



Universidad de las Ciencias Informáticas

Facultad 2

**Búsqueda de fragmentos similares pertenecientes a grafos
moleculares utilizando índices topográficos e híbridos**

**Trabajo de Diploma para Optar por el Título de
Ingeniero en Ciencias Informáticas**

Autores:

María Cecilia Hernández Govea

Juan Luis Paneque Pérez

Tutores:

MSc. Aurelio Antelo Collado

Dr. Ramón Carrasco Velar

La Habana, junio de 2015

“Año 57 de la Revolución”

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 _____.

María Cecilia Hernández Govea

Autor

Juan Luis Paneque Pérez

Autor

Dr. Ramón Carrasco Velar

Tutor

MSc. Aurelio Antelo Collado

Tutor

Resumen

En el presente trabajo se proponen métodos de búsqueda de subgrafos en colecciones de grafos moleculares mediante técnicas de cotejo inexacto de grafos utilizando descriptores ponderados por propiedades químico-físicas. Se realiza la fragmentación del grafo químico utilizando una forma de grafo reducido que emplea agrupaciones de átomos que llamamos centros descriptores y la combinación de ellos utilizando teoría combinatoria. Para la descripción de las estructuras moleculares se emplearon los índices de Estado Refractotopográfico, Electrotopográfico y Lipotopográfico para Átomos. Se definió el concepto de Propiedad Máxima Común como criterio de identificación de moléculas con fragmentos que poseyeran valores similares de la propiedad descrita por los índices. Se implementan los algoritmos de búsquedas de fragmentos con dos o más centros descriptores, validándose los mismos con diferentes funciones de similitud obteniéndose resultados que demuestran la calidad de la búsqueda realizada. 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, cotejo inexacto de grafos.

Índice

Introducción.....	1
Capítulo 1: Fundamentación teórica.....	6
1.1 Diseño y obtención de fármacos.....	6
1.2 Grafo químico.....	7
1.3 Reducción del grafo químico.....	8
1.4 Búsqueda de similitud molecular.....	10
1.5 Medidas de similitud y de distancia.....	11
1.6 Descriptores moleculares.....	12
1.7 Índices topográficos para átomos.....	13
1.7.1 Índice del Estado Electrotopográfico.....	14
1.7.2 Índice de Estado Lipotopográfico.....	14
1.7.3 Índice de Estado Refractotopográfico.....	15
1.8 Cotejo de grafos.....	15
1.9 Técnicas de cotejo inexacto de grafos.....	16
1.10 Conclusiones del capítulo.....	17
Capítulo 2: Materiales y métodos.....	18
2.1 Reducción del grafo químico.....	18
2.2 Fragmentación del grafo químico.....	19
2.3 Descriptores utilizados.....	20
2.4 Funciones de similitud utilizadas.....	20
2.5 Propiedad Máxima Común.....	20
2.6 Algoritmos implementados.....	21
2.6.1 Búsqueda de fragmentos similares de segundo grado por índices topográficos e híbridos.....	21

2.6.2	Búsqueda de moléculas similares por fragmento de Propiedad Máxima Común.....	23
2.7	Lenguaje de programación: Java	26
2.8	Entorno de Desarrollo Integrado: Eclipse	26
2.9	Librerías utilizadas	27
2.9.1	Jmol	27
2.9.2	Chemistry Development Kit (CDK).....	27
2.10	Conclusiones del capítulo.....	28
Capítulo 3: Resultados y discusión.....		29
3.1	Cálculo de los umbrales de similitud a utilizar.	29
3.2	Utilización de la fragmentación y las propiedades químico-físicas.....	33
3.3	Búsquedas de fragmentos simples.....	36
3.4	Búsquedas de fragmentos de Propiedad Máxima Común.....	38
3.5	Conclusiones del capítulo.....	41
Conclusiones generales		42
Recomendaciones.....		43
Referencias Bibliográficas		44
Glosario de términos		48

Índice de figuras

Figura 1. Primera forma de reducción del grafo químico.	8
Figura 2. Segunda forma de reducción del grafo químico.....	9
Figura 3. Tercera forma de reducción del grafo químico.....	9
Figura 4. Cuarta forma de reducción del grafo químico.	10
Figura 5. Agrupaciones de átomos utilizadas.	18
Figura 6. Quinta forma de reducción del grafo químico.	19
Figura 7. Valores del Índice del Estado Refractotopográfico de Átomos en fragmentos.	34
Figura 8. Valores del Índice del Estado Refractopográfico para Átomos en fragmentos moleculares con un átomo de oxígeno (en rojo).	34
Figura 9. Fragmentos moleculares similares estructuralmente.	34
Figura 10. Resultados obtenidos por cada una de las funciones de similitud.	38
Figura 11. Molécula 100335 perteneciente al ensayo AID941.	39

Índice de tablas

Tabla 1. Funciones de similitud y de distancia.....	20
Tabla 2. Algoritmo para la búsqueda de fragmentos similares en moléculas.....	23
Tabla 3. Algoritmo para obtener los fragmentos PMC entre dos moléculas.....	24
Tabla 4. Algoritmo para buscar los fragmento PMC entre dos moléculas.....	25
Tabla 5. Ajuste del umbral para el coeficiente de Tanimoto.....	30
Tabla 6. Ajuste de la cota para el índice de Jaccard.....	31
Tabla 7. Ajuste de la cota para el coeficiente de Dice-Sørensen.....	31
Tabla 8. Ajuste del umbral para el coeficiente de Sørensen.....	31
Tabla 9. Ajuste de la cota para el coeficiente de Soergel.....	32
Tabla 10. Ajuste del umbral para el coeficiente de Ruzicka.....	32
Tabla 11. Ajuste del umbral para el coeficiente de Czekanowski.....	32
Tabla 12. Umbrales calculados para las funciones de similitud/distancia.....	33
Tabla 13. Fragmentos moleculares y su tipo genérico.....	35
Tabla 14. Fragmentos moleculares y sus propiedades químico-físicas.....	35
Tabla 15. Búsquedas de fragmentos en diferentes posiciones.....	36
Tabla 16. Descripción del fragmento A6-HCl1 perteneciente a la molécula 12005721.....	37
Tabla 17. Fragmentos comunes para cada una de las funciones de similitud.....	38
Tabla 18. Resultados de la búsqueda realizada.....	41

Introducción

Desde su surgimiento el hombre sufre de enfermedades que en ocasiones le causan la muerte, lo que ha conllevado a buscar incesantemente sustancias que lo curen, o al menos, disminuyan los síntomas de sus padecimientos. Tradicionalmente, las medicinas se obtenían a partir de hierbas y otras sustancias naturales, de donde se extraían los principios activos responsables de sus efectos curativos atribuidos empíricamente. Posteriormente, no sólo se logró identificar a los compuestos responsables de la actividad farmacológica, sino que además se obtuvieron derivados químicos sintetizados que de cierto modo imitan la acción del producto natural.

En la actualidad, el mundo se ve amenazado por una gran cantidad de enfermedades, haciéndose indispensable la búsqueda de nuevos tratamientos y fármacos que las combatan. El diseño, desarrollo y fabricación de los mismos, se logra y se consolida gracias al conocimiento conjunto de diversas disciplinas, las que están integradas por el trabajo de diferentes grupos multidisciplinarios, para llevar a la sustancia que es potencialmente un principio activo, hasta la presentación farmacéutica que se conoce como medicamento, en la que la capacidad de deducción, la intuición y en muchos casos, la suerte, han jugado un papel fundamental. (1)

Hoy día la ciencia utiliza herramientas que mejoran los procedimientos de extracción y síntesis de fármacos. La química computacional desarrolla la simulación de los requerimientos químico-físicos de estos para lograr especificidad sobre receptores o enzimas que correspondan a los sitios de acción, y la química combinatoria a su vez, permite generar un gran número de derivados en corto tiempo, a diferencia de las síntesis realizadas en laboratorios convencionales.

En todo este largo proceso de prueba y error a lo largo del ciclo de desarrollo de un medicamento, tanto en la fase de descubrimiento, como en la fase de desarrollo se incurre en gastos significativos que llegan a ser multimillonarios por cada nuevo medicamento que sale al mercado. Para el desarrollo de un nuevo medicamento desde su obtención en el laboratorio hasta su uso en la terapéutica puede llegar a alcanzar un costo de 360-500 millones de USD y en un período promedio de doce a quince años, donde de se necesita ensayar sobre 10 mil compuestos, de los cuales sólo 10 pasan los ensayos y solo 1 llega a convertirse en medicamento útil en la terapéutica. Esta situación evidencia la necesidad de encontrar y desarrollar métodos rápidos, eficientes y de bajo

costo computacional para optimizar el proceso de obtención de nuevos fármacos (2) (3), principalmente en la etapa inicial de su concepción.

Un elemento a tener en cuenta es el principio de similitud molecular, que plantea que las moléculas que son similares en su estructura deben tener actividad biológica semejante (4). 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 (5). 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.

El estudio de las moléculas a partir de la representación de ellas como grafos almacenados en bases de datos, es un procedimiento actual que tiene en cuenta, que la asociación de las tecnologías informáticas con los datos biológicos y químicos, generan un gran volumen de información que con las herramientas adecuadas, impulsan el descubrimiento de conocimiento relevante no intuitivo, ni observable a primera vista. Este gran volumen de información químico-biológico-estructural almacenado en grandes bases de datos y el proceso de extracción de nuevos conocimientos dieron origen al nacimiento del término bioinformática.

La bioinformática se puede definir entonces como la disciplina científica que utiliza la tecnología de la información para almacenar, organizar, recuperar, analizar, y extraer conocimientos a partir de la información químico-biológica-estructural almacenada de genes, proteínas y compuestos químicos con la finalidad de responder preguntas complejas en los campos de la química, bioquímica y biología. (6) Por tal razón, la combinación inteligente de las ciencias biológicas, químicas y de la computación, favorecen la generación de herramientas para el avance en los estudio de la genómica, proteómica y el diseño computacional de fármacos.

Algunos descriptores moleculares tienen una base experimental que los limita en cuanto al alcance y uniformidad de los datos generados en laboratorios independientes. Un enfoque en los estudios moleculares que contribuye a reducir estas discrepancias es cuando se describe la molécula a partir de una descripción paramétrica basada en la teoría del grafo químico. Los descriptores que se obtienen son el resultado de la aplicación de un procedimiento lógico y matemático, el cual transforma la información química codificada en forma de matrices de conectividad o de distancia entre los átomos (7). Estos descriptores moleculares proveerán la información estructural asociable

a las propiedades químico-físicas de cada compuesto de un modo práctico que permite un mejor tratamiento computacional.

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.

Hemos visto que las moléculas pueden ser representadas como grafos, donde los átomos representan los vértices y las aristas representan el tipo de enlace entre los átomos. Además, los grafos reducidos pueden derivarse de los grafos moleculares y constituir otra forma de representación de las moléculas. De esta forma, un fragmento molecular se estudia como un subgrafo del grafo químico asociado a la molécula. Así es posible realizar comparaciones entre distintos fragmentos moleculares y aplicar búsquedas de patrones similares entre ellos. Desde el punto de vista del grafo químico se traduce en encontrar el conjunto de subgrafos dentro de una colección de grafos que sean semejantes entre si y que, si incluimos la actividad biológica dentro del análisis, pueden ser potencialmente responsables de dicha actividad.

Los algoritmos que abordan el problema de identificar subgrafos frecuentes (SF) en grafos difieren entre sí por: la estrategia de búsqueda que emplean (en amplitud o en profundidad), la forma en que generan candidatos a patrones (extender o combinar patrones encontrados), la naturaleza de los grafos que examinan (si es una colección de grafos o un solo grafo) y el conjunto de subgrafos que encuentran (todos o parte de ellos) (8).

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 (9), Gaston (10), gRed (11) y GraphSig (12). Estos algoritmos permiten encontrar el conjunto completo de subgrafos frecuentes en una colección de grafos, que puede ser muy grande para una colección moderada de grafos. Además, identifican la ocurrencia de un subgrafo candidato resolviendo el problema del isomorfismo de grafos (o subgrafos), ya que dos grafos son isomorfos si entre ellos solo varía la apariencia (13). El isomorfismo entre dos subgrafos representa la tarea que más recursos computacionales demanda en estos algoritmos.

Con el objetivo de tratar de encontrar un subconjunto significativo, otros autores desarrollaron los siguientes algoritmos: CloseGraph (14), SPIN (15) y MARGIN (16). Los mismos se concentran en la búsqueda de subgrafos maximales o cerrados, pero a pesar de la reducción de la cantidad de subgrafos encontrados con estos algoritmos 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 (17), APGM (18), GraMi (19) y AGraP (20). Los mismos encuentran un subconjunto significativo, es decir, una menor cantidad de subgrafos que, de alguna manera, retienen la información del conjunto completo de subgrafos.

Aunque la atención en los últimos años en la minería de subgrafos se ha desplazado, de encontrar el conjunto completo de subgrafos, a encontrar un subconjunto significativo, que retenga la información del conjunto completo de subgrafos, existen algoritmos que permiten encontrar subgrafos con diferencias estructurales entre vértices y aristas basado en el cotejo exacto de grafos.

Por otra parte, existen descriptores híbridos, llamados así por estar ponderados los vértices por propiedades químico físicas, que permiten diferenciar subgrafos por esas propiedades y no solamente por su estructura. (21) (22) Sin embargo, la comparación entre este tipo de subgrafos no puede realizarse mediante el cotejo exacto de grafos y eso conduce a un campo de actualidad que es el cotejo inexacto de grafos. (23) Estas técnicas permiten establecer comparaciones entre grafos ponderados sin que la topología de las moléculas constituya una restricción.

La detección de subgrafos o fragmentos relevantes en conjuntos de moléculas evaluadas biológicamente constituye un problema de actualidad ya que encontrar la parte de la molécula que potencialmente es la responsable de la actividad biológica se le denomina detección del farmacóforo y sigue siendo un campo de investigación abierto en el diseño racional de fármacos.

A partir de lo anteriormente descrito, surge el siguiente **problema científico** a resolver: *¿Cómo realizar búsquedas de subgrafos moleculares ponderados en colecciones de grafos?*

Para esto se tendrá como **objeto de estudio** la *búsqueda de subgrafos en grafos moleculares*, y como **campo de acción** la *búsqueda de subgrafos moleculares en colecciones de grafos*.

Para dar solución al **problema** se traza como **objetivo general**: *Desarrollar métodos de búsqueda de subgrafos en colecciones de grafos moleculares utilizando índices topográficos e híbridos*.

Para dar cumplimiento al objetivo general se realizaron las siguientes **tareas de investigación**:

1. Definición de una forma de grafo reducido ponderado por propiedades químico-físicas.
2. Selección de una medida de similitud entre grafos reducidos ponderados por propiedades químico-físicas que admita diferencias estructurales de agrupaciones de vértices y aristas.
3. Implementación de algoritmos para encontrar subgrafos similares en una colección de grafos, representados mediante fragmentos simples y complejos, utilizando la medida de similitud definida.
4. Validación de los métodos de búsqueda implementados en diferentes ensayos.

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 relacionados con el objeto de estudio.

Empíricos:

- **Consulta de las fuentes de información:** se emplea 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 los algoritmos de búsquedas propuestos.
- **Pruebas:** se utilizan para comprobar si los algoritmos de búsquedas propuestos obtienen resultados aceptables.

El aporte principal de la presente investigación, es la definición del concepto de Propiedad Máxima Común y la propuesta de los algoritmos de búsquedas de fragmentos simples y complejos en colecciones de grafos utilizando propiedades químico-físicas.

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

Capítulo 1: Fundamentación Teórica. En este capítulo se presenta un resumen de la investigación realizada sobre la búsqueda de fragmentos en estructuras químicas. Se aborda el desarrollo de estas técnicas en el diseño y obtención de fármacos. 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 de los algoritmos y métodos.

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 de las implicaciones, trascendencia y beneficios de estos resultados, y las posibles aplicaciones de los algoritmos implementados y los conceptos introducidos en la investigación.

Capítulo 1: Fundamentación teórica.

En este capítulo se ofrece una perspectiva general del estudio molecular y su aplicación en la obtención de fármacos. Se exponen los principios y bases teóricas del cotejo de grafos en la búsqueda de similitud molecular a través de los estudios estructura - actividad, así como su papel en el diseño y obtención de medicamentos. Se define el concepto de grafo químico, su aplicación en la química molecular y los diferentes tipos de reducción presentes en la bibliografía. Se exponen los conceptos asociados a los descriptores atómicos - moleculares, específicamente los índices híbridos y se definen los índices electrotopográficos, refractotopográficos y lipotopográficos. Finalmente se presenta una recopilación de las funciones de similitud / diferencia, para estudios de alineación de estructuras, presentes en la literatura.

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.

Existen dos enfoques diferentes para encontrar un nuevo compuesto: 1) diseño de nuevos compuestos para ser sintetizados en un laboratorio químico y posteriormente evaluados farmacológicamente; 2) encontrar compuestos conocidos, con otras actividades o usos, en bases de datos químicas y probar su actividad experimentalmente. Ambos enfoques son importantes y muy utilizados en la práctica farmacéutica, el primero permite el diseño de nuevas cabezas de serie con la actividad deseada, pero ellos necesitan ser primeramente sintetizados, evaluados farmacológicamente y finalmente tienen que pasar a través de rigurosos ensayos toxicológicos y farmacodinámicos. Sin embargo, los compuestos seleccionados con el segundo enfoque ya tienen métodos de síntesis bien establecidos y en muchos casos su comportamiento toxicológico y farmacodinámico es bien conocido, sobre todo para el caso de compuestos comercializados como fármacos (2).

El método tradicional de búsqueda de nuevos principios activos basado en el sistema de prueba y error a través de ensayos masivos de gran número de sustancias químicas, es ineficiente. Las vías

alternativas en las que la industria farmacéutica ha depositado su confianza, y también una gran cantidad de recursos, son la síntesis combinatoria y el diseño computacional de fármacos.

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 (2). 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 Grafo químico.

La teoría de grafos es una herramienta matemática útil para los estudios moleculares de estructura actividad y está basada en la asociación de las estructuras químicas a los grafos matemáticos, a partir de la cual se pueden definir diversos descriptores que relacionen las propiedades químico-físicas de las moléculas y la estructura tridimensional de las mismas (24). En la teoría del grafo químico se define este como un grafo conexo y no dirigido, en el que los nodos son los átomos y las aristas son los enlaces entre estos.

Una gran variedad de objetos pueden ser representados, en una forma simplificada, como grafos: grupos de átomos, moléculas, conjuntos de moléculas, polímeros, reacciones y mecanismos de reacción. La representación como un grafo de objetos químicos puede retener las principales características de la propiedad investigada y ofrecer conclusiones cualitativas o cuantitativas en concordancia con las que son ofrecidas por métodos más rigurosos. La relación entre los diferentes tipos de vértices del grafo químico está dada por las aristas del grafo, que son los enlaces químicos, conexiones entre grupos de átomos o transformaciones de grupos funcionales (25).

Muchos 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. 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). A partir de estas representaciones se pueden modelar las

estructuras químicas y de ahí surgen, basados en las invariantes de conectividad molecular, los índices topológicos y topográficos como descriptores moleculares y atómicos (22).

1.3 Reducción del grafo químico.

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. Debe tenerse en consideración, que aunque los grupos de átomos pueden estar separados entre sí, todos los elementos de ese fragmento pertenecen a la molécula original y por tanto existe un camino que enlaza al menos un átomo perteneciente a uno de los grupos y otro átomo en el segundo grupo dentro del grafo totalmente conexo al que pertenecen. Un vértice en un grafo reducido puede representar un sistema de anillos, anillos aromáticos, anillos alifáticos o grupos funcionales (27).

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 (28).

Existen diversos sistemas para transformar una molécula en un grafo reducido, destacando y agrupando diferentes subestructuras en un mismo compuesto químico. A continuación se muestran las diferentes formas de reducción de un grafo químico que existen en la bibliografía, todos los tipos de grafos reducidos parten del mismo grafo molecular y se reducen a diferentes representaciones simplificadas.

Tipo 1: Los vértices en el grafo reducido corresponden a sistemas de anillos (R) y componentes acíclicos conectados (Ac). En la figura 1 se muestra, a la izquierda, el sistema de anillos R (encerrado en un círculo), correspondiente al vértice central en el grafo reducido, mostrado a la derecha de la figura:

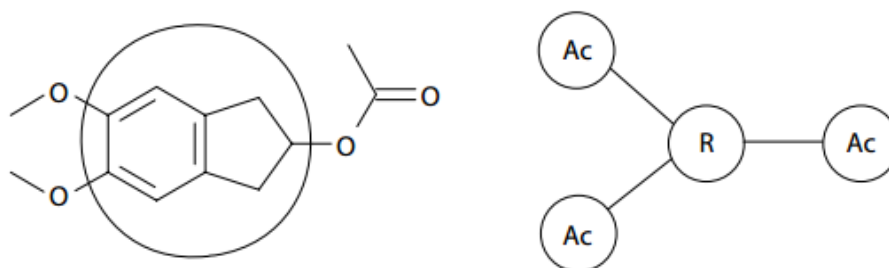


Figura 1. Primera forma de reducción del grafo químico.

Tipo 2: Los vértices en el grafo reducido corresponden a átomos de carbono conectados (C) y están enlazados a heteroátomos (H). En figura 2, en la parte superior, se muestran los heteroátomos (encerrados en una elipse), y en la parte inferior, se presenta el grafo reducido que se obtiene mediante este tipo de reducción:

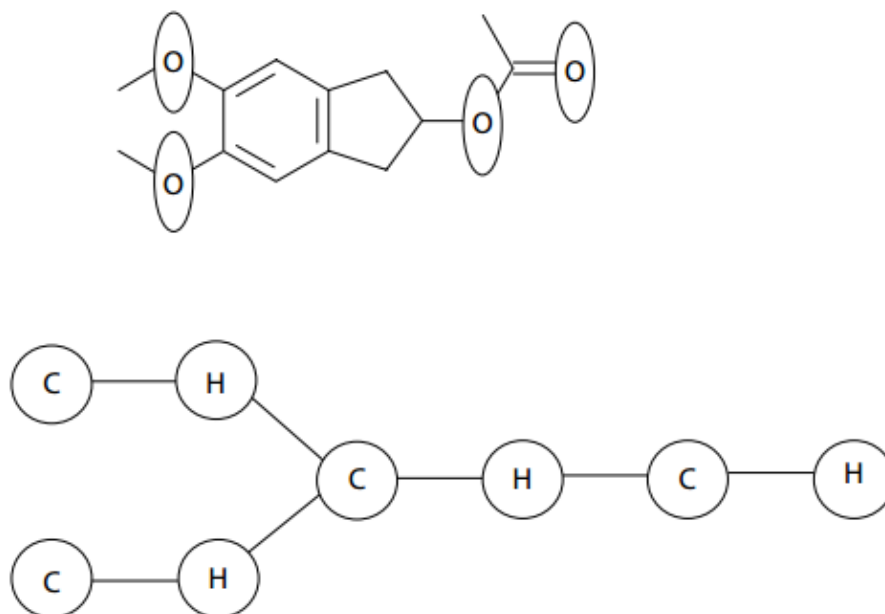


Figura 2. Segunda forma de reducción del grafo químico.

Tipo 3: Los vértices en el grafo reducido corresponden a anillos aromáticos (Ar), anillos alifáticos (R) y grupos funcionales (F). En la figura 3 se muestran, a la izquierda, los grupos funcionales (encerrados en círculos), y a la derecha se muestra la representación final del grafo reducido que se obtiene con este tipo de reducción:

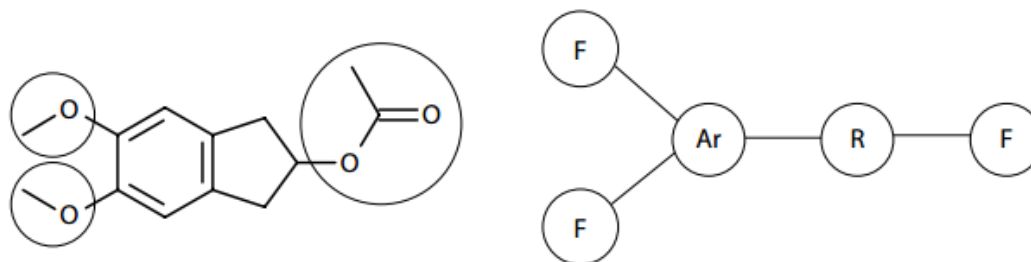


Figura 3. Tercera forma de reducción del grafo químico.

Tipo 4: Los vértices en el grafo reducido corresponden a anillos aromáticos (Ar), grupos funcionales (F) y grupos conectados (L). En la figura 4 se muestra cada grupo funcional (encerrados en un círculo) y los grupos conectados (encerrados en elipses). El grafo reducido correspondiente (mostrado a la derecha), tiene la misma topología que en la figura 3, pero con un tipo de fragmento diferente para el vértice entre los anillos aromáticos y los grupos funcionales.

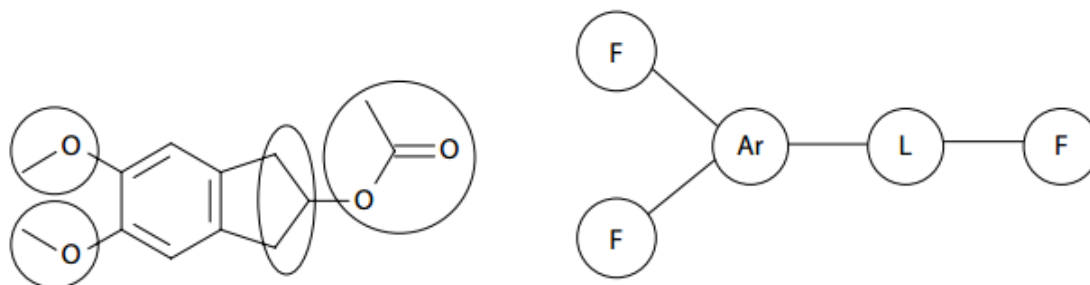


Figura 4. Cuarta forma de reducción del grafo químico.

1.4 Búsqueda de similitud molecular.

En los últimos años, los avances en el desarrollo de nuevos modelos matemáticos que describen fenómenos químicos, el desarrollo de programas más intuitivos y la disponibilidad de computadoras más veloces, han provisto a los científicos experimentalistas de herramientas computacionales que utilizadas de conjunto con técnicas de investigación tradicionales permiten, tanto examinar las propiedades estructurales de compuestos existentes, como para predecir propiedades y actividades de nuevas entidades químicas (5).

Las estructuras químicas comenzaron entonces a ser procesadas y almacenadas por los diversos servicios especializados de información pública y privada, relacionados con la industria química y químico-farmacéutica. Es así como grandes empresas farmacéuticas diseñaron sistemas de procesamiento y almacenamiento de esta información vital para su trabajo de investigación y desarrollo. Este trabajo encontró amplia aceptación entre los usuarios especializados que en muchas ocasiones prefieren hacer búsquedas mediante la estructura en sí misma, y no mediante los nombres del compuesto químico, el cual puede identificarse con diferentes nombres en dependencia de la base de datos y del sinónimo que adopte.

La búsqueda de similitud molecular es aquella técnica de recuperación de información mediante la cual, a partir de una estructura química definida por el especialista, se identifican aquellas moléculas en una base de datos que son más semejantes a la molécula de referencia usando medidas cuantitativas de similitud intermolecular. Esta técnica de recuperación de información estructural es aplicable a las bases de datos de moléculas que almacenan tanto información en 2D como 3D de las estructuras. Para esto, dos clases de medidas de similitud de estructuras químicas han sido desarrolladas, las medidas globales y las locales. Las medidas de clase global son aquellas que asignan un valor numérico para determinar la completa similitud entre dos moléculas. Las medidas del tipo local son aquellas medidas que proveen información de similitud física, como resultado de la alineación de una molécula con otra; por ejemplo, estos sistemas producen un mapeo de las características de la molécula con aquellos rasgos estructurales de las moléculas existentes en la base de datos, de forma que este proceso constituye una superposición de una molécula sobre otra.

El cálculo de similitud entre moléculas debe ser efectivo, su aplicación debe dar resultados útiles al usuario; y debe ser eficiente, lo cual se traduce en utilizar los requerimientos de computación necesarios para lograr detectar similitud entre moléculas en bases de datos de tamaño considerable en tiempos razonables. Los trabajos de búsquedas de similitud crean las bases para un mejor desarrollo de los proyectos de investigación a la hora de evaluar nuevas variantes de moléculas para ser utilizadas en sustitución de otras; creando todo un campo de gestión de información dentro de las estructuras químicas con directa implicación en los proyectos de investigación y desarrollo de fármacos.

1.5 Medidas de similitud y de distancia.

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 (29).

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.

1.6 Descriptores moleculares.

Una forma de describir el comportamiento y la estructura que caracterizan a compuestos químicos con un alto nivel de precisión, es el cálculo de descriptores moleculares. 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. (30)

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. (31)

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 (7).

1.7 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 (32). 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 (22). 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 híbrida, también está incluido en esta investigación es el Índice de Estado Electrotopográfico para Átomos (33). 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.

1.7.1 Índice del Estado Electrotopográfico.

El Índice de Estado Electrotopográfico para átomos (33) (S_{state} , S_i), se basa en el efecto electrónico de cada átomo sobre los otros átomos en la molécula. El mismo se calcula por la expresión $S_i = I_i + \Delta I_i$, donde S_i es el valor del índice para el átomo i , I_i 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 ($\sigma - h$). 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:

$$\Delta I_i = \sum (I_i - I_j) / r_{i,j}^2$$

donde $r_{i,j}^2$ es la distancia euclídeana 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 S_{state} 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.

1.7.2 Índice de Estado Lipotopográfico.

El Índice de Estado Lipotopográfico para átomos (33) ($L_{state3D}$, 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

$$\Delta AL_{ij} = \sum_{j=1}^N (AL_i + AL_j)/r_{i,j}^2$$

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 $r_{i,j}^2$ es la distancia euclidiana entre los átomos i y j , calculado a partir de la estructura optimizada con algún método semiempírico.

1.7.3 Índice de Estado Refractotopográfico.

El Índice de Estado Refractotopográfico para Átomos (33) ($R\text{-state3D}$, \mathfrak{R}_{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:

$$\mathfrak{R}_{3D} = AR_i + \Delta AR_i$$

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

$$\Delta AR_{ij} = \sum_{j=1}^N (AR_i + AR_j)/r_{i,j}^2$$

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 $r_{i,j}^2$ es la distancia euclidiana entre los átomos i y j , calculado a partir de la estructura optimizada con algún método semiempírico.

1.8 Cotejo de grafos.

En los últimos años se ha detectado la necesidad de convertir grandes volúmenes de datos en información útil. Como consecuencia se han ido desarrollando técnicas y métodos que permiten procesar este gran volumen de información. Ejemplo de tales técnicas lo constituye el descubrimiento de patrones frecuentes, en especial la detección de subgrafos frecuentes en colecciones de grafos donde evaluar la similitud estructural de los grafos es una de las tareas más costosas en tiempo de ejecución, conocida como cotejo de grafo. Existen dos formas de abordar el cotejo de grafos: cotejo exacto y cotejo inexacto. (23)

El cotejo exacto consiste en determinar si las etiquetas y la estructura de dos grafos son idénticas. Para que dos grafos sean considerados equivalentes se requiere que exista una igualdad total entre

vértices y aristas (y etiquetas en el caso de grafos etiquetados) (8). El concepto de la igualdad entre grafos se conoce como isomorfismo. Dos grafos son isomorfos si existe una correspondencia biunívoca entre V_1 y V_2 , que mantiene las aristas y todas las etiquetas. Debido a que hay que tolerar cierto nivel de distorsiones geométricas, variaciones semánticas o desajustes entre vértices o aristas mientras se realiza la búsqueda de los subgrafos frecuentes, se ha hecho necesario evaluar la similitud entre grafos admitiendo algunas diferencias estructurales, o sea, mediante técnicas de cotejo inexacto.

El cotejo inexacto consiste en encontrar la mejor adecuación entre vértices o aristas de dos grafos para determinar su similitud admitiendo diferencias entre estructuras y etiquetas. Sobre esta base, se ha planteado la necesidad de realizar la minería de SF utilizando cotejo inexacto de grafos, además, es válido mencionar que el cotejo inexacto de grafos ha sido aplicado con éxito en dominios de la ciencia como el análisis de estructuras bioquímicas (23).

Estudiar los grafos químicos utilizando conceptos de búsquedas de patrones en grafos analizados como estructuras de datos, permitiría conocer una molécula y realizar comparaciones entre ellas, atendiendo a los descriptores moleculares con propiedades químico-físicas particionadas como contenido de información adicional.

1.9 Técnicas de cotejo inexacto de grafos.

Existen aplicaciones donde el cotejo exacto de grafos no es aplicable con éxito, debido a la existencia de distorsiones geométricas, variaciones semánticas o desajustes entre vértices y aristas. Por lo antes planteado se ha hecho necesario emplear técnicas de cotejo inexacto entre grafos permitiendo diferencias en la estructura. (23)

Entre las técnicas conocidas se encuentran la Distancia de Edición de Grafos (DEG) que se le conoce como el conjunto de operaciones de edición estándar sobre un grafo. La β arista isomorfismo, mediante la cual se calcula la similitud entre dos grafos considerando la ausencia o presencia de aristas entre un par de grafos, sin modificar los vértices. Derivada de esta está la (β arista sub-isomorfismo) que en la práctica permite la identificación de subgrafos comunes entre grafos (23)

Una técnica utilizada para calcular la similitud entre dos grafos es la Distancia de Edición de Grafos (DEG); esta técnica se le conoce como el conjunto de operaciones de edición estándar sobre un grafo y el proceso a seguir sería re-etiquetar o sustituir, e insertar un vértice o una arista del grafo seleccionado. Además existen también otras operaciones de edición como la fusión o división de

vértices o aristas donde mientras más pequeña sea la DEG resultante el costo de edición será menor, siendo muy similares los grafos en término de estructura y etiquetas. Otras definiciones como el homeomorfismo con vértices/aristas disjuntas y las basadas en las probabilidades de sustitución, pueden encontrarse más amplias y con el rigor y formulismo matemático necesario, en la referencia citada. (23)

1.10 Conclusiones del capítulo.

En este capítulo se presentaron las tendencias actuales en el diseño y obtención de fármacos y el papel que desempeña la bioinformática en el estudio molecular y en la búsqueda de sustancias con principios activos equivalentes, que combinadas permitan la elaboración y comercialización de medicamentos más efectivos, menos costosos y con una disminución de los efectos colaterales. Se definieron los principales conceptos asociados a los estudios QSAR, el cotejo de grafos y la búsqueda de similitud molecular, áreas en las 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 y las funciones de similitud (distancia) para el cálculo de semejanzas entre estructuras reportadas.

Capítulo 2: Materiales y métodos

En este capítulo se expone el método de reducción del grafo molecular utilizado, el procedimiento de fragmentación del grafo, los descriptores y funciones de similitud empleados y los algoritmos implementados. Igualmente se explican los procedimientos de búsqueda de fragmentos similares y las herramientas utilizadas en la investigación, definiéndose el lenguaje de programación y la plataforma con la que se codificó el prototipo implementado, así como las librerías utilizadas en los estudios. Al final del capítulo se presentan los algoritmos implementados para dar solución al problema planteado.

2.1 Reducción del grafo químico.

Como se mencionó en el Capítulo 1, la reducción del grafo químico, consiste en obtener una representación más abstracta y reducida de la estructura molecular que a la vez mantenga las características y propiedades principales. A partir de estudios realizados (34), se propone un método de reducción del grafo químico, basado en agrupar los átomos en estructuras típicas de la fragmentación grafo-teórica (clúster, ciclos, vértices y grupos funcionales seleccionados). En esta forma se emplea la reducción del grafo químico en la que los vértices en el grafo reducido corresponden agrupaciones de átomos, topológicamente iguales (35). Estos corresponden a heteroátomos (H), formados por átomos de oxígeno, nitrógeno, azufre, bromo y cloro; anillos de diferente tamaño (A); clúster de orden 3 y 4 (C); metilo (M3); metileno (M2) y metino (M), los cuales se muestran en la figura 5.

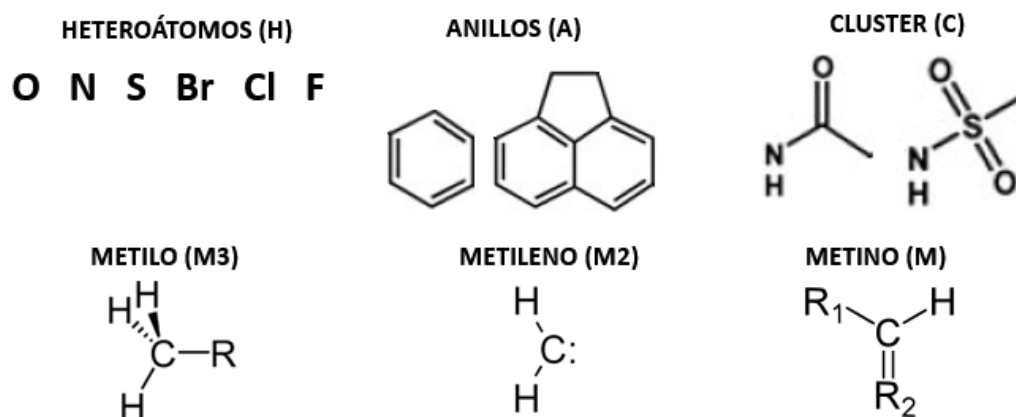


Figura 5. Agrupaciones de átomos utilizadas.

En la figura 6 se muestra, en la parte izquierda, los nuevos vértices del grafo reducido, identificados (encerrados en círculos), y a la derecha se muestra el grafo reducido que se obtiene mediante esta técnica de reducción.

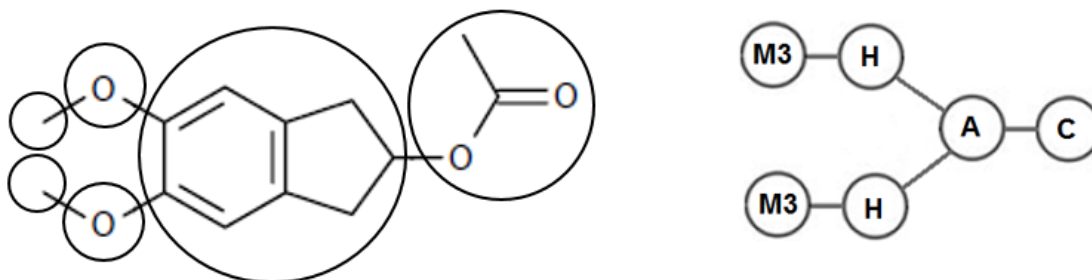


Figura 6. Quinta forma de reducción del grafo químico.

2.2 Fragmentación del grafo químico.

Para estudiar mejor una molécula o macromolécula, a menudo es preferible fragmentarla en piezas menores y estudiar cada una de ellas por separado. Esto simplifica el trabajo a costa de perder quizás alguna información sobre las interacciones entre las partes de la molécula, sin embargo, a pesar de ello esta técnica cada día alcanza mayor popularidad.

A partir del total de fragmentos reducidos obtenidos mediante la forma de reducción del grafo químico ya mencionada, fragmentos a los cuales denominaremos centros descriptores (CD), se define un fragmento molecular de segundo orden a la combinación de un par de CD relacionados entre sí por la distancia euclidiana entre sus respectivos centros de masas (CM). El número total de fragmentos de segundo orden en una molécula fragmentada por este procedimiento, se obtiene entonces como la combinación de todos los pares de CD en los que se divide la molécula, a los que se les asigna la distancia euclidiana entre los CM de cada par.

La diferencia principal entre los diferentes CM estará dada por la ponderación que se haga de los átomos que los componen. Como criterio de ponderación de los átomos se utilizarán los índices de Estado mencionados en el epígrafe 1.7. De esta forma, un fragmento de 2º orden puede definirse como el vector:

$$v = (S_{i_{CD1}}, \mathfrak{R}_{3D_{CD1}}, \Lambda_{3D_{CD1}}, S_{i_{CD2}}, \mathfrak{R}_{3D_{CD2}}, \Lambda_{3D_{CD2}}, d_E(CD_1, CD_2)) \text{ donde } v \in \mathbb{R}^7.$$

lo cual facilita el estudio de las similitudes entre estos. Es posible definir fragmentos de orden superior, es decir, relaciones entre 3 o más CDs. En el procedimiento de búsqueda, se tendrá en cuenta la distancia euclidiana entre los CM de los CDs.

2.3 Descriptores utilizados.

Para el desarrollo del presente trabajo se emplearon como descriptores atómicos topográficos los Índices de Estado Electrotopográfico, Refractotopográfico y Lipotopográfico reportados recientemente. (33)

2.4 Funciones de similitud utilizadas.

A partir del estudio de las diferentes funciones de similitud / diferencia presentes en la bibliografía, se seleccionaron aquellas que mejor se adaptan al cálculo de similitud molecular, basados en vectores de valores. En la tabla 1 se presentan los coeficientes de similitud o distancia encontrados en la bibliografía (29), y que serán empleados en esta investigación.

No	Nombre	Fórmula	Intervalo
1	Sørensen	$\frac{\sum a_i - b_i }{\sum (a_i + b_i)}$	[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	$\frac{\sum a_i - b_i }{\sum \max(a_i, b_i)}$	[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]

Tabla 1. Funciones de similitud y de distancia.

2.5 Propiedad Máxima Común.

Antes de definir el concepto de Propiedad Máxima Común (PMC) 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 $(v_i, v_j) \in A(G_m)$ que conecta los vértices v_i y v_j en G_m de forma tal que $v_i, v_j \in V(G_m)$.

Dos grafos G_q y G_t se dice que son isomorfos si hay una correspondencia uno a uno entre sus conjuntos de vértices $f: V(G_q) \rightarrow V(G_t)$, preservándose la adyacencia entre estos (dos vértices u y v son adyacentes si y solo si $f(u)$ y $f(v)$ son adyacentes en G_t), la relación en sí misma es llamada isomorfismo.

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 $u, v \in A$ 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) (36).

Un grafo reducido se define como el conjunto de centros descriptores (CDs), obtenidos a partir del método de reducción propuesto en la sección 2.1, cada uno de los CDs están ponderados mediante el cálculo de los índices híbridos para átomos (Electrotopográfico, Refractotopográfico y Lipotopográfico, definidos en la sección 1.7). 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 2.2), 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 $PMC(G_1, G_2)$, a los subgrafos de G_1 y G_2 que presentan la máxima similitud en las propiedades químico-físicas representadas por los índices ya señalados, entre los Centros Descriptores y la distancia euclidiana entre sus centros de masa.

2.6 Algoritmos implementados.

2.6.1 Búsqueda de fragmentos similares de segundo grado por índices topográficos e híbridos.

El objetivo de este algoritmo es realizar una comparación entre un fragmento molecular de segundo orden y los diferentes fragmentos de este tipo que se obtienen de la reducción y fragmentación de la molécula objeto de análisis, encontrando todos los fragmentos de estructura y propiedades similares, basado en los valores de los índices topográficos e híbridos (ponderados por propiedades químico-físicas) presentes en los átomos que los conforman, así como, en semejanzas en la estructura tridimensional de los CDs que los conforman, y en la distancia euclidiana entre los CDs correspondientes. El algoritmo implementado permite establecer las bases para el establecimiento de relaciones entre la estructura y la actividad biológica y puede describirse como un conjunto de pasos sencillos:

Primer paso: se define la función de similitud y una cota de mínima similitud la cual dependerá de la función seleccionada.

Segundo paso: se calcula el valor de los índices atómicos a cada compuesto, se reduce el grafo a los CDs según el procedimiento establecido y se calcula el valor total del índice de cada CD para cada índice.

Tercer paso: se obtienen los fragmentos moleculares de segundo orden y se construyen los vectores correspondientes, el cual estará formado por los índices Electrotopográfico, Refractotopográfico y Lipotopográfico de los centros descriptores que forman al fragmento, así como por la distancia euclidiana que separa a los centros de masa de cada uno de ellos.

Cuarto paso: se compara cada uno de los vectores obtenidos con el vector del fragmento seleccionado para el estudio y se calcula la semejanza existente entre los vectores obtenidos, utilizando la función de similitud seleccionada y se compara el valor alcanzado con la cota especificada. Si dicho valor sobrepasa la cota prefijada, se puede afirmar que los fragmentos analizados poseen semejanzas en cuanto a los descriptores empleados, independientemente de que no exista un macheo exacto entre las estructuras.

Algoritmo FragmentosSimilares(FM, M, u)

Inicio

1. $GR \leftarrow \text{GrafoReducido}(M)$
2. $LF \leftarrow \text{ObtenerFragmentosOrden2}(GR)$
3. Para cada F en LF Hacer
4. $V1 \leftarrow \{ \text{vector topográfico del fragmento } F \}$
5. $V2 \leftarrow \{ \text{vector topográfico del fragmento } FM \}$
6. $CS \leftarrow \text{CalculaSemejanza}(V1, V2)$
7. si $CS \geq u$ entonces
8. $FS \leftarrow \text{Adiciona}(F)$
9. fin si
10. fin Para

11. retorna FS
Fin

Tabla 2. Algoritmo para la búsqueda de fragmentos similares en moléculas

En la tabla 2 se muestra el pseudocódigo del algoritmo para la búsqueda de los fragmentos similares en una molécula dado un determinado fragmento. En primera línea, GR constituye el grafo reducido asociado a la molécula especificada. En la siguiente, línea LF contiene los fragmentos moleculares de segundo orden que son obtenidos a partir de las N en 2 combinaciones entre los centros descriptores del grafo reducido GR. A continuación, en las líneas 3-10 se realiza la comparación de cada uno de los fragmentos presentes en LF con el fragmento FM pasado por parámetros, para lo cual se obtienen los vectores topográficos de los dos fragmentos y se calcula la similitud existente. Si la misma supera el umbral especificado, el fragmento analizado se adiciona a la lista FS, que contiene todos los fragmentos de la molécula que cumplen la condición de semejanza. Finalmente el algoritmo retorna la lista de fragmentos similares encontrados.

2.6.2 Búsqueda de moléculas similares por fragmento de Propiedad Máxima Común.

El algoritmo para la búsqueda de moléculas similares por fragmento de 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.

Primer paso: se define la función de similitud y una cota de mínima similitud la cual dependerá de la función seleccionada.

Segundo paso: se calcula el valor de los índices atómicos a cada compuesto, se reduce el grafo a los CDs según el procedimiento establecido y se calcula el valor total del índice de cada CD para cada índice.

Tercer paso: se construye la matriz de semejanza por propiedades químico-físicas entre la molécula diana y la molécula a estudiar, donde la i-ésima fila contiene el centro descriptor i-ésimo de la primera molécula (molécula diana), la j-ésima columna representa el j-ésimo centro descriptor de la segunda molécula y la intercepción fila-columna muestra el índice de similitud entre CDs de ambas moléculas, calculado a partir del vector de propiedades del CD, utilizando la función de similitud seleccionada para la búsqueda.

Cuarto paso: se identifica la celda que posea el mayor valor de similitud. Si este es mayor o igual a la cota prefijada, se extraen ambos CDs y se asignan a la PMC. Este paso se hace iterativamente hasta que el valor máximo encontrado entre pares de CDs resulta menor que la cota prefijada.

Quinto paso: Una vez determinado el conjunto de centros descriptores con el mayor grado de semejanza en cada una de las moléculas, se construyen los vectores de distancia euclidiana utilizando una de las siguientes estrategias:

Selección lineal: consiste en construir el vector con las distancias asociadas a los centros descriptores adyacentes tomando el de la posición 1 con el 2, el 2 con el 3 y así sucesivamente hasta parear el de la posición N-1 con el de la posición N.

Selección circular: el vector se construye con las distancias entre cada uno de los centros descriptores tomando el de la posición 1 con la 2, la 2 con la 3 y así sucesivamente hasta calcular la distancia entre el de la posición N con el de la posición 1 nuevamente

Selección completa: aquí el vector se construye a partir del grafo completo, incluyendo las distancias euclidianas de cada CD con cada uno de los restantes CDs del grafo.

Sexto paso: se calcula el índice de semejanza existente entre los vectores de distancias euclidianas de los fragmentos identificados en los pasos anteriores utilizando el coeficiente de similitud seleccionado, se compara este valor obtenido con la cota definida para la búsqueda y en caso de que sea mayor, entonces los subgrafos obtenidos constituyen los fragmentos de Propiedad Máxima Común entre las moléculas estudiadas.

Algoritmo MoléculasSemejantesPMC(M1, M2, u)

Inicio

1. GR1 \leftarrow **GrafoReducido(M1)**
2. GR2 \leftarrow **GrafoReducido(M2)**
3. CD1 \leftarrow { centros descriptores del grafo reducido GR1 }
4. CD2 \leftarrow { centros descriptores del grafo reducido GR2 }
5. (F1, F2) \leftarrow **BuscarFragmentosPQF(CD1, CD2, u)**
6. V1 \leftarrow { vector de distancia euclideana del fragmento F1 de la molécula M1 }
7. V2 \leftarrow { vector de distancia euclideana del fragmento F2 de la molécula M2 }
8. CS \leftarrow **CalculaSemejanza(V1, V2)**
9. si CS \geq u entonces
10. retornar F
11. si no
12. retornar nulo

Fin

Tabla 3. Algoritmo para obtener los fragmentos PMC entre dos moléculas

En la tabla 3 se muestra el pseudocódigo del algoritmo para determinar si dos moléculas poseen los fragmentos de Propiedad Máxima Común y por consiguiente comparten cierto grado de semejanza

estructural y por propiedades químico-físicas. En las líneas 1-2 se obtienen los grafos químicos pertenecientes a las moléculas M1 y M2 y se calculan los grafos reducidos asociados, luego se almacenan en las variables CD1 y CD2 los centros descriptores obtenidos a partir del grafo reducido calculado. En la línea 5 a las variables F1 y F2 se le asignan los fragmentos similares por propiedades químico-físicas de cada una de las moléculas obtenidos del método BuscarFragmentosPQF, a continuación en las líneas 6-7 se generan los vectores de distancia euclídeana V1 y V2, que contendrán las distancias entre los centros descriptores de cada uno de los fragmentos, empleando una de las estrategias descritas anteriormente. En la línea 8 se calcula el coeficiente de similitud entre los vectores mediante una de las funciones de similitud seleccionadas. Finalmente, se verifica si el índice de semejanza calculado supera la cota de mínima similitud definido para la búsqueda y en caso afirmativo F1 y F2 constituyen los fragmentos que contienen la Propiedad Máxima Común entre las dos moléculas estudiadas.

Algoritmo BuscarFragmentosPQF(CD1, CD2, u)	
	Inicio
1.	MS \leftarrow ConstruyeMatrizSemejanza(CD1, CD2)
2.	VM \leftarrow 0
3.	Hacer
4.	VM \leftarrow { mayor valor en la matriz de semejanza MS }
5.	(PX; PY) \leftarrow { <i>fila-columna de VM</i> (PX; PY) \notin E }
6.	F1 \leftarrow Adiciona ({cd centro de descriptor de CD1 en la posición PX })
7.	F2 \leftarrow Adiciona ({cd centro de descriptor de CD2 en la posición PY })
8.	E \leftarrow Adiciona(PX; PY)
9.	Mientras VM \geq u
10.	retorna (F1, F2)
	Fin

Tabla 4. Algoritmo para buscar los fragmento PMC entre dos moléculas

En la tabla 4 se muestra el algoritmo mediante el cual se buscan los fragmentos similares por propiedades químico-físicas, dados los centros descriptores correspondientes a cada una de las moléculas, para ello en la línea 1 se construye la matriz de similitud topográfica que almacena los índices de similitud por propiedades químico-físicas entre cada uno de los centros descriptores de ambas moléculas. En las líneas 3-9 se seleccionan iterativamente los pares de centros descriptores, pertenecientes a cada una de las moléculas, que poseen mayor índice de similitud topográfica y se van eliminando las filas y columnas que ya han sido utilizadas, todo este proceso mientras este índice no disminuya de la cota mínima especificada. Finalmente, en la línea 10 se retorna la tupla F1, F2, que constituyen los fragmentos de similitud máxima, formado por los centros descriptores con mayor analogía en las propiedades químico-físicas.

Cuando en lugar de criterios de similitud, en los que se trabaja con valores entre 0 y 1 para la mayor semejanza, se hace con criterios de distancia, en lugar de fijar una cota mínima de similitud, se establece un umbral de distancia máxima permisible entre los valores de la propiedad y el algoritmo se ejecuta bajo esa restricción.

2.7 Lenguaje de programación: Java

Para el desarrollo de este trabajo se seleccionó como lenguaje de programación Java el cual es muy extendido en la actualidad. Desarrollado por la compañía Sun Microsystems, Java es un lenguaje de propósito general, concurrente, basado en clases y orientado a objetos 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 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.

2.8 Entorno de Desarrollo Integrado: Eclipse

Para la implementación de los algoritmos y las pruebas realizadas se utilizó como entorno de desarrollo el Eclipse, plataforma extensible, basada en Java y liberada bajo Licencia Pública 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 inicialmente por Alphaworks, laboratorio de desarrollo de IBM y actualmente es desarrollado por la Fundación Eclipse (37), 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.

2.9 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.

2.9.1 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. (38)

2.9.2 Chemistry Development Kit (CDK).

Se utiliza el Chemistry Development Kit (CDK) (39) pues esta es una librería de código abierto programada en Java para la química computacional y la química 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.

2.10 Conclusiones del capítulo.

En este capítulo se expuso el método para la reducción del grafo químico basado en centros descriptores, se explicó el procedimiento de fragmentación utilizado y se definieron las funciones de similitud / distancia seleccionadas, dentro de la gran variedad de funciones presentes en la bibliografía. Una vez definidos los principios fundamentales asociados a la búsqueda de similitud molecular y la teoría de grafos, se introdujo el concepto de Propiedad Máxima Común, el cual constituye uno de los aportes fundamentales de esta investigación. Se presentaron además los algoritmos implementados para la búsqueda de fragmentos similares de segundo orden y de moléculas similares mediante la obtención del fragmento PMC entre ambas. 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.

Capítulo 3: Resultados y discusión

En este capítulo se exponen los resultados obtenidos en la investigación y los experimentos realizados para validar los métodos de búsquedas desarrollados, en cada uno de estos se expresan los elementos que se tienen en cuenta para la evaluación de los mismos. Se muestran los cálculos realizados para definir las cotas de mínima similitud y los umbrales de máxima distancia con los que se trabaja en la investigación. 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 fragmentos simples y mediante el empleo del concepto de Propiedad Máxima Común.

3.1 Cálculo de los umbrales de similitud a utilizar.

Un elemento a tener presente cuando se realizan cálculos de semejanza mediante el empleo de funciones de similitud o de distancia, lo constituye el grado de error permisible que se estará utilizando. Es por ello, que antes de realizar las pruebas para validar los algoritmos implementados, se hace indispensable determinar la cota de mínima similitud o el umbral de máxima diferencia (distancia) de cada una de las funciones seleccionadas, el cual se estará empleando en las búsquedas que se realizarán y constituirán las propuestas de criterio de detección en esta investigación; aunque el criterio de determinación puede cambiar, por lo que queda a la decisión del investigador establecer el valor correspondiente que desee para sus experimentos. Para determinar las cotas y los umbrales se escogieron aleatoriamente once fragmentos moleculares de segundo orden, pertenecientes a moléculas presentes en el ensayo AID941 (40), con los cuales se realizaron búsquedas para ajustar los límites deseados. Cada uno de los fragmentos seleccionados se comparó con cada uno de los fragmentos presentes en las 330 moléculas contenidas en el ensayo AID941. En los experimentos realizados se considera un resultado correcto cuando los fragmentos obtenidos presentan valores similares con respecto al fragmento seleccionado e incorrecto en caso contrario.

En la tabla 5 se muestran los resultados del procedimiento realizado para el coeficiente de Tanimoto. Inicialmente se realiza la búsqueda para el fragmento compuesto por un anillo y un heteroátomo de

oxígeno, fijando un umbral de 0,2 para que se obtuvieran resultados tanto correctos como incorrectos. Se encontraron 77 fragmentos considerados similares, de los cuales 76 presentaron una estructura de anillo de diferentes dimensiones (6, 9, 10, 13 y 14) relacionado a un heteroátomo, excepto en un caso en el que, en lugar de un heteroátomo se identificó un clúster de orden 3 (Coef. Tanimoto=0,1996). En primera instancia, si el valor del umbral fue seleccionado correctamente, este resultado indicaría que entre estos fragmentos existe una equivalencia entre sus propiedades descritas por los índices empleados y que por lo tanto, esto pudiera considerarse la identificación de isostería no clásica basada en un conjunto de propiedades, a diferencia de las isosterías convencionales en las que se califican como de volumen (H vs F; Metilo vs TrifluoroMetilo; etc), o electrostáticas (TrifluoroMetilo vs Nitro) entre otras.

Ese valor de 0,1996 indujo a reducir el umbral de distancia 0,199, y con este nuevo umbral se realiza otra búsqueda. Se tomó un fragmento compuesto por un anillo y un heteroátomo de nitrógeno, obteniéndose 33 correctos y 69 incorrectos (con valores entre 0,0911 y 0,1987 de Coef. Tanimoto), por lo que se hace necesario un nuevo ajuste del umbral de forma tal que sea menor que el menor valor de diferencia obtenido dentro de los resultados incorrectos, fijándose el nuevo umbral hasta 0,091. Con esta nueva restricción, se realiza otra búsqueda, ésta vez con un fragmento compuesto por un anillo y un heteroátomo de cloro, encontrándose solo 2 fragmentos correctos, por lo que no es necesario disminuir más el umbral. Este procedimiento se realiza para cada uno de los 7 restantes fragmentos, modificando el valor si se encuentran resultados incorrectos y manteniéndolo igual en otro caso.

Tanimoto	Fragmento	Umbral	Bien		Mal			Molecula
			Cant	Max	Cant	Min	Max	
			Anillo-Heteroatomo[O]	0,200	76	0,1999	1	
Anillo-Heteroatomo[N]	0,199	33	0,1955	69	0,0911	0,1987	815515.mol	
Anillo-Heteroatomo[Cl]	0,091	2	0,0130	0			16189701.mol	
Anillo-Heteroatomo[Br]	0,091	5	0,0910	0			203396.mol	
Anillo-Heteroatomo[S]	0,091	5	0,0850	0			1253074.mol	
Anillo-Metilo	0,091	4	0,0850	1	0,0790	0,0790	664134.mol	
Anillo-Heteroatomo[F]	0,078	22	0,0670	0			675202.mol	
Anillo-Anillo	0,078	3	0,0680	0			665132.mol	
Anillo-Cluster	0,078	2	0,0500	0			661810.mol	
Cluster-Heteroatomo[N]	0,078	3	0,0767	0			1005277.mol	
Cluster-Cluster	0,078	1	0,0000	0			661810.mol	

Tabla 5. Ajuste del umbral para el coeficiente de Tanimoto.

El método descrito anteriormente se realizó para cada una de las funciones de distancia y de similitud seleccionadas para esta investigación, tomando como base los fragmentos seleccionados. Cabe destacar que en el caso de las funciones de similitud, el ajuste de la cota inferior se realizó tomando un valor mayor que el mayor valor de similitud obtenido dentro de los resultados incorrectos y en el

caso de las funciones de distancia, se realizó tal como se explicó anteriormente para el coeficiente de Tanimoto. En las tablas 6-11 se muestran los valores obtenidos en los experimentos realizados.

	Fragmento	Umbral	Bien		Mal			Molecula
			Cant	Min	Cant	Min	Max	
Jaccard	Anillo-Heteroatomo[O]	0,950	168	0,9503	17	0,9510	0,9779	701067.mol
	Anillo-Heteroatomo[N]	0,978	13	0,9782	6	0,9820	0,9946	5336330.mol
	Anillo-Heteroatomo[Cl]	0,995	2	0,9998	0			16189701.mol
	Anillo-Heteroatomo[Br]	0,995	4	0,9990	0			1487485.mol
	Anillo-Heteroatomo[S]	0,995	2	0,9971	0			3243696.mol
	Anillo-Metilo	0,995	3	0,9952	0			664134.mol
	Anillo-Heteroatomo[F]	0,995	20	0,9958	0			2136935.mol
	Anillo-Anillo	0,995	10	0,9967	0			2136935.mol
	Anillo-Cluster	0,995	1	1,0000	0			3243630.mol
	Cluster-Heteroatomo[N]	0,995	1	1,0000	0			5336330.mol
	Cluster-Cluster	0,995	1	1,0000	0			5336330.mol

Tabla 6. Ajuste de la cota para el índice de Jaccard.

	Fragmento	Umbral	Bien		Mal			Molecula
			Cant	Min	Cant	Min	Max	
Dice-Sørensen	Anillo-Heteroatomo[O]	0,980	134	0,9806	6	0,9801	0,9888	701067.mol
	Anillo-Heteroatomo[N]	0,989	12	0,9894	6	0,9909	0,9973	5336330.mol
	Anillo-Heteroatomo[Cl]	0,997	5	0,9974	0			2905495.mol
	Anillo-Heteroatomo[Br]	0,997	4	0,9994	0			203396.mol
	Anillo-Heteroatomo[S]	0,997	4	0,9970	0			649054.mol
	Anillo-Metilo	0,997	7	0,9975	0			649054.mol
	Anillo-Heteroatomo[F]	0,997	28	0,9971	0			5014184.mol
	Anillo-Anillo	0,997	1	1,0000	4	0,9972	0,9976	649054.mol
	Anillo-Cluster	0,998	2	0,9990	0			661810.mol
	Cluster-Heteroatomo[N]	0,998	2	0,9992	0			11835305.mol
	Cluster-Cluster	0,998	1	1,0000	0			661810.mol

Tabla 7. Ajuste de la cota para el coeficiente de Dice-Sørensen.

	Fragmento	Umbral	Bien		Mal			Molecula
			Cant	Max	Cant	Min	Max	
Sørensen	Anillo-Heteroatomo[O]	0,200	243	0,1999	41	0,1214	0,1988	2136882.mol
	Anillo-Heteroatomo[N]	0,120	229	0,1199	14	0,0825	0,1188	11839287.mol
	Anillo-Heteroatomo[Cl]	0,080	1	0,0000	0			727631.mol
	Anillo-Heteroatomo[Br]	0,080	3	0,0788	0			1893291.mol
	Anillo-Heteroatomo[S]	0,080	24	0,0775	1	0,0706	0,0706	2377612.mol
	Anillo-Metilo	0,070	81	0,0700	4	0,0484	0,0683	5771267.mol
	Anillo-Heteroatomo[F]	0,048	21	0,0422	0			6260437.mol
	Anillo-Anillo	0,048	4	0,0430	0			665132.mol
	Anillo-Cluster	0,048	6	0,0358	2	0,0471	0,0476	1215664.mol
	Cluster-Heteroatomo[N]	0,047	5	0,0456	0			1005277.mol
	Cluster-Cluster	0,047	2	0,0079	0			12005721.mol

Tabla 8. Ajuste del umbral para el coeficiente de Sørensen.

	Fragmento	Umbral	Bien		Mal			Molecula
			Cant	Max	Cant	Min	Max	
			Soergel	Anillo-Heteroatomo[O]	0,300	180	0,2999	
Anillo-Heteroatomo[N]	0,217	246		0,2167	14	0,1524	0,2123	11839287.mol
Anillo-Heteroatomo[Cl]	0,153	1		0,0000	0			727631.mol
Anillo-Heteroatomo[Br]	0,153	3		0,1460	0			1893291.mol
Anillo-Heteroatomo[S]	0,153	27		0,1525	1	0,1319	0,1319	2377612.mol
Anillo-Metilo	0,130	79		0,1296	4	0,0924	0,1279	5771267.mol
Anillo-Heteroatomo[F]	0,092	21		0,0809	0			6260437.mol
Anillo-Anillo	0,092	4		0,0824	0			665132.mol
Anillo-Cluster	0,092	6		0,0691	2	0,0900	0,0909	1215664.mol
Cluster-Heteroatomo[N]	0,090	5		0,0873	0			1005277.mol
Cluster-Cluster	0,090	2	0,0157	0			12005721.mol	

Tabla 9. Ajuste de la cota para el coeficiente de Soergel.

	Fragmento	Umbral	Bien		Mal			Molecula
			Cant	Min	Cant	Min	Max	
			Ruzicka	Anillo-Heteroatomo[O]	0,750	101	0,7534	
Anillo-Heteroatomo[N]	0,784	238		0,7842	14	0,7877	0,8476	11839287.mol
Anillo-Heteroatomo[Cl]	0,848	1		1,0000	0			727631.mol
Anillo-Heteroatomo[Br]	0,848	3		0,8526	0			1893291.mol
Anillo-Heteroatomo[S]	0,848	26		0,8500	1	0,8681	0,8681	2377612.mol
Anillo-Metilo	0,869	83		0,8690	5	0,8690	0,9076	5771267.mol
Anillo-Heteroatomo[F]	0,908	21		0,9191	0			6260437.mol
Anillo-Anillo	0,908	4		0,9176	0			665132.mol
Anillo-Cluster	0,908	6		0,9309	2	0,9091	0,9100	1215664.mol
Cluster-Heteroatomo[N]	0,910	5		0,9127	0			1005277.mol
Cluster-Cluster	0,910	2	0,9843	0			12005721.mol	

Tabla 10. Ajuste del umbral para el coeficiente de Ruzicka.

	Fragmento	Umbral	Bien		Mal			Molecula
			Cant	Min	Cant	Min	Max	
			Czekanowski	Anillo-Heteroatomo[O]	0,850	123	0,8503	
Anillo-Heteroatomo[N]	0,879	238		0,8790	14	0,8812	0,9175	11839287.mol
Anillo-Heteroatomo[Cl]	0,918	1		1,0000	0			727631.mol
Anillo-Heteroatomo[Br]	0,918	3		0,9205	0			1893291.mol
Anillo-Heteroatomo[S]	0,918	26		0,9189	1	0,9294	0,9294	2377612.mol
Anillo-Metilo	0,930	81		0,9300	4	0,9317	0,9516	5771267.mol
Anillo-Heteroatomo[F]	0,952	21		0,9578	0			6260437.mol
Anillo-Anillo	0,952	4		0,9570	0			665132.mol
Anillo-Cluster	0,952	6		0,9642	2	0,9524	0,9529	1215664.mol
Cluster-Heteroatomo[N]	0,953	5		0,9544	0			1005277.mol
Cluster-Cluster	0,953	2	0,9921	0			12005721.mol	

Tabla 11. Ajuste del umbral para el coeficiente de Czekanowski.

Una vez realizado el método de ajuste de las cotas y los umbrales de similitud y distancia para cada una de las 7 funciones de similitud/distancia seleccionadas para esta investigación, los valores finales calculados se muestran en la siguiente tabla.

Función	Umbral
Jaccard	0,995
Tanimoto	0,078
Dice-Sorensen	0,998
Sørensen	0,047
Soergel	0,09
Ruzicka	0,91
Czekanowski	0,953

Tabla 12. Umbrales calculados para las funciones de similitud/distancia.

Es necesario aclarar que los valores de cota/umbral calculados, fueron obtenidos a partir del algoritmo de búsqueda de fragmentos simples implementado y teniendo en cuenta los tres índices ponderados por propiedades químico-físicas para cada uno de los fragmentos analizados. Estos valores pueden cambiar para las búsquedas en las que solo se tenga en cuenta uno de los índices.

3.2 Utilización de la fragmentación y las propiedades químico-físicas.

Independientemente del principio de similitud molecular planteado por Johnson y Maggiora (4) de que sustancias semejantes presentan propiedades similares, se encuentran múltiples paradojas estructurales en las que moléculas estructuralmente similares exhiben propiedades biológicas diferentes y moléculas diferentes estructuralmente exhiben propiedades biológicas similares. Por lo antes planteado, el cotejo exacto de grafos no constituye la mejor elección a la hora de realizar búsquedas de similitud molecular. La necesidad de realizar comparaciones entre moléculas y encontrar un método para determinar similitud molecular, conlleva al empleo de descriptores capaces de diferenciar fragmentos y moléculas, no por el simple macheo de la estructura, sino por las propiedades que presenten. Esta capacidad de diferenciar átomos y grupos de átomos topológicamente idénticos es una capacidad demostrada de los índices híbridos ponderados por propiedades químico-físicas, los cuales fueron definidos anteriormente en el epígrafe 1.7.

En la figura 7 se muestran fragmentos moleculares que topológicamente son iguales y de acuerdo a la teoría de grafos son homólogos, debido a que presentan la misma cantidad de vértices y aristas. Sin embargo, a partir del cálculo de los valores del Índice del Estado Refractotopográfico, por ejemplo, de cada uno de átomos se encuentran diferencias entre cada uno de estos, las cuales están dadas por la ubicación espacial y el efecto del resto de los átomos de su entorno. Un ejemplo evidente de lo antes planteado lo constituye el átomo central en la Fig. 7B, debido a que en los tres fragmentos que se muestran, la propiedad calculada varía considerablemente, y esta variación está

dada por la influencia que ejercen el resto de los átomos sobre el mismo y la estructura que presentan.

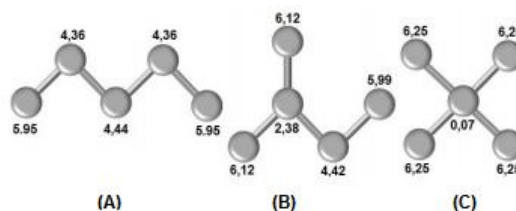


Figura 7. Valores del Índice del Estado Refractotopográfico de Átomos en fragmentos.

Otro ejemplo que ilustra lo planteado anteriormente lo constituyen los fragmentos mostrados en la figura 8, donde se muestra que los valores del Índice Refractotopográfico para Átomos varían con respecto a los mostrados anteriormente al intercambiar uno de los átomos de carbono por oxígeno (en rojo), evidenciándose que las propiedades químico-físicas cambian en dependencia del tipo de átomo y no solo de la estructura que presenten los fragmentos. Se demuestra además la importancia del empleo de estas propiedades a la hora de realizar un estudio de similitud molecular.

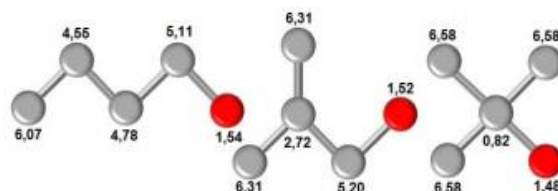


Figura 8. Valores del Índice del Estado Refractotopográfico para Átomos en fragmentos moleculares con un átomo de oxígeno (en rojo).

En la figura 9 se muestran otros fragmentos estructuralmente iguales, que también difieren debido a la existencia de átomos pesados en la misma posición relativa (F, Cl, Br, I). El cálculo correspondiente permite corroborar que los valores del índice del Estado Refractotopográfico para Átomos difieren al comparar átomo a átomo dada la influencia de cada uno sobre el resto de su entorno.

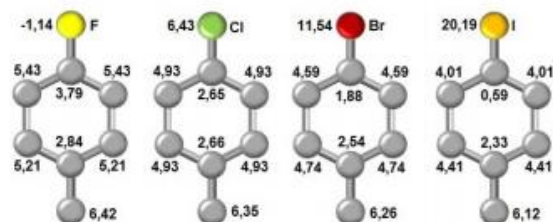


Figura 9. Fragmentos moleculares similares estructuralmente.

Otro elemento fundamental en esta investigación lo constituye la fragmentación molecular utilizando los centros descriptores definidos en el epígrafe 2.1. Como se aprecia en la tabla 13, los subgrafos mostrados son topológicamente idénticos y se obtienen fragmentos del mismo tipo formados por un anillo y un heteroátomo, excepto el cuarto que está formado por un anillo y un metilo siendo las propiedades de un metilo diferentes a las de un heteroátomo, y aunque sea un carbono aislado, como tiene hidrógenos enlazados, cambian sus propiedades químico-físicas, por lo tanto no se puede analizar de igual forma. Lo antes planteado demuestra la importancia de la fragmentación basada en CDs, debido a que la misma brinda un mayor contenido de información numérica.

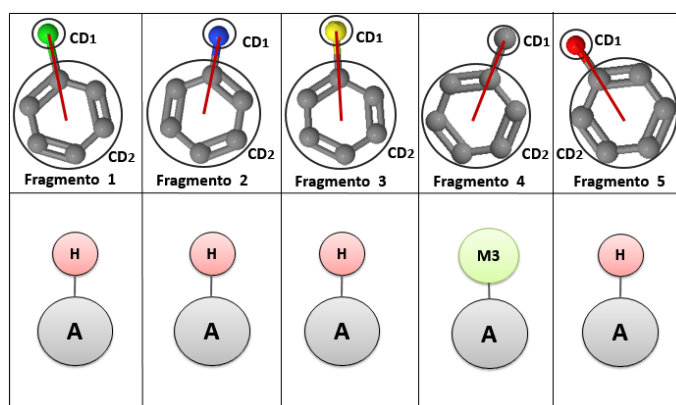


Tabla 13. Fragmentos moleculares y su tipo genérico.

Profundizando aún más en los fragmentos mostrados anteriormente, en la tabla 14 se muestran los valores de los Índices de cada uno de ellos, los cuales difieren átomo a átomo pero, en dependencia de la función y la cota o el umbral de similitud/distancia seleccionados, pueden ser diferentes o parecidos, al establecer las comparaciones entre fragmentos utilizando los índices topográficos e híbridos, el método de fragmentación por CDs y funciones de similitud/distancia.

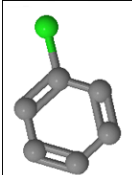
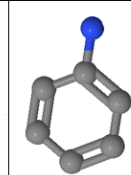
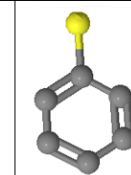
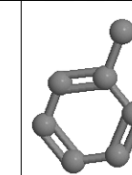
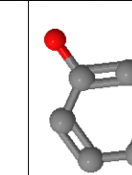
				
Fragmento 1	Fragmento 2	Fragmento 3	Fragmento 4	Fragmento 5
CD1: Heteroátomo ETPG : -0,1628 RTPG : 6,7320 LTPG : 1,4640	CD1: Heteroátomo ETPG : 1,8693 RTPG : 8,2311 LTPG : -0,7067	CD1: Heteroátomo ETPG : -5,9910 RTPG : 13,6820 LTPG : -1,5306	CD1: Metilo ETPG : 9,5916 RTPG : 8,2640 LTPG : 0,8924	CD1: Heteroátomo ETPG : 4,9754 RTPG : -0,4700 LTPG : -0,5504
CD2: Anillo ETPG : 14,5466 RTPG : 23,8616 LTPG : 1,7429	CD2: Anillo ETPG : 14,7585 RTPG : 30,5861 LTPG : 2,2154	CD2: Anillo ETPG : 13,6019 RTPG : 30,3202 LTPG : 3,3898	CD2: Anillo ETPG : 6,4621 RTPG : 24,8518 LTPG : 2,3127	CD2: Anillo ETPG : 12,5159 RTPG : 30,9916 LTPG : 3,3593
Dist: 3,1172	Dist: 2,6166	Dist: 3,1235	Dist: 2,2642	Dist: 2,1945

Tabla 14. Fragmentos moleculares y sus propiedades químico-físicas.

Otro de los criterios que permite validar la utilización de la fragmentación propuesta, en los estudios de similitud molecular se evidencia en la tabla 15. En la misma se presentan cuatro búsquedas tomando un mismo fragmento con diferentes valores de los índices Electrotopográfico, Refractotopográfico y Lipotopográfico como se observa en la parte inferior de la misma, además se puede observar que la cantidad de resultados encontrados para cada fragmento (A, B, C y D) es distinta, esta diferencia se debe fundamentalmente a la influencia que ejercen el resto de los átomos sobre el fragmento seleccionado para realizar la búsqueda y la estructura que presentan las moléculas en la colección de grafos moleculares.

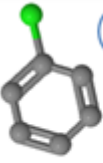
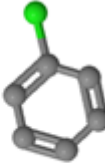
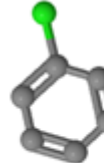
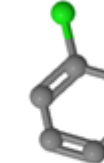
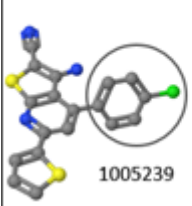
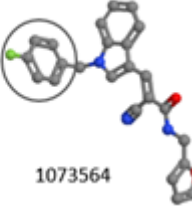
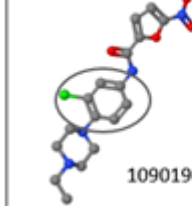
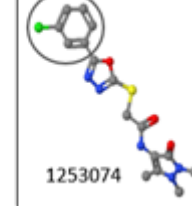
			
10	20	7	7
A	B	C	D
CD1: Heteroátomo [-0,1628 6,7320 1,4640]	CD1: Heteroátomo [1,8775 -3,7555 0,5759]	CD1: Heteroátomo [-0,5302 7,6128 1,8279]	CD1: Heteroátomo [-0,0643 6,8677 1,4694]
CD2: Anillo [14,5466 23,8616 1,7429]	CD2: Anillo [13,1968 29,5066 1,8947]	CD2: Anillo [12,6794 26,6897 2,5196]	CD2: Anillo [15,1821 26,9642 1,8990]
			
1005239	1073564	1090196	1253074

Tabla 15. Búsquedas de fragmentos en diferentes posiciones.

Todo los elementos planteados anteriormente evidencian que para realizar estudios de similitud molecular no se puede tener en cuenta solamente métodos que trabajen la estructura de las moléculas, sino que es necesario explorar alternativas como la propuesta en esta investigación mediante el empleo de una fragmentación que aporte información estructural y de descriptores híbridos ponderados con propiedades químico físicas que permita establecer distinciones entre las moléculas estudiadas.

3.3 Búsquedas de fragmentos simples.

Los fragmentos simples, como se explicó anteriormente son los subgrafos del grafo reducido que están formados por dos Centros Descriptores. A partir de estos fragmentos se forma un vector formado por los valores de los Índices de Estado Electrotopográfico, Refractotopográfico y Lipotopográfico para Átomos de cada Centro de Descriptor y la distancia entre los centro de masa

de cada uno de ellos. Esta representación en vectores de números reales, permite la aplicación de funciones matemáticas, que determinen la semejanza entre ellos.

En la Tabla 16 se muestra el fragmento A6-HCl1 perteneciente a la molécula 12005721 del ensayo AID941. En la misma aparecen los valores de índices y la distancia entre los CDs que lo forman.

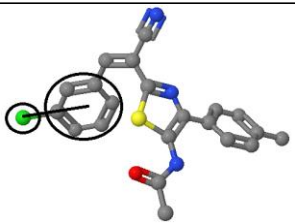
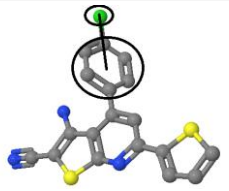
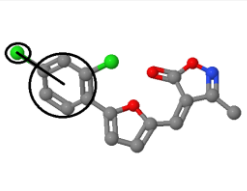
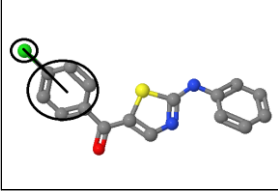
		Anillo6	Cl1
	ETPG	14,498	-0,227
	RTPG	24,123	6,866
	LTPG	1,720	1,499
	Dist	3,123	

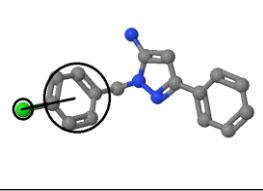
Tabla 16. Descripción del fragmento A6-HCl1 perteneciente a la molécula 12005721.

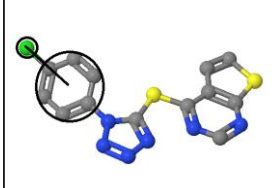
Tomando como objetivo el fragmento mostrado en la tabla 16, se realizaron varias ejecuciones del algoritmo implementado para la búsqueda de fragmentos similares (definido en el epígrafe 2.5.1) tomando cada una de las funciones de similitud/distancia seleccionadas para esta investigación. En la tabla 17 se muestran *los fragmentos comunes* encontrados por cada una de las funciones de similitud/distancia utilizadas.

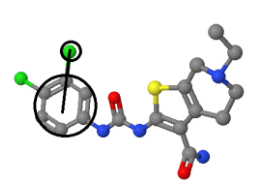
		Anillo6	Cl1
	ETPG	14,547	-0,163
	RTPG	23,862	6,732
	LTPG	1,743	1,464
	Dist	3,117	

		Anillo6	Cl2
	ETPG	13,900	-0,011
	RTPG	24,701	6,853
	LTPG	0,617	1,414
	Dist	3,117	

		Anillo6	Cl1
	ETPG	15,061	-0,030
	RTPG	24,904	6,686
	LTPG	1,550	1,398
	Dist	3,117	

		Anillo6	Cl1
	ETPG	13,866	-0,104
	RTPG	24,396	6,701
	LTPG	1,741	1,418
	Dist	3,119	

		Anillo6	Cl1
	ETPG	16,286	-0,001
	RTPG	24,618	6,692
	LTPG	1,621	1,422
	Dist	3,118	

		Anillo6	Cl1
	ETPG	13,770	-0,073
	RTPG	25,577	7,122
	LTPG	1,577	1,541
	Dist	3,121	

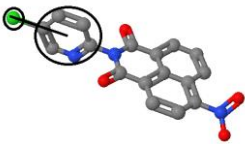
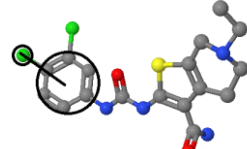
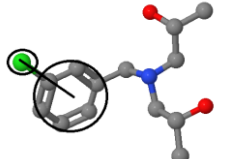
		Anillo6	Cl1
	ETPG	13,967	-0,104
	RTPG	25,075	7,183
	LTPG	1,585	1,541
	Dist	3,125	
		Anillo6	Cl2
	ETPG	13,770	0,031
	RTPG	25,577	6,913
	LTPG	1,577	1,460
	Dist	3,119	
		Anillo6	Cl1
	ETPG	12,955	-0,203
	RTPG	24,034	6,736
	LTPG	2,085	1,481
	Dist	3,119	

Tabla 17. Fragmentos comunes para cada una de las funciones de similitud.

A partir del análisis de los resultados, se evidencia que el algoritmo implementado permite encontrar aquellos fragmentos dentro del ensayo que poseen gran semejanza de sus propiedades químico-físicas expresadas por los índices, con el fragmento objetivo y a la vez presentan una notable semejanza estructural. En la figura 10 se muestra la cantidad de resultados devueltos por el algoritmo, por cada una de las funciones de similitud seleccionadas.

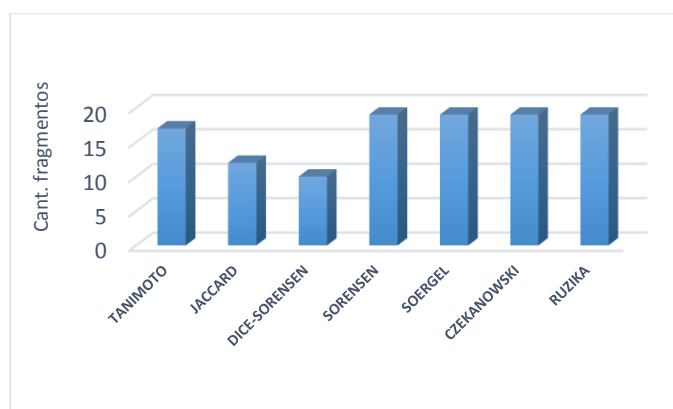


Figura 10. Resultados obtenidos por cada una de las funciones de similitud.

3.4 Búsquedas de fragmentos de Propiedad Máxima Común.

Quando se realizan estudios de similitud molecular, es necesario encontrar la parte entre dos moléculas que más se asemejan. Esta parte no constituye siempre un fragmento simple, es por ello que surge el concepto de fragmento de orden superior, los cuales, como se explicó en el epígrafe 2.2, son los subgrafos del grafo reducido que está compuesto por más de dos centros descriptores. Por otra parte, la Propiedad Máxima Común entre dos moléculas, es el máximo valor común de la propiedad de un par de fragmentos similares de cualquier orden presentes en moléculas diferentes, que se calcula a partir de los índices descritos en esta tesis.

En la figura 11 se muestra la molécula 100335, perteneciente al ensayo molecular AID941, en la que se señalan los diferentes centros descriptores a los que se reduce la misma y se muestran los valores de los índices de cada uno de ellos.

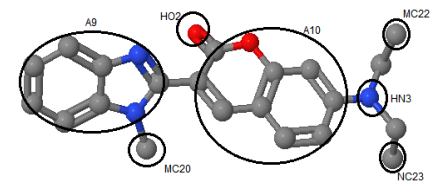
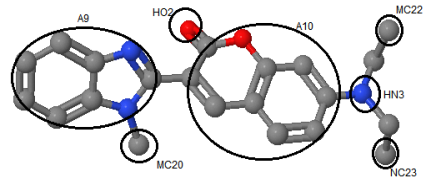
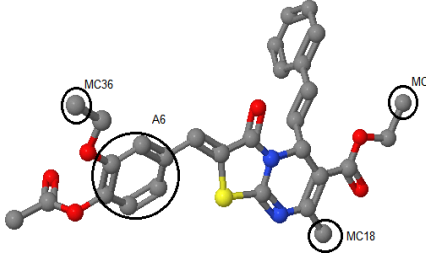
Molécula: 100335				
	CD	ETPG	RTPG	LTPG
	A-9	13,32172	28,92321	1,83878
	A-10	14,47367	40,59444	2,06244
	H-O2	5,21552	-2,88369	-0,16061
	H-N3	-1,89187	1,69889	0,09245
	M-20	10,44022	9,58385	0,03903
	M-22	8,47965	6,7708	0,13414
	M-23	8,39183	6,67828	0,12774


Figura 11. Molécula 100335 perteneciente al ensayo AID941.

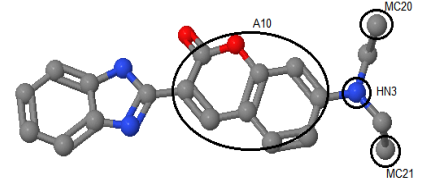
Tomando como diana la molécula mostrada en la figura 11, se realizó una búsqueda de similitud utilizando el algoritmo implementado para la búsqueda de moléculas similares por fragmento de Propiedad Máxima Común (definido en el epígrafe 2.5.2), tomando el coeficiente de Tanimoto como función de similitud y la estrategia de selección completa, para obtener los vectores de distancias entre los centros descriptores. La búsqueda se realizó en las 330 moléculas presentes en el ensayo molecular AID941, los resultados obtenidos se muestran en la tabla 19.

En la tabla 19 se muestra como el algoritmo implementado es capaz de encontrar fragmentos, además de en la propia molécula seleccionada demostrándose la exactitud del mismo, otras moléculas que lo contienen en un rango de valores razonablemente lógico. En la misma se señalan los CDs encontrados por cada una de las moléculas, así como los valores de los índices topográficos e híbridos asociados a cada uno de ellos.

Molécula: 100335					Molécula: 100335
	CD	ETPG	RTPG	LTPG	
	A-9	13,32172	28,92321	1,83878	
	A-10	14,47367	40,59444	2,06244	
	H-O2	5,21552	-2,88369	-0,16061	
	H-N3	-1,89187	1,69889	0,09245	
	M-20	10,44022	9,58385	0,03903	
	M-22	8,47965	6,7708	0,13414	
M-23	8,39183	6,67828	0,12774		

Molécula: 11957307					Molécula: 100335
	CD	ETPG	RTPG	LTPG	
	M-36	8,40014	7,21344	0,16929	
	A-6	11,45031	28,91058	2,70509	
	M-18	10,41034	9,55584	0,59352	
	M-33	8,82852	7,40027	0,17869	
	M-22	8,47965	6,7708	0,13414	
	M-23	8,39183	6,67828	0,12774	

Molécula: 3243384					Molécula: 100335
	CD	ETPG	RTPG	LTPG	
	A-10	14,4333	44,09157	2,46372	
	M-19	8,30347	6,67386	0,15047	
	M-18	8,46377	6,81715	0,15776	
	A-10	14,47367	40,59444	2,06244	
	M-22	8,47965	6,7708	0,13414	
	M-23	8,39183	6,67828	0,12774	

Molécula: 94381					Molécula: 100335
	CD	ETPG	RTPG	LTPG	
	A-10	15,55965	40,98299	2,10978	
	M-20	8,47289	6,7665	0,13498	
	H-N3	-1,84854	1,71768	0,09385	
	M-21	8,38483	6,67408	0,1289	
	A-10	14,47367	40,59444	2,06244	
	M-23	8,39183	6,67828	0,12774	

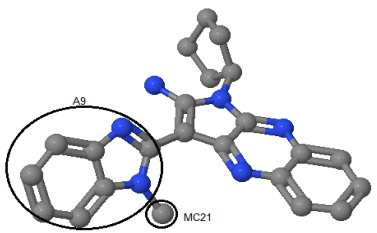
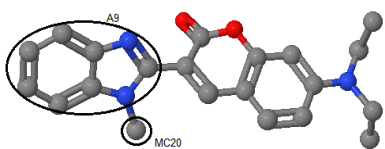
Molécula: 666746					Molécula: 100335				
	CD	ETPG	RTPG	LTPG		CD	ETPG	RTPG	LTPG
	A-9	13,458	29,67264	2,29144		A-9	13,32172	28,92321	1,83878
M-21	11,21911	10,31396	0,09136	M-20	10,44022	9,58385	0,03903		

Tabla 18. Resultados de la búsqueda realizada.

Como se evidencia en los resultados, los fragmentos encontrados en cada una de las moléculas, presentan valores de las propiedades químico-físicas similares, y a la vez poseen una distribución espacial semejante, verificándose en las distancias que separan a cada uno de los centros descriptores señalados en cada una de las moléculas, las cuales muestran homogeneidad en los fragmentos encontrados.

3.5 Conclusiones del capítulo.

En este capítulo se explicó el procedimiento para el cálculo de los umbrales de similitud para cada una de las funciones de similitud/distancia seleccionadas para la investigación y se presentaron los valores finales obtenidos para cada uno de ellos, los que constituyen una propuesta de este trabajo, aunque cada investigador es libre de especificar sus propios valores a la hora de realizar las búsquedas con los algoritmos implementados. Se expusieron los elementos fundamentales que validan el empleo de la fragmentación utilizando el grafo reducido mediante centros descriptores, así como de los descriptores híbridos ponderados por propiedades químico-físicas y la importancia de los mismos en los estudios de similitud molecular independientemente a las paradojas estructurales existentes. Se realizaron varios experimentos que permiten validar la efectividad de los algoritmos implementados y su importancia en la búsqueda de similitud molecular.

Conclusiones generales

- Se desarrollaron dos procedimientos de búsquedas en bases de datos de compuestos químicos por macheo inexacto de grafos basado en descriptores topográficos e híbridos que permitió la identificación de fragmentos con valores similares de las propiedades y revelar fragmentos similares en propiedades con estructuras diferentes.
- Se validó la pertinencia al empleo de índices topográficos e híbridos junto a la reducción del grafo molecular por CDs para la identificación de fragmentos semejantes equivalentes en sus propiedades.
- Se define el concepto de Propiedad Máxima Común como criterio de identificación de máxima similitud entre moléculas, a partir de descriptores topográficos e híbridos y la fragmentación por CDs.

Recomendaciones

- Extender las búsquedas a otros coeficientes de similitud para incrementar la generalización de los resultados obtenidos.
- Continuar optimizando los algoritmos implementados, mediante el empleo de técnicas de programación concurrente y de alto rendimiento.

Referencias Bibliográficas

1. **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.
2. **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.
3. *Modelos QSAR utilizando técnicas de Softcomputing*. **Molina Souto, Yania y Carrasco Velar, Ramón.** 2, La Habana : UCI, 2010, Vol. 3, pág. 2.
4. **Johnson, M. y G., Maggiora, [ed.].** *Concepts and Applications of Molecular Similarity*. New York : John Wiley & Sons, Inc., 1990. pág. 393.
5. **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.
6. *New trends in bioinformatics: from genome sequence to personalized medicine*. **Molidor, R., y otros.** 3, 2003, Exp Gerontol., Vol. 38, págs. 1031-6.
7. **Todeschini, R., y otros.** *Molecular Descriptors for Chemoinformatics*. [ed.] H. Kubinyi, G. Folkers R. Mannhold. Germany : Wiley-VCH, 2009.
8. **Flores Garrido, Marisol , Carrasco Ochoa , J. Ariel y Martínez Trinidad, José Fco. .** *Búsqueda de patrones interesantes en un solo grafo utilizando correspondencia inexacta*. Coordinación de Ciencias Computacionales, Instituto Nacional de Astrofísica, Óptica y Electrónica. Mexico : s.n., 2012. pág. 42, Técnico.
9. **Yan, Xifeng y Han, Jiawei.** *gSpan Graph-Based Substructure Pattern Mining*. Department of Computer Science, University of Illinois. Urbana-Champaign : s.n., 2002. Technical Report.
10. *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.
11. **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.

12. **Rathore, Àli.** *Data Mining of Chemical Compounds Using Functional Groups*. s.l., India : Chabot College, 2009.
13. **Politécnica, Universidad.** *DMATIC*. [En línea] 2014. [Citado el: 22 de marzo de 2015.] <http://www.dma.fi.upm.es/gregorio/grafos/SucGrafCertifArboles/html/Isomorfismo%20de%20grafos.htm>.
14. **Yan, Xifeng y Han, Jiawei .** *CloseGraph: Mining Closed Frequent Graph Patterns*. Illinois : s.n., 2003. Informe Técnico.
15. **Huan, Jun, y otros.** *SPIN Mining Maximal Frequent Subgraphs from Graph Databases*. 2004.
16. **Thomas, Lini T. , Valluri, Satyanarayana R. y Karlapalem, Kamalakar.** *MARGIN: Maximal frequent subgraph mining*. s.l. : Center For Data Engineering, IIIT, Hyderabad, 2010.
17. **Chen, Chen, y otros.** *gApprox: Mining Frequent Approximate Patterns from a Massive Network*. [ed.] University of Illinois. Illinois, USA : s.n., 2007.
18. *An efficient graph-mining method for complicated and noisy data with realworld applications.* **Jia, Y., Zhang, J. y Huan, J.** 2011, Knowl Inf Syst , Vol. 28, págs. 423–447.
19. **Saeedy, Mohammed El y Kalnis, Panos .** *GraMi: Generalized Frequent Pattern Mining in a Single Large Graph*. Division of Mathematical and Computer Sciences and Engineering, King Abdullah University of Science and Technology (KAUST). 2011. Technical Report.
20. *AGraP: an Algorithm for Mining Frequent Patterns in a Single Graph using Inexact Matching.* **Flores Garrido, Marisol, Carrasco Ochoa, Jesús Ariel y Martínez Trinidad, José Fco.** s.l. : Springer, 2014, Knowledge and Information Systems, pág. 1.
21. *Definition of a novel atomic index for QSAR: the refractotopological state.* **Carrasco-Velaz, R., Padrón, J. A. y Gálvez, J.** [ed.] Alberta University. 1, Alberta : Canadian Society for Pharmaceutical Sciences., 2004, J Pharm Pharm Sci. , Vol. 7, págs. 19-26.
22. **Carrasco-Velaz, Ramón.** *Nuevos descriptores atómicos y moleculares para estudios de estructura-actividad. Aplicaciones*. La Habana : Editorial Universitaria, 2007. 978-959-16-0646-4.
23. **Acosta Mendoza, Niusvel, Gago Alonso, Andrés y Medina Pagola, José E.** *Minería de subgrafos frecuentes utilizando cotejo inexacto de grafos*. CENATAV. La Habana : s.n., 2011. pág. 50, Informe Técnico.

24. *Una breve introducción a la teoría de grafos*. **Menéndez-Velázquez, Amador**. 28, Oviedo : Universidad de Oviedo, 1998, SUMA, págs. 11-26.
25. **Ivanciuc, Ovidiu y Balaban, Alexandru T**. Graph Theory in Chemistry. [ed.] N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, and P. R. Schreine P. v. R. Schleyer. *The Encyclopedia of Computational Chemistry*. Chichester : John Wiley & Son, 1998, págs. 1169-1190.
26. **Rodríguez-Blancas, J. L.** Un paseo por la topología en la Web. [En línea] [Citado el: 8 de Abril de 2015.] <http://www.ual.es/~jlrodri/Topgen5/introduccion.html>.
27. **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.
28. *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.
29. *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.
30. **Costales Leiva, Lien y Guirola González, Asnay. Tutores: R: Carrasco-Velaz y A. Antelo-Collado**. *Predicción de actividad anticancerígena de compuestos orgánicos partiendo de descriptores, utilizando programación genética*. La Habana : UCI, 2007.
31. **Todeschini , R y Consonni , V**. Handbook of Molecular Descriptors. 2000.
32. **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/>.
33. *Hybrid reduced graph for SAR studies*. **Carrasco Velaz, R., y otros**. s.l. : Taylor y Francis, 2013, SAR and QSAR in Environmental Research.
34. **Carrasco-Velaz, Ramón**. Comunicación personal. 2015.
35. *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, Pupo-Merino, Mario y Carrasco-Velaz, Ramón**. 3, La Habana : CENIC, 2011, Revista CENIC Ciencias Biológicas, Vol. 42, págs. 111-118.

36. **Asad Rahman, Syed, y otros.** *Small Molecule Subgraph Detector (SMSD) toolkit*. s.l. : Journal of Cheminformatics, 2009. págs. 2-3.
37. **The Eclipse Foundation.** Eclipse. [En línea] The Eclipse Foundation, 2004. [Citado el: 7 de Noviembre de 2014.] <https://eclipse.org/>.
38. **JMol.** JMOL. [En línea] 2015. [Citado el: 9 de Noviembre de 2014.] <http://jmol.sourceforge.net/>.
39. *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.
40. **National Center for Biotechnology Information.** <http://www.ncbi.nlm.nih.gov>. [En línea] NCBI, 2015. [Citado el: 15 de Enero de 2015.] <http://www.ncbi.nlm.nih.gov/guide/chemicals-bioassays/>.

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.
 - **Átomo:** Partícula más pequeña de un elemento químico que retiene las propiedades asociadas con ese elemento.
 - **Bioinformática:** El uso de las matemáticas aplicadas, la estadística y la ciencia de la informática para estudiar sistemas biológicos.
 - **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.
 - **Descriptor:** Número que describe la estructura química o una propiedad de la molécula o fragmento de ésta.
 - **Fármacos:** Término farmacológico para cualquier compuesto biológicamente activo, capaz de modificar el metabolismo de las células sobre las que hace efecto.
 - **Índice topológico:** Número que se calcula generalmente a partir de la matriz de adyacencia o de distancias de los elementos de un grafo molecular.
 - **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.
 - **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.
 - **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.
 - **Molécula:** Es la partícula de una sustancia que retiene todas las propiedades de la misma y está compuesta por uno o más átomos.
-

- **Principios Activos:** Sustancias químicas o biológicas a las que se les atribuye una actividad determinada para constituir un medicamento.
 - **Propiedad Máxima Común:** Dados los grafos $G1$ y $G2$, se entiende por fragmentos de Propiedad Máxima Común $PMC(G1, G2)$, a los subgrafos maximales de $G1$ y $G2$ que presentan la mayor similitud en las propiedades químico-físicas entre los vértices y en la longitud de las aristas que los unen.
 - **Química Computacional:** Es una rama de la química que utiliza computadores para ayudar a resolver problemas químicos. La química computacional es ampliamente utilizada en el diseño de nuevas drogas y materiales.
 - **Sintetizar:** Proceso de obtención de un compuesto a partir de sustancias más sencillas.
 - **Topografía molecular:** Es la información que puede obtenerse de la distribución espacial de los átomos de una molécula.
 - **Topología molecular:** Es toda la información (y la única) que puede obtenerse de la conectividad mutua entre todos los pares de átomos en una molécula.
-