenido mediante la teoría combinatoria entre los CDs del grafo reducido. Por lo antes planteado y partiendo del concepto de MCS, en la definición 1 se expone que se entiende por Propiedad Máxima Común.

Definición 1: Dados los grafos G_1 y G_2, se entiende por fragmentos con Propiedad Máxima Común MCP_{HD} (G_1, G_2), a los subgrafos g_1 y g_1 de los grafos G_1 y G_2 que presentan la máxima similitud en las propiedades químico-físicas representadas por los índices (S_{3D}, A_{3D}, R_{3D}), entre los Centros Descriptores (CDn) y la distancia euclidiana entre sus centros de masa ($d_E(CD_1, CD_2)$).

Mediante las funciones de similitud propuestas a utilizar por el epígrafe anterior, se modificaron aplicando la Propiedad Máxima Común, las cuales se verán en la *Figura 6* .

$$\mathrm{D_{MCP}} = \frac{|\mathrm{MCP}(A,B)|_b}{1/2\left(|A|_b + |B|_b\right)}$$

$$\mathrm{Tc_{MCP}}(A,B) = \frac{|\mathrm{MCP}(A,B)|_b}{|A|_b + |B|_b - |\mathrm{MCP}(A,B)|_b}$$

Donde:

$|\mathrm{MCP}(A,B)|_b$: **Propiedad máxima común entre las moléculas A y B**

$|A|_b$: **Propiedad exclusiva de la molécula A**

$|B|_b$: **Propiedad exclusiva de la molécula B**

Figura 6. Funciones de similitud modificadas por Propiedad Máxima Común.

2.5.- Algoritmos implementados

2.5.1.- Calculo de similitud molecular por Propiedad Máxima Común.

El algoritmo para el cálculo de similitud molecular por Propiedad Máxima Común permite encontrar los fragmentos moleculares que posean valores similares de las propiedades dadas por los índices y una distribución atómica que pueden ser o no semejantes a la molécula diana. Se toma como ejemplo el empleo de funciones de similitud para la descripción de los pasos del algoritmo.

En la *Tabla 2* se muestra el pseudocódigo del algoritmo para determinar si dos moléculas poseen un fragmento de Propiedad Máxima Común y por consiguiente comparten cierto grado de semejanza estructural y por propiedades químico-físicas. En el paso 1 y 2 se calculan los descriptores (electrotopográfico, refractotopográfico y lipotopográfico utilizando las ecuaciones descritas en el epígrafe 1.5) de las moléculas M1 y M2 y se almacenan en los grafos G1 y G2 respectivamente, acto seguido se realiza el paso 3 y 4 donde se obtienen los grafos químicos reducidos GR1 y GR2 (utilizando los CD descriptos anteriormente) pertenecientes a las moléculas M1 y M2 respectivamente.

En los pasos 5 y 6 se obtienen la lista de CD presente en cada uno de los grafos GR1 y GR2 y se almacenan en las variables CD1 y CD2 respectivamente; en el paso 7 se normalizan los descriptores con el objetivo de eliminar los valores negativos, el resultado se almacena en los grafos GR1 y GR2. Luego en el paso 8 se obtiene el fragmento de los grafos GR1 y GR2 que presentan el valor máximo común de propiedad y se almacenan en las variables F1 y F2, respectivamente. A continuación, en los pasos 9 y 10 se generan los vectores de distancia euclideana V1 y V2, que contendrán las distancias entre los centros descriptores de cada uno de los fragmentos F1 y F2, respectivamente. Por último, en el paso 11 se calcula el coeficiente de similitud entre los vectores mediante una de las funciones de similitud seleccionadas y en el paso 12 se retorna dicho valor.

Tabla 2.Algoritmo del cálculo de similitud molecular por PMC.

```
Algoritmo molecularSimilarityMCP (M1, M2, u)
Inicio
    1. G1  ← descriptordCalculation(M1)
    2. G2  ← descriptordCalculation(M2)
    3. GR1 ←reducedGraph(G1)
    4. GR2 ← reducedGraph(G1)
    5. CD1 ← { centros descriptores del grafo reducido GR1 }
    6. CD2 ← { centros descriptores del grafo reducido GR2 }
    7. (GR1, GR2) ← descriptorNormalization(G1,G2)
    8. (F1, F2) ← searchFragmentMCP(CD1,CD2, sinFunc, error, index)
    9. V1  ← { vector distancia euclideana del fragmento F1 de la molécula M1 }
    10. V2  ← { vector distancia euclideana del fragmento F2 de la molécula M2 }
    11. CS  ← CalculateSimilarity(V1, V2)
    12. Retornar CS
Fin
```

2.5.2.- Búsqueda de fragmentos similares por Propiedad Máxima Común.

El algoritmo para la búsqueda de fragmentos por Propiedad Máxima Común permite encontrar los fragmentos moleculares F1 y F2 utilizando la lista de centro descriptores CD1 y CD2, el error permitido y el índice a utilizar. A continuación, describiremos paso a paso el algoritmo.

En la *Tabla 3* se muestra el seudocódigo mediante el cual se buscan los fragmentos similares utilizando el concepto de propiedades máxima común, ponderada por propiedades químico-físicas, para ello en el paso 1 se obtiene la matriz de similitud topográfica que almacena en cada celda el índice de similitud entre cada uno de los centros descriptores de ambas moléculas representado en cada fila y columna, ponderados por propiedades químico-físicas; este índice de similitud se calcula empleando el coeficiente de similitud de Tanimoto cuya ecuación fue abordada en el epígrafe 2.3. En el paso 2 se inicializa una variable contadora VM en cero para ser utilizada posteriormente.

En el paso 3 se inicia un ciclo que termina en el paso 8 cuando el valor de la variable contadora VM es mayor o igual al error introducido como parámetro; en este ciclo se obtiene el valor mayor de similitud almacenado en la matriz y se guardan en los vectores F1 y F2 los centros descriptores pertenecientes a la posición X e Y de la matriz, respectivamente. Este paso se repite hasta que se cumpla la condición descrita en el paso 8. Al finalizar el ciclo se obtienes los fragmentos F1 y F2, los cuales están compuestos por centros descriptores.

En el paso 9 se obtienen una matriz de similitud de distancia que almacena en cada celda el índice de similitud entre cada una de la distancia entre los centros descriptores presentes en los fragmentos F1 y F2 representado en cada fila y columna, este índice de similitud se calcula empleando el coeficiente de similitud de Tanimoto. En el paso 10 se inicializa una variable contadora VM en cero para ser utilizada posteriormente.

Del paso 11 al 16 se realizan las mismas operaciones descritas entre el paso 3 y 8, con la diferencia que los valores almacenados en la matriz son valores de similitud entre distancias y de este ciclo se obtienen los fragmentos FD1 y FD2, los cuales se retornan en el paso 17.

Tabla 3.Algoritmo buscar fragmento por PMC.

Algoritmo searchFragmentMCP (CD1, CD2, sinFunc, error, index)
Inicio
1. MCD ← createComparisonMatrix(CD1, CD2, sinFunc, error, index)
2. VM ← 0
3. Hacer
4. VM ← { mayor valor en la matriz de semejanza MCD }
5. (PX; PY)← { fila-columna de VM | (PX; PY) ∉ E}
6. F1 ← Adiciona ({ cd | centros descriptores de CD1 en la posición PX})
7. F2 ← Adiciona ({ cd | centros descriptores de CD2 en la posición PY})
8. Mientras VM ≥ error
9. MD ← createComparisonMatrix(F1, F2, sinFunc)
10. VM ← 0
11. Hacer
12. VM ← { mayor valor en la matriz de semejanza MD }
13. (PX; PY)← { fila-columna de VM | (PX; PY) ∉ E}
14. FD1 ← Adiciona ({ cd | centros descriptores de CD1 en la posición PX})
15. FD2 ← Adiciona ({ cd | centros descriptores de CD2 en la posición PY})
16. Mientras VM ≥ error
17. Retornar FD1, FD2
Fin

2.5.3.-Reducción del grafo molecular.

El algoritmo para reducción del grafo permite reducir el mismo generando cada centro de descriptor, a continuación, describiremos paso a paso el algoritmo.

En la *Tabla 4* se muestra el algoritmo mediante el cual se reduce el grafo químico, dado una molécula, para ello en paso 1 se obtienes en la lista R todos los CDs cuya clasificación es anillos, en este paso es necesario aclarar que se obtienen todos los anillos de orden 3 hasta orden n. Luego en el paso 2 se obtienen los cluster que pueden ser de orden 3 o 4 y se almacenan en la lista C. Seguidamente en el paso 3 se obtienen y se almacenan en la lista FG los grupos funcionales los cuales pueden ser metilo, metileno y metino. A continuación, en el paso 4 se obtienen los heteroátomos que no son más que los átomos pesados distintos de carbono (C) y se almacenan en la variable H, como penúltimo paso se

almacena en una variable nombrada GR todos los centros descriptores obtenidos en las variable R, C, FG y H y por último en el paso 6 se retorna dicha variable.

Tabla 4.Reducción del grafo químico.

Algoritmo reducedGraph (M1)
Inicio 1. R ← searchRings (M1) 2. C ← searchCluster (M1) 3. FG ←searchFuntionalGroup (M1) 4. H ← searchHeteroatoms (M1) 5. GR ← { centros descriptores del grafo M1 } 6. Retornar GR Fin

2.5.4.-Normalización de descriptores moleculares.

En la *Tabla 5* se muestra el algoritmo mediante el cual normalizan los descriptores, dado dos moléculas, para ello en el paso 1 se busca el mayor valor negativo de los descriptores calculados en la molécula M1 y son asignados a la variable B1. Luego en el paso 2 se realiza la misma operación, pero para la molécula M2 y el resultado se asigna a la variable B2. A continuación, en el paso 3 se realiza una comparación entre los valores de B1 y B2, en el caso que B1 sea mayor que B2 en el paso 4 se obtienen las moléculas M1 y M2 normalizadas utilizando el valor de B1; en el caso contrario se ejecuta el paso 6 normalizando las moléculas M1 y M2 con el valor de B2. En cualquiera de los casos se retorna el valor de las moléculas M1 y M2 con sus descriptores normalizados.

Tabla 5.Algoritmo de normalización de descriptores.

Algoritmo descriptorNormalization (M1, M2)
Inicio
1. B1 ← biggerNegative (M1)
2. B2 ← biggerNegative (M2)
3. si B1 > B2 entonces
4. (M1, M2) ← normalization (M1, M2, B1)
5. si no
6. (M1, M2) ← normalization (M1, M2, B2)
7. Retornar M1, M2
Fin

2.5.5.-Calculo de similitud.

En la *Tabla 6* se muestra el algoritmo mediante el cual se calcula la similitud, entre las moléculas M1 y M2, para ello en paso 1 se obtiene el valor de la suma de los descriptores ponderados por propiedades químico-físicas del fragmento de la molécula M1 obtenido en el algoritmo searchFragmentMCP descrito en el epígrafe 2.5.2 y se almacena en la variable PMCA. En el paso 2 se realiza la misma operación, pero con el fragmento perteneciente a la molécula M2 y se almacena en la variable PMCB. Luego, se calcula la propiedad máxima común entre los valores de PMCA y PMCB utilizando la fórmula de la media aritmética y se almacena en el variable PMC, esto se realiza en el paso 3 del algoritmo.

A continuación, en los pasos 4 y 5 se obtienen las variable PTA y PTB mediante la suma de los descriptores ponderados por propiedades químico-físicas de los átomos pertenecientes a las moléculas M1 y M2 respectivamente. Por último, se calcula el valor de la similitud empleando un coeficiente de similitud y se almacena en la variable CS y acto seguido se retornado en el paso 7.

Tabla 6. Cálculo de la similitud.

Algoritmo calculateSimilarity (M1, M2)
Inicio 1. PMCA ← calculatePMCA (M1) 2. PMCB ← calculatePMCB (M2) 3. PMC ← calculatePMC (PMCA, PMCB) 4. PTA ← calculatePTA (M1) 5. PTB ← calculatePTB(M2) 6. CS ← calculateCS (PTA, PTB, PMC) 7. Retornar CS Fin

2.6.- Lenguaje de programación: Java.

Para el desarrollo de este trabajo se seleccionó como lenguaje de programación Java el cual es muy utilizado en la actualidad. Desarrollado por la compañía Sun Microsystems, Java es un lenguaje de propósito general, concurrente, basado en clases y orientado a objetos, que fue diseñado específicamente para tener tan pocas dependencias de implementación como fuera posible con una sintaxis fácilmente accesible y cómoda de desarrollar, elaborado a partir de los lenguajes C y C++, de donde hereda sus características principales, a la vez que elimina otras para mantener reducidas las especificaciones del lenguaje, llegando a reducir a la mitad los errores más comunes de programación en estos lenguajes.

Como características generales presenta una disponibilidad de un amplio conjunto de bibliotecas, gestión avanzada de memoria, trabaja con sus datos como objetos y con interfaces a estos y soporta las tres características propias del paradigma de la programación orientada a objetos: encapsulación, herencia y polimorfismo. Posee una arquitectura neutral, es decir, su compilador compila su código a un fichero objeto de formato independiente de la arquitectura de la máquina en que se ejecutará, es portable, multihilo, multiplataforma (Windows, Linux, Mac). Además, construye sus interfaces de usuario a través de un sistema abstracto de ventanas de forma que las ventanas puedan ser implantadas en los diferentes sistemas operativos.

2.7.- Entorno de Desarrollo Integrado: Eclipse.

Para la implementación de los algoritmos y las pruebas realizadas se utilizó como entorno de desarrollo el Eclipse, plataforma de software compuesto por un conjunto de herramientas de programación de

código abierto multiplataforma, extensible, basada en Java y liberada bajo Licencia Publica Eclipse (EPL). La misma es una potente herramienta universal de entorno de desarrollo de software hecha en Java y lo usa como lenguaje de programación principal, aunque permite plugins para varios lenguajes. Eclipse fue desarrollado originalmente por IBM y actualmente es desarrollado por la Fundación Eclipse (53), organización que fomenta una comunidad de código abierto y un conjunto de productos complementarios, capacidades y servicios. Eclipse es un software multiplataforma por lo que se puede ejecutar en diversos sistemas operativos incluyendo Windows y Linux y posee la capacidad de ser soportado para distintas arquitecturas. Su misión consiste en evitar tareas repetitivas, facilitar la escritura de código correcto, disminuir el tiempo de depuración e incrementar la productividad del desarrollador. Otra de las características destacables de esta herramienta es que soporta la programación orientada a objetos (POO).

Eclipse posee un editor de código visual que ofrece compilación incremental de código, autocompletado, tabulador de un bloque de código seleccionado, resaltado de sintaxis, un potente depurador (que permite establecer puntos de interrupción, modificar e inspeccionar valores de variables), un navegador de clases, un gestor de archivos y proyectos y asistentes (wizards): para la creación, exportación e importación de proyectos, así como para generar esqueletos de códigos (templates). Por todas estas características fue seleccionado Eclipse como entorno de desarrollo para la programación de este trabajo.

2.8.- Librerías utilizadas

En la bioinformática, debido a la cantidad de datos a procesar, el tiempo computacional y la velocidad de ejecución son elementos que no se deben despreciar. Aprovechando las potencialidades de Java, otros autores han creado librerías programadas en este lenguaje para facilitar el manejo visual eficiente de estructuras químicas.

2.8.1.- Jmol.

Es un visualizador de Java de código abierto para estructuras químicas en tercera dimensión que realiza representación gráfica tridimensional de alto rendimiento sin grandes requerimientos de hardware, pues solo precisa de la instalación de la Máquina Virtual de Java. Es multiplataforma, compatible con sistemas operativos Windows, Mac OS y Linux/Unix. Se destaca por ofrecer numerosas funcionalidades nuevas en la representación y análisis de estructuras. Reconoce numerosos formatos moleculares. Ofrece funcionalidades para la representación de estructuras secundarias de biomoléculas, pudiéndose obtenerse interactivamente parámetros esenciales como distancia, ángulo y ángulo de torsión. Exporta

los resultados procesados a .jpg, .png, .ppm, .pdf y PovRay. Puede ser utilizado como librería para incluirlo en otras aplicaciones. (54)

2.8.2.- Chemistry Development Kit (CDK).

Se utiliza el Chemistry Development Kit (CDK) (55) pues esta es una librería de código abierto programada en Java para la química computacional y la quimio y bioinformática, disponible en Windows, Unix y Mac OS. Es desarrollada por más de 40 programadores alrededor del mundo y usado en más de 10 proyectos académicos e industriales diferentes de todo el mundo. En los últimos años, la biblioteca de CDK ha evolucionado hasta convertirse en un potente paquete de quimioinformática completo. Entre sus bondades se puede destacar la capacidad de generar y editar diagramas de estructuras en dos dimensiones, así como generación de geometría en tres dimensiones, búsqueda de subestructuras y cálculo de descriptores para QSAR.

2.9.- Dataset utilizados.

El conjunto de datos de trabajo es el ensayo AID-941 pertenecientes a la base de datos PubChem BioAssay (56), perteneciente al Centro Nacional de Información Biotecnológica (NCBI) (57).

AID941 (58) es un ensayo de cribado de alto rendimiento confirmatorio basado en células para inhibidores de la unión de TLR4-MyD88 y pertenece al proyecto de ensayo "Resumen de los esfuerzos de desarrollo de la sonda para identificar inhibidores de Toll-Like Receptor 4 (TLR4); el mismo está formado por un conjunto de 330 molécula sintetizadas de las cuales 176 resultaron activas y 173 inactivas. Los ficheros de datos estructurales de los compuestos presentes en el ensayo, se encuentran en el formato mol o sdf.

2.10.- Aplicación desarrollada.

A continuación, veremos imágenes de la aplicación desarrollada donde podemos observar su funcionamiento, en la Figura 7 muestra dos moléculas a la cual se le busco su similitud, las partes de color amarillo se observan los fragmentos similares de cada molécula.

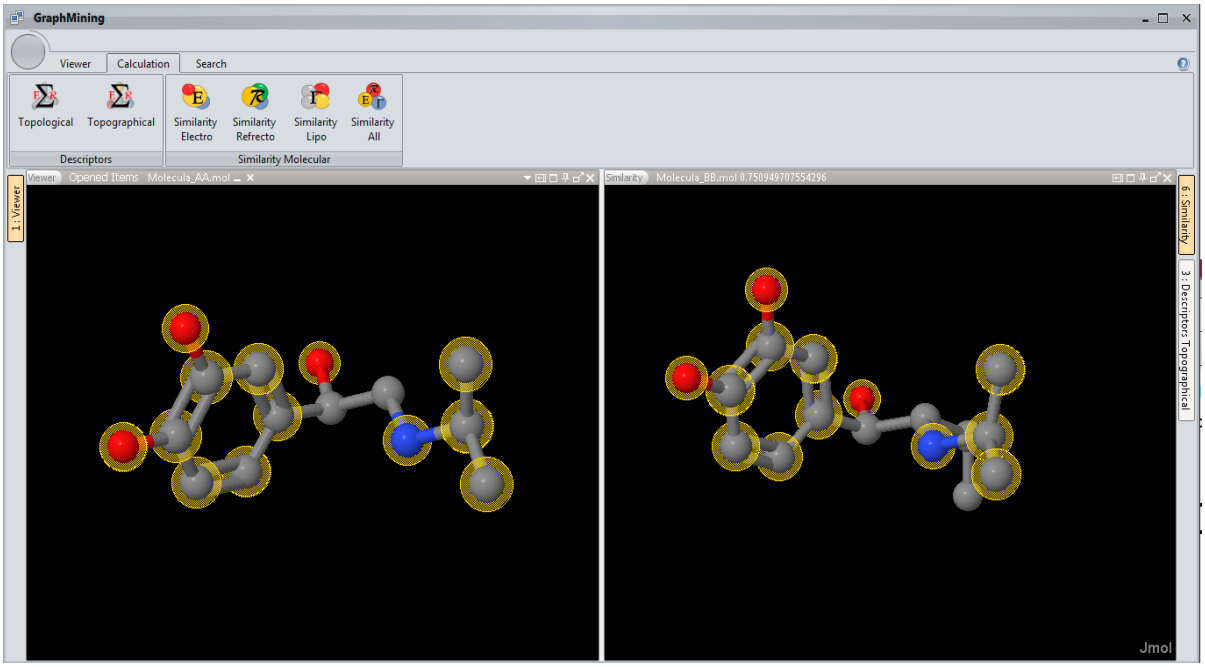


Figura 7. Representación del cálculo de la similitud molecular.

En la Figura 8, se representa la visualización de una molécula utilizando el índice del estado electrotopográfico. Para realizar la visualización primeramente se deben calcular los descriptores y acto seguido en el menú de vistas se debe seleccionar con un clic la opción identificada en rojo.

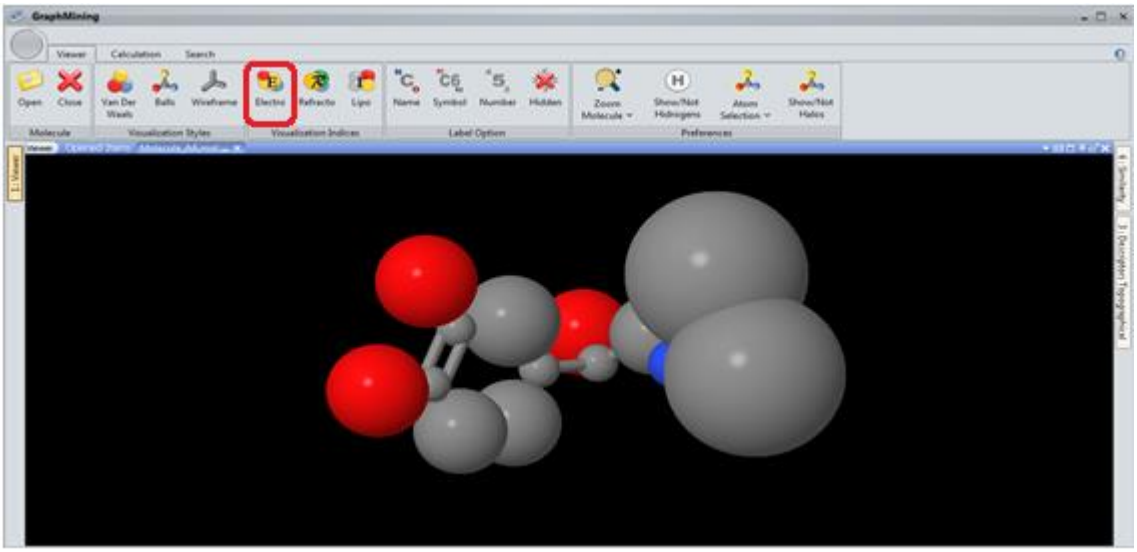


Figura 8. Representación del índice de estado Electrotopográfico.

En la *Figura 9*, se representa una molécula visualizada empleando el índice estado r*efractotopográfico*. Para realizar la visualización se utilizan los mismos pasos descritos anteriormente con la diferencia que se selecciona la opción que se muestra en la figura.

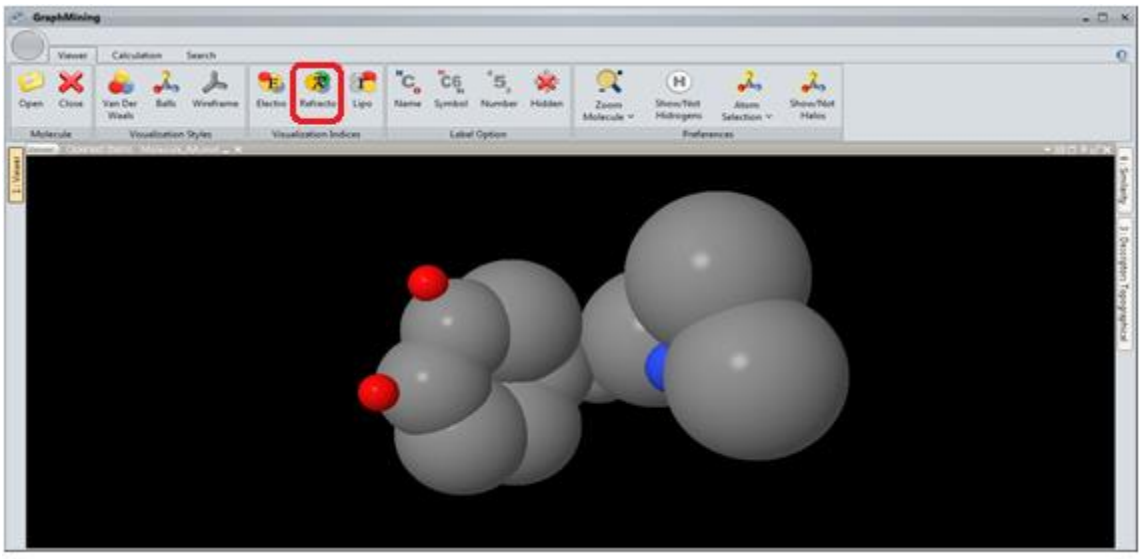


Figura 9.Representación del índice de estado refracto Refractotopográfico.

En la *Figura 10*, se representa una molécula visualizada empleando el índice estado lipotopográfico. Para realizar la visualización se utilizan los mismos pasos descritos anteriormente con la diferencia que se selecciona la opción que se muestra en la figura.

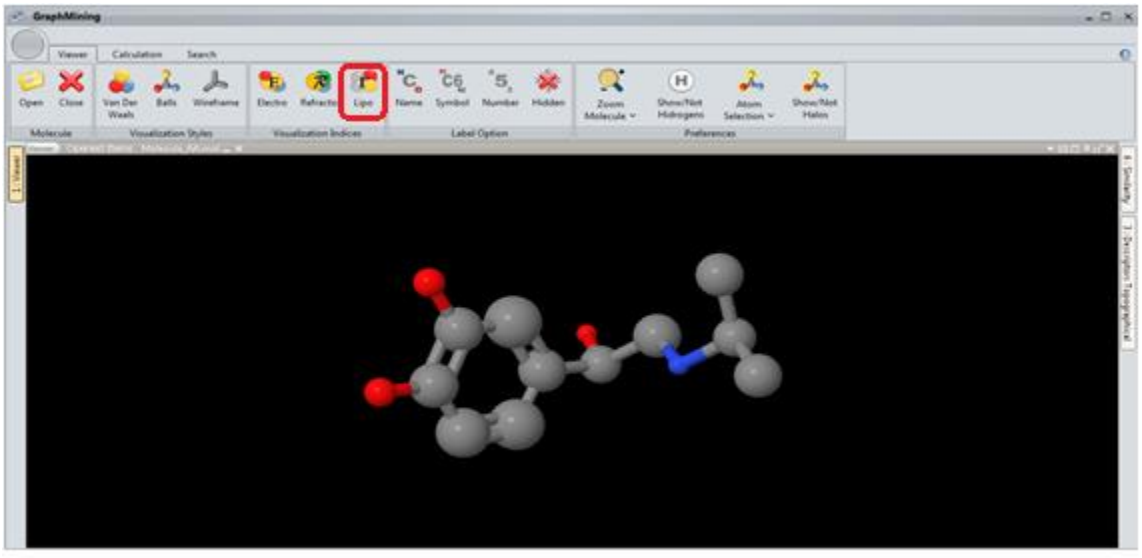


Figura 10.Representación del índice de estado Lipotopográfico .

En la *Figura 11*, se muestra el resultado obtenido al calcular la similitud molecular, en este caso las moléculas se visualizan utilizando uno de los índices calculados y además se muestra el valor de la similitud obtenido entre los dos compuestos químicos. Lo mismo ocurre en las *Figura 12* y *Figura 13* con la diferencia que se visualizan los compuestos utilizando los índices refractotopográfico y lipotopográfico respectivamente.

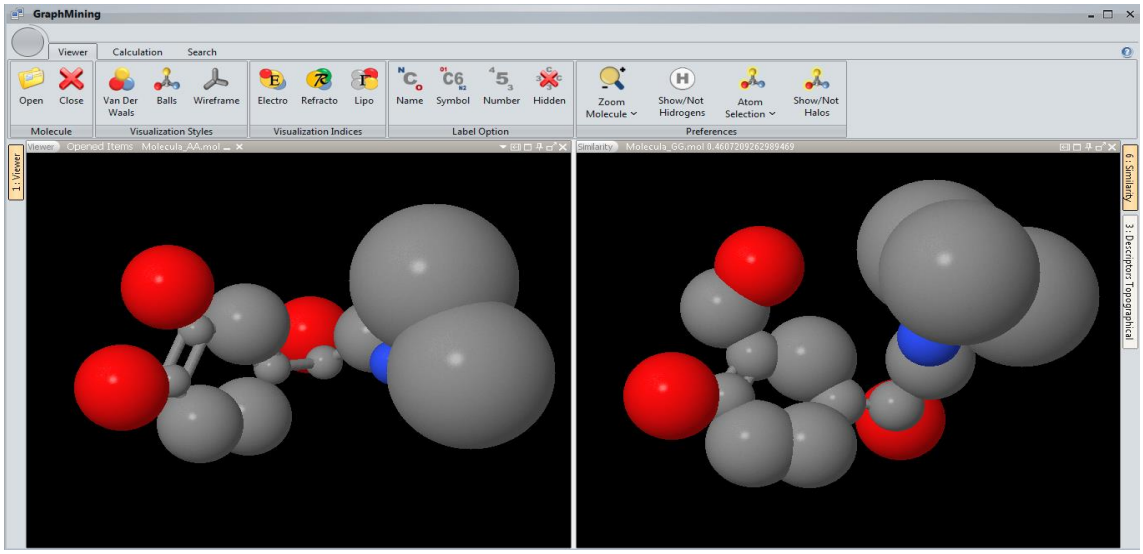


Figura 11. Representación de la similitud molecular, visualizado por el índice de estado Electrotopográfico.

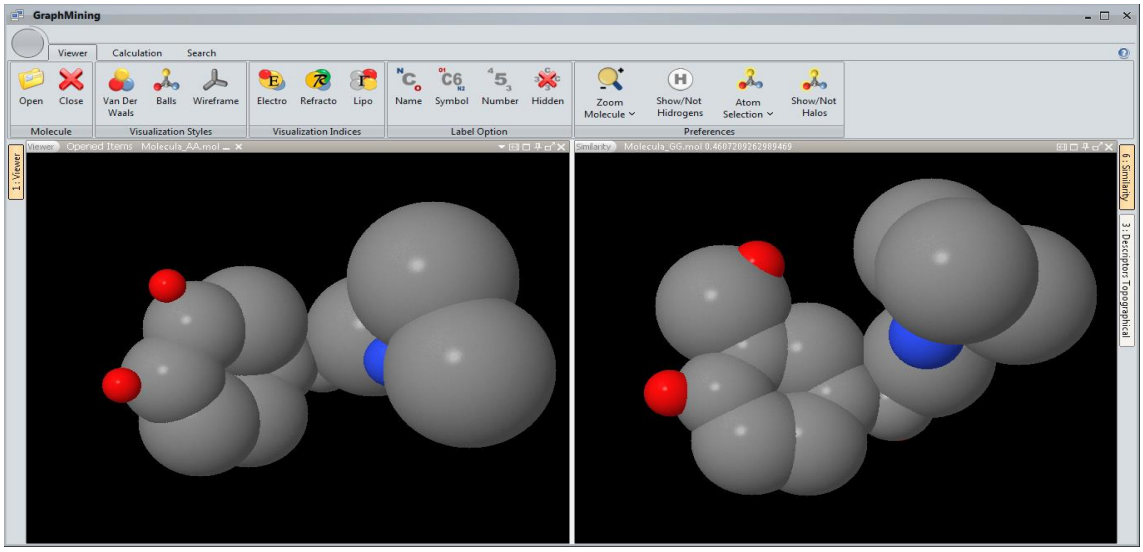


Figura 12. Representación de la similitud molecular, visualizado por el índice de estado Refractotopográfico.

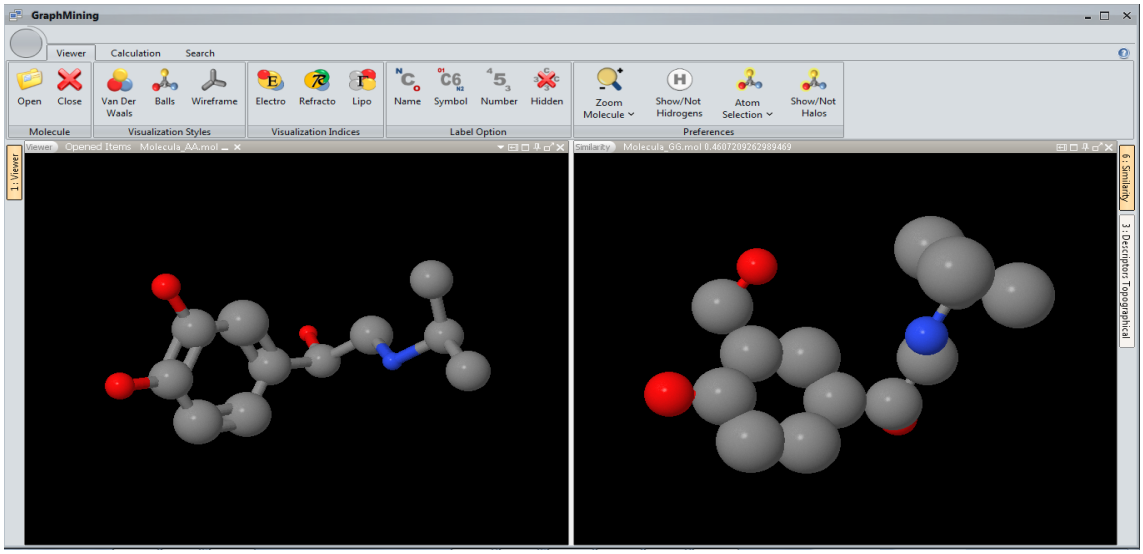


Figura 13. Representación de la similitud molecular, visualizado por el índice de estado Lipotopográfico.

2.11 Conclusiones del capítulo.

En este capítulo se explicó el procedimiento de fragmentación de un grafo químico donde se seleccionó la forma de reducción del grafo químico basado en Centros Descriptores (CD), además se decidió utilizar para el cálculo de las propiedades químico-físicas los descriptores híbridos atómicos (S_{3D}, $\Re_{3D}$ y Λ_{3D}). Se seleccionaron para obtener el valor de similitud molecular los coeficientes de Tanimoto y Dice. Estas funciones fueron modificadas por el concepto de Propiedad Máxima Común (MCP), el cual constituye uno de los aportes fundamentales de esta investigación. Se presentaron además los algoritmos implementados para el cálculo de similitud molecular por Propiedad Máxima Común(PMC), la búsqueda de fragmentos similares por PMC, la reducción del grafo molecular, normalización de descriptores moleculares y por último el algoritmo cálculo de similitud. Finalmente se presentaron las características principales que motivaron a la selección del lenguaje de programación y el entorno de desarrollo, así como las librerías a utilizar en la implementación de los algoritmos de búsqueda, y los dataset utilizados.

CAPITULO 3. RESULTADOS Y DISCUSIÓN

En este capítulo se exponen los resultados obtenidos en la investigación y los diferentes experimentos realizados para poder refrendar los métodos de búsquedas desarrollados. Se presentan los resultados alcanzados mediante el empleo de los descriptores topológicos e híbridos ponderados por propiedades químico-físicas, validando la utilización de estos, independientemente de las diferencias estructurales que se detecten. Finalmente se exhiben los resultados de aplicar los métodos de búsquedas implementados en la exploración de similitud entre compuestos mediante el empleo del concepto de Propiedad Máxima Común.

3.1.- Selección del coeficiente de similitud basado en la Propiedad Máxima Común.

Para la selección del coeficiente de similitud basado en el concepto de propiedad máxima común de los descrito en el epígrafe 2.4, se empleó como juego de datos 7 compuestos moleculares utilizados por Bajorath en su trabajo "Diseño de redes espaciales químicas utilizando una variante de similitud Tanimoto basada en las subestructuras máximas comunes" publicado en el 2015 (59). En la *Tabla 7*, se muestra el resultado de los cálculos de similitud de los dos coeficientes de similitud Tanimoto y Dice basado en MCPhd utilizando los descriptores híbridos S_{3D}, $\Re_{3D}$ y Λ_{3D}. Como se puede apreciar los dos coeficientes arrojaron resultados aceptables para cada uno de los descriptores de forma individual o con la combinación de todos.

Tabla 7. Resultados del cálculo de similitud utilizando los coeficientes de similitud Tanimoto y Dice basados en MCPhd.

	Coeficientes de similitud							
	Tanimoto				**Dice**			
Moléculas	**S_{3D}**	**$\Re_{3D}$**	**Λ_{3D}**	**$S_{3D}, \Re_{3D}, \Lambda_{3D}$**	**S_{3D}**	**$\Re_{3D}$**	**Λ_{3D}**	**$S_{3D}, \Re_{3D}, \Lambda_{3D}$**
A-B	0,71	0,67	0,89	0,69	0,54	0,53	0,62	0,53
A-C	0,84	0,94	0,90	0,90	0,61	0,65	0,63	0,63
A-D	0,58	0,67	0,74	0,61	0,50	0,55	0,58	0,52
A-E	0,47	0,53	0,40	0,51	0,42	0,46	0,38	0,44
A-F	0,28	0,65	0,74	0,65	0,30	0,55	0,59	0,55
A-G	0,45	0,50	0,13	0,48	0,40	0,43	0,15	0,42

Pero si se realiza una análisis más detallado de los resultados obtenidos al emplear cada uno de los índices, se puede observar que los resultados de similitud obtenidos al utilizar el descriptor S_{3D} son superiores al emplear el coeficiente Tanimoto como se muestra en la *Figura 14*. Lo mismo ocurre con los índices $\Re_{3D}$ y Λ_{3D} como se puede observar en las *Figura 15* y*Figura 16*, respectivamente.

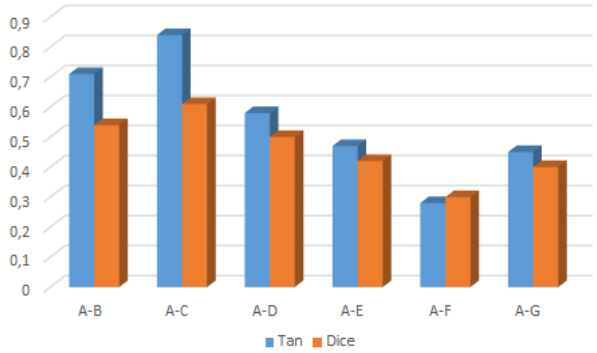


Figura 14. Comparación los valores de similitud obtenidos con los coeficientes Tanimoto y Dice basado en MCPhd utilizando el índice S_{3D}.

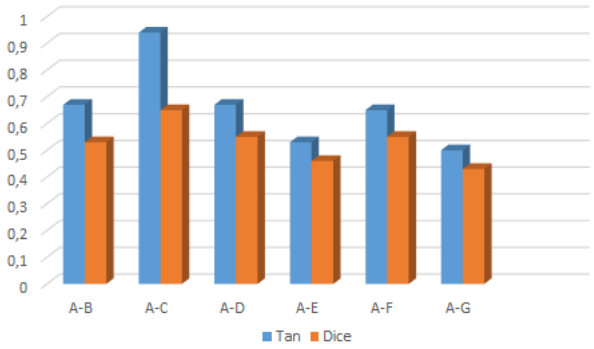


Figura 15. . Comparación los valores de similitud obtenidos con los coeficientes Tanimoto y Dice basado en MCPhd utilizando el índice $\Re_{3D}$.

Al combinar los valores de los tres descriptores antes descritos los valores de similitud obtenidos siguen desmotrando que coeficiente Tanimoto es el que mejores valores obtiene como se puede apreciar en la *Figura 17*.

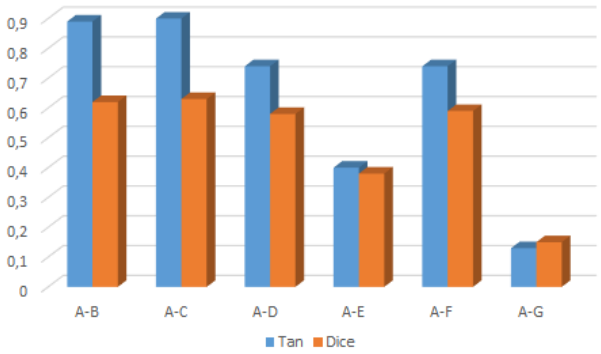


Figura 16. Comparación los valores de similitud obtenidos con los coeficientes Tanimoto y Dice basado en MCPhd utilizando el índice Λ_{3D}.

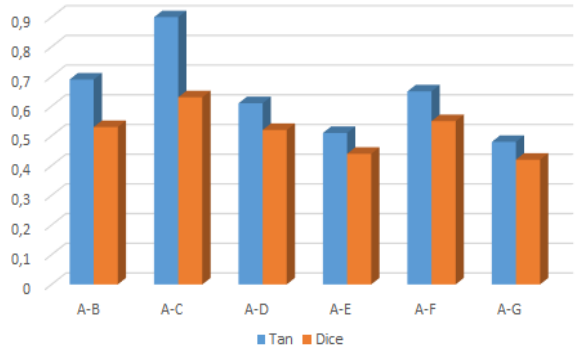


Figura 17. Comparación los valores de similitud obtenidos con los coeficientes Tanimoto y Dice basado en MCPhd utilizando los índices S_{3D}, $\Re_{3D}$ y Λ_{3D}.

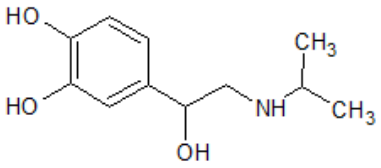
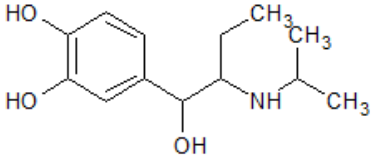
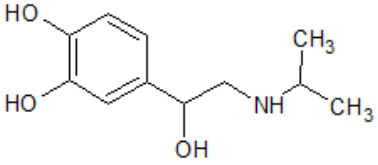
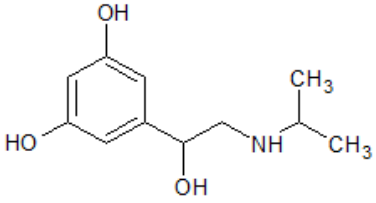
3.2.- Comparación del coeficiente de Tanimoto utilizando diferentes algoritmos.

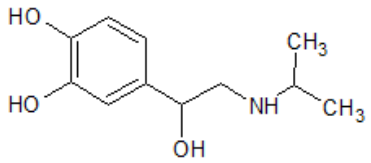
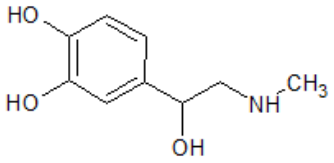
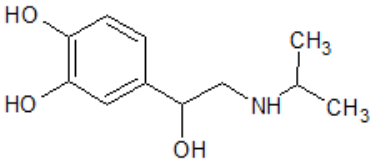
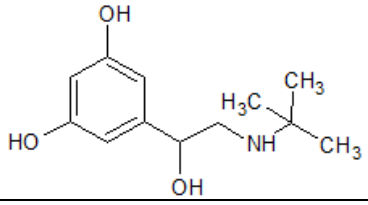
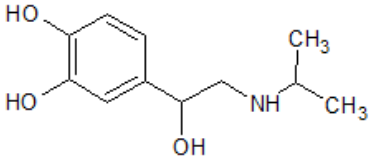
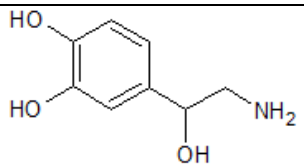
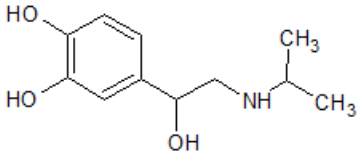
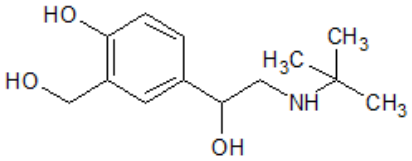
En este epígrafe se realizará una comparación del coeficiente de similitud Tanimoto basado en Propiedad máxima común (MCPhd), que fue el seleccionado en el epígrafe anterior con dos algoritmos de las más empleadas en la búsqueda de similitud en el descubrimiento de fármacos: Huella digital (Fingerprint) de conectividad extendida diámetro 4 (ECFP4) y Subestructura máxima común (MCS). Esta comparación se realizó con el objetivo de validar la efectividad del algoritmo MCPhd desarrollado en la presente investigación.

Antes de comenzar con la comparación es necesario explicar en que consisten los algoritmos ECFP4 y MCS. Las ECFP4 son huellas dactilares topológicas circulares de diametro 4 diseñadas para la caracterización molecular, la búsqueda de similitud y el modelado de estructura y actividad. Se encuentran entre los algoritmos de búsqueda de similitud más populares en el descubrimiento de fármacos y se utilizan con eficacia en una amplia variedad de aplicaciones(60). Mientras que, MCS es un algoritmo para búsqueda de similitud química y predicciones de actividad, representa la subestructura más grande que aparece al comparar dos estructuras. Su uso para medir la similitud de las estructuras químicas tiene varias ventajas. En primer lugar, es intuitivo, ya que es probable que la subestructura común más grande de fármacos estructuralmente relacionados sea un componente importante de sus actividades y en segundo lugar, la coincidencia se puede visualizar resaltando el máximo común subgráfico entre dos estructuras químicas. (61)

En la *Tabla 8*, se muestran los resultados de los calculos de similitud molecular mediente el coeficiente de Tanimoto con los tres algortimos, empleando el juego de dato de los 7 compuestos moleculares utilizados por Bajorath referenciados en el epigrafe anterior. De los mismos se puede apreciar que algortimo MCPhd para cada uno de los descriptores híbridos (S_{3D}, $\Re_{3D}$ y Λ_{3D}) y la combinación de ellos, obtienen valores de similitud superiores al algoritmo ECFP4. Lo que significa que el algoritmo propuesto en la presente investigación es más eficiente que uno de los algoritmos de búsqueda más populares en el descubrimiento de fármacos ECFP4.

Tabla 8. Resultado del cálculo de similitud molecular utilizando los métodos ECFP4, MCS y MCPhd.

	**B**	
Tc_{ECFP4} 0,51	Tc_{MCS} 0,88	Tc_{MCPhd} (S_{3D}) 0,71 ($\Re_{3D}$) 0,67 (Λ_{3D}) 0,89 (All) 0,69
	**C**	
Tc_{ECFP4} 0,61	Tc_{MCS} 0,88	Tc_{MCPhd} (S_{3D}) 0,84 ($\Re_{3D}$) 0,94 (Λ_{3D}) 0,90 (All) 0,90

	**D**	
Tc_{ECFP4} 0,51	Tc_{MCS} 0,87	Tc_{MCPhd} (S_{3D}) 0,58 ($\Re_{3D}$) 0,67 (Λ_{3D}) 0,74 (All) 0,61
	**E**	
Tc_{ECFP4} 0,35	Tc_{MCS} 0,82	Tc_{MCPhd} (S_{3D}) 0,47 ($\Re_{3D}$) 0,53 (Λ_{3D}) 0,40 (All) 0,51
	**F**	
Tc_{ECFP4} 0,55	Tc_{MCS} 0,80	Tc_{MCPhd} (S_{3D}) 0,28 ($\Re_{3D}$) 0,65 (Λ_{3D}) 0,74 (All) 0,65
	**G**	
Tc_{ECFP4} 0,45	Tc_{MCS} 0,78	Tc_{MCPhd} (S_{3D}) 0,45 ($\Re_{3D}$) 0,50 (Λ_{3D}) 0,13 (All) 0,48

Por otra parte, con respecto al algoritmo MCS los resultados obtenidos con el algoritmo propuestos utilizando los diferentes descriptores se encuentran por debajo en algunos índices y en otros por encima, por lo que no se puede definir a simple vista cuál de los dos algoritmos muestra mejores resultados de similitud.

Si se analiza la visualización del índice S_{3D} que se muestra en la *Figura 18*, podemos apreciar que los compuestos que más se asemejan son el A-C y luego el A-B con una similitud de 0,84 y 0,71 respectivamente, mientras que por el algoritmo MCS los compuestos A-B y A-C presentan una similitud de 0,88 en ambos casos, pero se puede apreciar que estructuralmente A-B y A-C son diferentes, por lo que el algoritmo MCPhd lograr una mayor diferenciación. Además con ambos algoritmos los compuestos E, F y G son los menos similares al compuesto A, pero con la diferencia que por el algoritmo MCS se obtienes similitudes por encima de 0,70, mientras que por el MCPhd utilizando el descriptor S_{3D} los valores de similitud se encuentran por debajo de 0,50. Si analizamos, la visualización de sus estructuras

podemos ver que son los tres compuestos que más se diferencian estructuralmente con respecto al A y también a los compuestos B, C y D, pero si analizamos los valores de similitud obtenidos por el algoritmo MCS se puede apreciar que la diferencia de similitud entre los compuestos E, F y G con respecto a B, C y D no son tan distantes, por lo que consideramos que el algoritmo MCPhd utilizando el índice S_{3D} obtiene valores de similitud más lógicos, por lo que demuestra que el índice S_{3D} logra diferenciar estructuras moleculares y por lo tanto se puede emplear para realizar búsqueda de similitud.

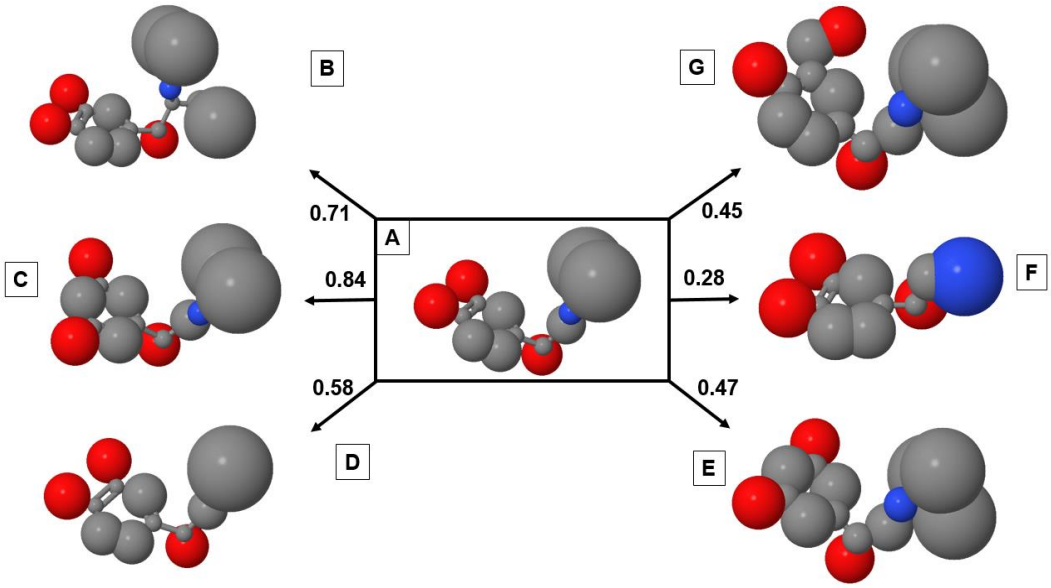


Figura 18. Comparación de las relaciones de similitud. Para un compuesto A y seis moléculas relacionadas B-G, las relaciones de similitud se comparan sobre la base de los valores TcMCPhd utilizando el índice S_{3D}.

Además, si analizamos los resultados obtenidos por el algoritmo MCPhd utilizando el índice $\Re_{3D}$ con respecto a los obtenidos con el algortimo MCS, se puede ver que existe semejansa con el análisis realizado con los valores obtenidos al aplicar el algortimo MCPhd con el indice S_{3D} para los compuestos B y C, así como con los compuestos E, F y G. Por lo que se puede afirmar que el algoritmo MCPhd lograr una mayor diferenciación en los valores de similitud obtenidos, lo cual se puede reafirmar con la visualización del índice $\Re_{3D}$ mostrada en la *Figura 19*.

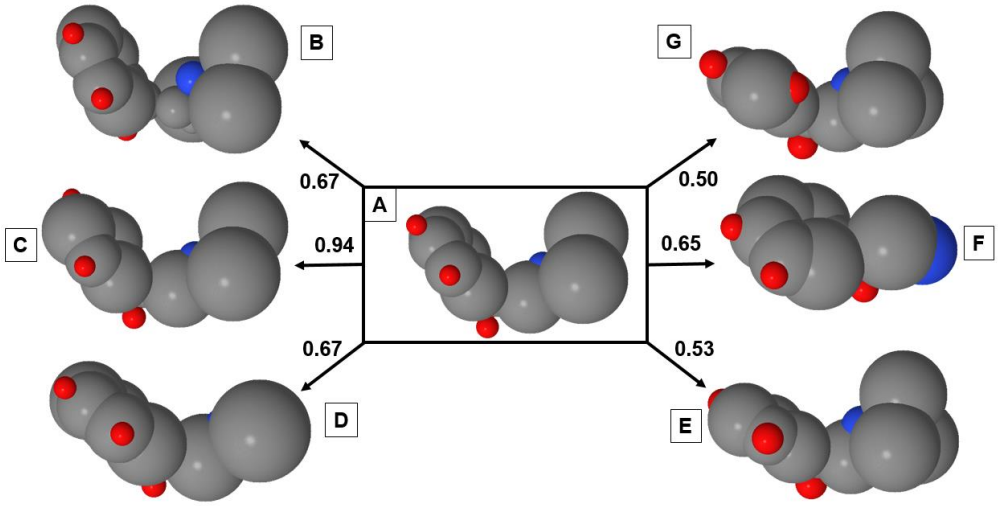


Figura 19. Comparación de las relaciones de similitud. Para un compuesto A y seis moléculas relacionadas B-G, las relaciones de similitud se comparan sobre la base de los valores TcMCPhd utilizando el índice $\Re_{3D}$.

Lo mismo ocurre con los valores obtenidos al utilizar el índice Λ_{3D}, como se muestra en la *Figura 20*. Analizando lo anteriormente expuesto se demostro que el algortimo MCPhd utilizando los índices S_{3D}, $\Re_{3D}$ y Λ_{3D} indistintamente obtiene resultados de similitud entre moleculas más aceptables que los algortimos ECFP4 y MCS, debido a que logra hacer una mayor diferenciación en los valores de similitud obtenidos, los cuales coinciden con la visualización estructural de cada índice.

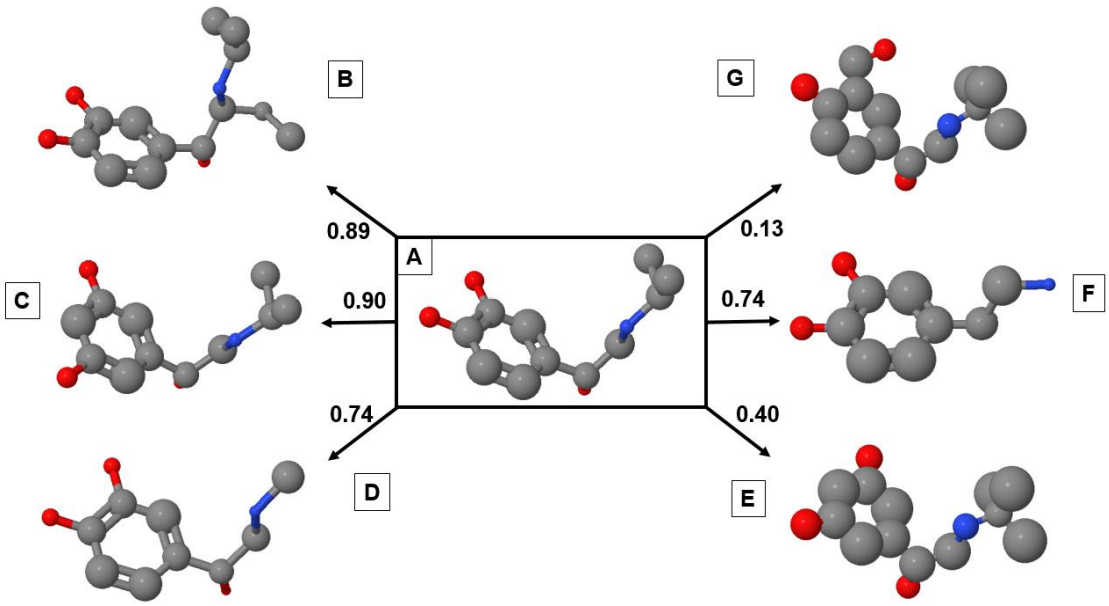


Figura 20. Comparación de las relaciones de similitud. Para un compuesto A y seis moléculas relacionadas B-G, las relaciones de similitud se comparan sobre la base de los valores TcMCPhd utilizando el índice Λ_{3D}.

3.3.- Cálculo de Similitud molecular utilizando el concepto de Propiedad Máxima Común.

Después de haber demostrado que el algoritmo MCPhd utilizando los descriptores híbridos S_{3D}, $\Re_{3D}$ y Λ_{3D} obtienes valore más aceptables de semejanza, es hora de demostrar si guarda alguna relación con la actividad biológica de los compuestos analizados. Para ellos se utilizará el ensayo ADI-941 perteneciente a la base de datos PubChem BioAssay, el cual presenta 349 compuestos moleculares divididos en 176 compuestos activos y 173 inactivos a deferentes tipos de cáncer. Para ello se realizó el cálculo de todas las moléculas perteneciente en el ensayo contra todas para cada uno de los índices y la combinación de ellos.

En la *Tabla 10*, se muestra el resultado de los cálculos de similitud de las 10 moléculas más activas del ensayo sobre la base de los valores de TcMCPhd utilizando el índice $\Re_{3D}$. De los mismo se puede concluir que el algoritmo logra encontrar un conjunto pequeño de moléculas similares, es decir, logra realizar un tamizaje adecuado para ser empleado en el descubrimiento de fármacos y además logra encontrar un porciento más elevado de molecular activas con respecto a las moléculas inactivas, lo que demuestra que el índice $\Re_{3D}$ guarda relación con la actividad biológica de las moléculas presente en el ensayo.

Tabla 9. Resultados del cálculo de similitud sobre la base de los valores TcMCPhd utilizando el índice $\Re_{3D}$ de las 10 moléculas más activas del ensayo AID941.

Moléculas	Total	% Actividad	> 50%	Active	Inactive	% Active	% Inactive
2957085	329	100	24	19	5	0,79	0,21
1218173	329	98	24	20	4	0,83	0,17
704848	329	98	19	14	5	0,74	0,26
1973785	329	97	10	5	5	0,50	0,50
2221657	329	94	21	14	7	0,67	0,33
756377	329	94	21	14	7	0,67	0,33
185146	329	93	7	7	0	1,00	0,00
15944846	329	91	15	13	2	0,87	0,13
305322	329	89	15	13	2	0,87	0,13
6031948	329	70	3	3	0	1,00	0,00

Cuando se analizan los resultados de las 10 moléculas menos activas del ensayo mostrados en la *Tabla 10*, podemos ver que se obtiene un porciento de moléculas inactivas mayor a las moléculas activas encontradas, lo que reafirma lo planteado anteriormente.

Tabla 10. . Resultados del cálculo de similitud sobre la base de los valores TcMCPhd utilizando el índice $\Re_{3D}$ de las 10 moléculas más inactivas del ensayo AID941.

Moléculas	Total	% Actividad	< 10%	Active	Inactive	% Inactive	% Active
647846	329	0	50	19	31	0,62	0,38
657803	329	0	34	12	22	0,65	0,35
661889	329	0	26	10	16	0,62	0,38
665132	329	0	85	41	44	0,52	0,48
2998049	329	0	11	5	6	0,55	0,45
2079329	329	0	133	65	68	0,51	0,49
2666619	329	0	92	42	50	0,54	0,46
741237	329	0	40	18	22	0,55	0,45
3243630	329	0	49	17	32	0,65	0,35
3244608	329	0	18	6	12	0,67	0,33

Lo mismo ocurre cuando se analizan los resultados obtenidos con los índices S3D y Λ3D, los cuales de muestran en las *Tabla 11Tabla 12*, *Tabla 13* y*Tabla 14*. Por lo que se puede afirmar que los índices analizados guardan relación con la actividad biológica presente en el ensayo y por lo tanto queda demostrado que se pueden utilizar para realizar búsqueda de similitud molecular para una actividad biológica determinada.

Tabla 11. Resultados del cálculo de similitud sobre la base de los valores TcMCPhd utilizando el índice S_{3D} de las 10 moléculas más activas del ensayo AID941.

Moléculas	Total	% Actividad	> 50%	Active	Inactive	% Active	% Inactive
2957085	329	100	17	16	1	0,94	0,06
1218173	329	98	23	16	7	0,70	0,30
704848	329	98	22	17	5	0,77	0,23
1973785	329	97	8	6	2	0,75	0,25
2221657	329	94	18	13	5	0,72	0,28
756377	329	94	12	10	2	0,83	0,17
185146	329	93	6	6	0	1,00	0,00
15944846	329	91	7	7	0	1,00	0,00
305322	329	89	11	9	2	0,82	0,18
6031948	329	70	6	4	2	0,67	0,33

Tabla 12. Resultados del cálculo de similitud sobre la base de los valores TcMCPhd utilizando el índice S_{3D} de las 10 moléculas más inactivas del ensayo AID941.

Moléculas	Total	% Actividad	< 10%	Active	Inactive	% Inactive	% Active
647846	329	0	42	19	23	0,55	0,45
657803	329	0	19	6	13	0,68	0,32
661889	329	0	27	6	21	0,78	0,22
665132	329	0	79	47	32	0,41	0,59
2998049	329	0	7	2	5	0,71	0,29

2079329	329	0	71	36	35	0,49	0,51
2666619	329	0	62	37	25	0,40	0,60
741237	329	0	53	31	22	0,42	0,58
3243630	329	0	33	12	21	0,64	0,36
3244608	329	0	13	7	6	0,46	0,54

Tabla 13. Resultados del cálculo de similitud sobre la base de los valores TcMCPhd utilizando el índice Λ_{3D} de las 10 moléculas más activas del ensayo AID941.

Moléculas	Total	% Actividad	> 50%	Active	Inactive	% Active	% Inactive
2957085	329	100	29	23	6	0,79	0,21
1218173	329	98	20	13	7	0,65	0,35
704848	329	98	19	14	5	0,74	0,26
1973785	329	97	6	5	1	0,83	0,17
2221657	329	94	18	15	3	0,83	0,17
756377	329	94	16	12	4	0,75	0,25
185146	329	93	12	10	2	0,83	0,17
15944846	329	91	11	9	2	0,82	0,18
305322	329	89	17	13	4	0,76	0,24
6031948	329	70	15	4	2	0,67	0,33

Tabla 14. Resultados del cálculo de similitud sobre la base de los valores TcMCPhd utilizando el índice Λ_{3D} de las 10 moléculas más inactivas del ensayo AID941.

Moléculas	Total	% Actividad	< 10%	Active	Inactive	% Inactive	% Active
647846	329	0	52	19	33	0,63	0,37
657803	329	0	31	11	20	0,65	0,35
661889	329	0	25	8	17	0,68	0,32
665132	329	0	59	30	29	0,49	0,51
2998049	329	0	14	6	8	0,57	0,43
2079329	329	0	49	18	31	0,63	0,37
2666619	329	0	92	47	45	0,49	0,51
741237	329	0	59	34	25	0,42	0,58
3243630	329	0	46	18	28	0,61	0,39
3244608	329	0	10	4	6	0,60	0,40

Por último se analizaron los cálculos realizados con la combinación de los tres índices los cuales se muestran en la *Tabla 15* y*Tabla 16*. De los mismos se puede concluir que al combinar los tres índices se obtienen valores semejantes a los arrojados al utilizar los índices por separados, por lo que se puede concluir que los tres índices se relaciones entre sí, demostrándose lo planteado por Brown y Fraser en 1868 sobre el diseño de nuevo fármacos, que la bioactividad de una droga depende de las propiedades físico-químicas: electrónicas, estéricas y lipofílicas.

Tabla 15. Resultados del cálculo de similitud sobre la base de los valores TcMCPhd utilizando los índices S_{3D}, $\Re_{3D}$ y Λ_{3D} de las 10 moléculas más activas del ensayo AID941.

Moléculas	Total	% Actividad	> 50%	Active	Inactive	% Active	% Inactive
2957085	329	100	18	15	3	0,83	0,17
1218173	329	98	23	18	5	0,78	0,22
704848	329	98	15	9	6	0,60	0,40
1973785	329	97	12	10	2	0,83	0,17
2221657	329	94	18	13	5	0,72	0,28
756377	329	94	13	11	2	0,85	0,15
185146	329	93	6	6	0	1,00	0,00
15944846	329	91	8	8	0	1,00	0,00
305322	329	89	15	13	2	0,87	0,13
6031948	329	70	5	3	2	0,60	0,40

Tabla 16. Resultados del cálculo de similitud sobre la base de los valores MCPhd utilizando los índices S_{3D}, $\Re_{3D}$ y Λ_{3D} de las 10 moléculas más inactivas del ensayo AID941.

Moléculas	Total	% Actividad	< 10%	Active	Inactive	% Inactive	% Active
647846	329	0	50	18	32	0,64	0,36
657803	329	0	31	8	23	0,74	0,26
661889	329	0	22	5	17	0,77	0,23
665132	329	0	70	33	37	0,53	0,47
2998049	329	0	16	2	14	0,88	0,13
2079329	329	0	100	50	50	0,50	0,50
2666619	329	0	77	36	41	0,53	0,47
741237	329	0	34	14	20	0,59	0,41
3243630	329	0	41	17	24	0,59	0,41
3244608	329	0	17	9	8	0,47	0,53

3.4.- Conclusiones del capítulo.

En este capítulo se explicaron los cálculos de los coeficientes de similitud basado en la Propiedad Máxima Común mediante un juego de datos de siete moléculas, para la validación de la función de similitud Tanimoto se escoge por obtener mejores resultados que Dice. Además, se hizo una comparación de los coeficientes de similitud de Tanimoto utilizando diferentes algoritmos basados en Subestructura Máxima Común, dicho cálculo reafirma la validación del algoritmo propuesto ya que se obtienen una mayor diferenciación en los valores de similitud obtenidos. Por último, se hizo el cálculo de similitud molecular utilizando el concepto de Propiedad Máxima Común en el ensayo ADI-941 para cada uno de los índices, por lo que se puede afirmar que los índices analizados guardan relación con la actividad biológica presente en el ensayo y por lo tanto queda demostrado que se pueden utilizar para realizar búsqueda de similitud molecular para una actividad biológica determinada.

CONCLUSIONES GENERALES.

- Se diseñó e implemento el algoritmo MCPhd basado en el concepto de propiedad máxima común para encontrar moléculas similares en una colección de grafos, utilizando los índices híbridos: Refractotopográfico (S_{3D}), Electrotopográfico ($\mathcal{R}_{3D}$) y Lipotopográfico (Λ_{3D}).
- Se validó los resultados del algoritmo implementado MCPhd con los algoritmos más utilizados en el diseño de fármacos EFCP4 y MCS, demostrándose la efectividad del algoritmo MCPhd.
- Se demostró con la utilización del algoritmo MCPhd en el ensayo AID-941, que los índices híbridos: Refractotopográfico (S_{3D}), Electrotopográfico ($\mathcal{R}_{3D}$) y Lipotopográfico (Λ_{3D}) guardan relación con la actividad biológica presente en el ensayo.

RECOMENDACIONES.

- Aplicar el algoritmo MCPhd a otros ensayos moleculares.
- Trabajar en la optimización del algoritmo MCPhd, para ser aplicado eficientemente a colecciones grandes de compuesto.

REFERENCIAS BIBLIOGRAFÍCAS

1. Vergara Mardones, Hernán. Fármacos, salud y vida: las armas y metas de la farmacia. -- [Santiago: s.n., 2011] (Salesianos impresores). -- 202 p
2. UK Department of Health (UKDH). Medicines and older people: implementing medicines-related aspects of the NSF for older people. 2001.
3. US National Institute of Allergy and Infectious Diseases (USNIAID). The problem of antimicrobial resistance (2006). Acceso: sept. 2012. http://www.niaid.nih.gov/ factsheets/antimicro.htm.
4. GRABOWSKI.H. "The effect of pharmacoeconomics on company research and development decisions ". Pharmacoeconomics (vol 5): 389-397. 1997.
5. Galduf J, Gil A. Revisión crítica del estudio de P. Dazon: precios y disponibilidad de medicamentos en nueve países. Revista Economía de la Salud. 2006; 5(1): 22-30pp.
6. Escalona, Julio. C., Carrasco, Ramón y Padrón, Juan A. Introducción al diseño racional de fármacos. La Habana: Editorial Universitaria, 2008. pág. 4. ISBN 978-959-16-0647-1.
7. Kubinyi, H. Similarity and dissimilarity – a medicinal chemist's view. Perspect Drug Discov. Perspectives in Drug Discovery and Design. s.l.: Kluwer Academy Publishers, 1998.
8. 4. Yan, Xifeng y Han, Jiawei. gSpan Graph-Based Substructure Pattern Minning. Department of Computer Science, University of Illinois. Urbana-Champaign: s.n., 2002. Technical Report.
9. The Gaston Tool for Frequent Subgraph Mining. Proceedings of the International Workshop on Graph-Based Tools (GraBaTs 2004). Nijssen, Siegfried y Kok, Joost N. . 2005. Vol. 127, págs. 77-87.
10. Gago Alonso, Andrés, y otros. Minería de subgrafos conexos frecuentes reduciendo el número de candidatos. CENATAV. La Habana: s.n., 2008. pág. 50, RT-017.
11. Rathore, Àli. Data Mining of Chemical Compounds Using Functional Groups. s.l., India: Chabot College, 2009.
12. X. Yan and J. Han. Closegraph: mining closed frequent graph patterns. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '03, pages 286–295. ACM, 2003.
13. J. Huan,W.Wang, J. Prins, and J. Yang. Spin: mining maximal frequent subgraphs from graph databases. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '04, pages 581–586. ACM, 2004.
14. L. T. Thomas, S. R. Valluri, and K. Karlapalem. Margin: Maximal frequent subgraph mining. ACM Trans. Knowl. Discov. Data, 4(3):10:1–10:42, 2010.

15. C. Chen, X. Yan, F. Zhu, and J. Han. gapprox: Mining frequent approximate patterns from a massive network. In ICDM, pages 445–450. IEEE Computer Society, 2007.
16. Y. Jia, J. Zhang, and J. Huan. An efficient graph-mining method for complicated and noisy data with real-world applications. Knowl. Inf. Syst., 28(2):423–447, 2011.
17. M. E. Saeedy and P. Kalnis. GraMi: generalized frequent pattern mining in a single large graph. Technical report, Division of Mathematical and Computer Sciences and Engineering, King Abdullah University of Science and Technology, 2011.
18. Flores-Garrido, M., Carrasco-Ochoa, JA. & Martínez-Trinidad, J.F. Knowl Inf Syst (2015) 44: 385.
19. J.J. McGregor, "Backtrack Search Algorithms and the Maximal Common Subgraph Problem", Software Practice and Experience, Vol. 12, pp. 23-34, 1982.
20. Asad Rahman S, Bashton M, Holliday G, Schrader R, Thornton J. Small Molecule Subgraph Detector (SMSD) Toolkit. J. Cheminform. 2009.
21. R. Hariharan, A. Janakiraman, R. Nilakantan, B. Singh, S. Varghese, G. A. Landrum and A. Schuffenhauer: MultiMCS: A Fast Algorithm for the Maximum Common Substructure Problem on Multiple Molecules. Journal of Chemical Information and Modeling 51(4): 788-806, 2011.
22. Dalke and Hastings: FMCS: a novel algorithm for the multiple MCS problem. Journal of Cheminformatics 2013.
23. R.Carrasco and J. A. Padrón: Definición de un índice atómico novedoso para QSAR: el estado refractotopológico. Canadian Society for Pharmaceutical Sciences. 2004.
24. Marrero, Yovani, y otros. TOMOCOMD-CARDD: Un Novedoso Enfoque para el Diseño 'Racional In-Silico' de Fármacos Antimaláricos. Santa Clara: Editorial de la Universidad Marta Abreu, 2006. pág. 27.
25. A. Sirageldin, A. Selamat, R. Ibrahim, Graph–based simulated annealing and support vector machine in malware detection, in: M. F. Harun, A. Selamat (Eds.), 5th Malaysian Conference in Software Engineering (MySEC), IEEE Comp. Soc., Johor Bahru, 2011, pp. 512–515.
26. E. Duesbury, J. D. Holliday, P. Willett, Maximum common substructure–based data fusion in similarity searching, J. Chem. Inf. Model. 55 (2015) 222–230.
27. Y. Cao, T. Jiang, T. Girke, A maximum common substructure–based algorithm for searching and predicting drug–like compounds, Bioinformatics 24 (2008) i366–i374.
28. Ivanciuc, Ovidiu. Representing Two Dimensional (2D) Chemical Structures with Molecular Graphs. Handbook of Chemoinformatics Algorithms. Boca Raton, Fl : Chapman y Hall/CRC Taylor y Francis, 2010, págs. 1-36.

29. Aplicación de un algoritmo de reducción de grafos al Método de los Grafos Dicromáticos. Rodríguez Puente, Rafael, Marrero Osorio, Sergio A. y Lazo Cortés, Manuel S. 2, La Habana: s.n., Mayo-Agosto de 2012, Vol. 15, págs. 158-168. ISSN 1815-5944.
30. Modelos de predicción de actividad citotóxica en células SK-N-SH mediante técnicas de softcomputing en una muestra heterogénea de compuestos. Prieto-Entenza, Julio Omar, PupoMerino, Mario y Carrasco-Velar, Ramón. 3, La Habana : CENIC, 2011, Revista CENIC Ciencias Biológicas, Vol. 42, págs. 111 -118.
31. Maggiora, G; Vogt, M; Stumpfe, D; Bajorath, J. (2013) Molecular Similarity in Medicinal Chemistry. J. Med. Chem. 57:3186–3204.
32. Kunimoto, R; Vogt, M; Bajorath, J. (2016) Maximum common substructure-based Tversky index: an asymmetric hybrid similarity measure. J Comput Aided Mol Des. 30:523–531.
33. Raymond, J. W.; Willett, P. Maximum Common Subgraph Isomorphism Algorithms for the Matching of Chemical Structures. J. Comput.-Aided Mol. Des. 2002, 16, 521−533.
34. MACCS Structural Keys; Accelrys: San Diego, CA.
35. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. J. Chem. Inf. Model. 2010, 50, 742−754.
36. Good, A. C.; Richards, W. G. Explicit Calculation of 3D Molecular Similarity. Perspect. Drug Discovery Des. 1998, 9−11, 321− 338.
37. Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein−Protein Interaction. J. Med. Chem. 2005, 48, 1489−1495.
38. Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand−Receptor Binding. J. Chem. Inf. Model. 1997, 37, 1−9.
39. McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. J. Chem. Inf. Model. 2007, 47, 1504−1519.
40. Fliri, A.; Loging, W.; Thadeio, P. F; Volkmann, R. Biological Spectra Analysis: Linking Biological Activity Profiles to Molecular Structure. Proc. Natl. Acad. Sci. U.S.A. 2005, 102, 261−266.
41. Petrone, P. M.; Simms, B.; Nigsch, F.; Lounkine, E.; Kuthukian, P.; Cornett, A.; Deng, Z.; Davies, J. W.; Jenkins, J. L.; Glick, M. Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity. ACS Chem. Biol. 2012, 7, 1399−1409.
42. Hu, Y.; Bajorath, J. Compound Promiscuity: What Can We Learn from Current Data? Drug Discovery Today 2013, 18, 644−650.
43. Vogt M, Stumpfe D, Geppert H, Bajorath J (2010) Scaffold hopping using two-dimensional fingerprints: true potential, black magic, or a hopeless endeavor? Guidelines for virtual screening. J. Med. Chem. 12:5707–5715.

44. Gardiner EJ, Holliday JD, O'Dowd C, Willett P (2011) Effectiveness of 2D fingerprints for scaffold hoping. Future Med. Chem. 3:405–414.
45. Costales Leiva, Lien y Guirola González, Asnay. Tutores: R: Carrasco-Velar y A. Antelo-Collado. Predicción de actividad anticancerígena de compuestos orgánicos partiendo de descriptores, utilizando programación genética. La Habana: UCI, 2007.
46. Todeschini , R y Consonni , V. Handbook of Molecular Descriptors. 2000.
47. Todeschini, R., y otros. Molecular Descriptors for Chemoinformatics. [ed.] H. Kubinyi, G. Folkers R. Mannhold. Germany : Wiley-VCH, 2009.
48. al., R. Todeschini et y DRAGON. TALETE s.r.l. [En línea] 2013. [Citado el: 9 de abril de 2015.] http://www.talete.mi.it/..
49. Carrasco-Velar, Ramón. Nuevos descriptores atómicos y moleculares para estudios de estructura-actividad. Aplicaciones. La Habana: Editorial Universitaria, 2007. 978-959-16-0646-4.
50. Hybrid reduced graph for SAR studies. Carrasco Velar, R., y otros. s.l: Taylor y Francis, 2013, SAR and QSAR in Environmental Research.
51. Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. Sung Hyuk, Cha. 4, 2007, International Journal of Mathematical Models and Methods in Applied Sciences, Vol. 1, págs. 300-307.
52. Asad Rahman, Syed, y otros. Small Molecule Subgraph Detector (SMSD) toolkit. s.l. : Journal of Cheminformatics, 2009. págs. 2-3.
53. The Eclipse Foundation. Eclipse. [En línea] The Eclipse Foundation, 2004. [Citado el: 7 de Noviembre de 2014.] https://eclipse.org/.
54. JMol. JMOL. [En línea] 2015. [Citado el: 9 de Noviembre de 2014.] http://jmol.sourceforge.net/.
55. The Chemistry Development Kit (CDK): an open-source Java library for Chemo-and Bioinformatics. Steinbeck, Christoph, y otros. 43, s.l. : American Chemical Society, 2 de Noviembre de 2003, Journal of Chemical Information and Computer Science, págs. 493-500.
56. PubChem BioAssay Database http://www.ncbi.nlm.nih.gov/pcassay.
57. The National Center for Biotechnology Information https://www.ncbi.nlm.nih.gov/.
58. Ensayo AID941 https://pubchem.ncbi.nlm.nih.gov/bioassay/941#section=Top.
59. Desing of chemical space networks using a Tanimoto similarity variant based upon máximum common substructures. Zhang B, Vogt M, Maggiora GM, Bajorath J. 2015. J. Comput Aided Mol. Des. Vol 29, págs. 937-950.
60. 1. Extended-Connectivity Fingerprints. Rogers, D. y Hahn, M. American Chemical Society., 2010, J. Chem. Inf. Model. 2010, vol 50, págs. 742–754.

61. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. Cao Y, Jiang T, Girke T. Bioinformatics. 2008, vol 24, págs. 366–374.

GLOSARIO DE TÉRMINOS.

Actividad Biológica: Capacidad inherente de una sustancia, tal como un fármaco o una toxina, para alterar una o más funciones químicas o fisiológicas de una célula.

Algoritmo: Es una lista que, dado un estado inicial y una entrada, propone pasos sucesivos para arribar a un estado final obteniendo una solución.

Bioinformática: El uso de las matemáticas aplicadas, la estadística y la ciencia de la informática para estudiar sistemas biológicos.

Átomo: Partícula más pequeña de un elemento químico que retiene las propiedades asociadas con ese elemento.

Centro de masa: En un sistema discreto o continuo es el punto geométrico que dinámicamente se comporta como si en él estuviera aplicada la resultante de las fuerzas externas al sistema. De manera análoga, se puede decir que el sistema formado por toda la masa concentrada en el centro de masas es un sistema equivalente al original.

Fármacos: Término farmacológico para cualquier compuesto biológicamente activo, capaz de modificar el metabolismo de las células sobre las que hace efecto.

Descriptor: Número que describe la estructura química o una propiedad de la molécula o fragmento de ésta.

Grafo: Conjunto de objetos llamados vértices o nodos unidos por enlaces llamados aristas o arcos, que permiten representar relaciones entre elementos de un conjunto.

Grafo molecular: Representación pictórica de la topología molecular.

Molécula: Es la partícula de una sustancia que retiene todas las propiedades de la misma y está compuesta por uno o más átomos.

Índice topográfico: Número que se calcula generalmente a partir de la matriz de adyacencia o de distancias entre los elementos de un grafo que han sido ponderados por un valor numérico que contiene información tridimensional del grafo molecular.

Propiedad Máxima Común: Dados los grafos G_1 y G_2, se entiende por fragmentos con Propiedad Máxima Común MCP_{HD} (G_1, G_2), a los subgrafos g_1 y g_1 de los grafos G_1 y G_2 que presentan la máxima

similitud en las propiedades químico-físicas representadas por los índices (S_{3D}, A_{3D}, R_{3D}), entre los Centros Descriptores (CDn) y la distancia euclidiana entre sus centros de masa ($d_E(CD_1, CD_2)$).

Grafo ponderado: Pesado o con costos es un grafo donde cada arista tiene asociado un valor o etiqueta, para representar el costo, peso, longitud.

Isomorfismo: Dado los grafos G1 y G2, se busca si existe una función biyectiva tal que los vértices u y v en los grafos G1, son adyacentes si solo si f(u) y f(v) son adyacentes en el grafo G2.