Universidad de las Ciencias Informáticas

Facultad 2

Métodos de *clustering* para la obtención de fragmentos representativos en una colección de grafos moleculares utilizando descriptores híbridos

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

Autor: Mario Antonio Ruiz Acuña

Tutores: MSc. Aurelio Antelo Collado
Dr. Ramón Carrasco Velar

La Habana, junio de 2016

"Año 58 de la Revolución"

Declaración de autoría

Declaro ser autor de la presente tesis y reconozco 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 _____.

_____

Mario Antonio Ruiz Acuña

Autor

_____

Dr. Ramón Carrasco Velar

Tutor

_____

Msc. Aurelio Antelo Collado

Tutor

Agradecimientos

Quiero agradecer en primer lugar a Dios por quien trato de hacerlo todo, y por ser siempre el principal motivo de mi dedicación en las cosas y de manera especial en este trabajo: “Gracias Jesús”. También agradecer a toda mi familia por apoyarme en sus oraciones y ayudarme en todo cuanto necesite en estos cinco años de carrera. Gracias a mi mami por su amor y su apoyo, gran parte de este logro fue por ti, a mi hermano Pepo por ser un gran amigo, a mis abuelos por sus consejos, a mis Tíos Mario y Rebeca por ser mis padres en estos cinco años, a mi Tío Ernesto Ruiz por su ayuda material para poder realizar este trabajo y estudiar de manera eficaz esta carrera, a toda la familia Acuña. A mi futura esposa Sandra quien ha completado mi felicidad en esta etapa de mi vida, a todos mis compañeros en estos cinco años de carrera. A todos mis hermanos en la fe, gracias por sus oraciones, y a mis tutores por su colaboración en este trabajo. MUCHÍSIMAS GRACIAS A TODOS.

Resumen

En el presente trabajo se hace uso de los métodos de agrupamiento (*clustering*) para encontrar fragmentos representativos a partir de la fragmentación de las moléculas presentes en un ensayo (colección de grafos moleculares). Se proponen dos variantes de estructura de datos a analizar en los algoritmos de *clustering*, *Propiedad Máxima Común* y por rasgos, describiendo los grafos moleculares (moléculas) a través de descriptores híbridos (índice de Estado Electrotopográfico (Si), Lipotopográfico (Λ_{3D}) y Refractotopográfico ($\Re_{3D}$) ponderados para átomos). Se implementaron dos métodos de *clustering* (Basado en Grafo y Jarvis Patrick) y cuatro algoritmos para el cálculo de similitud entre fragmentos, tres por rasgos y uno utilizando *Propiedad Máxima Común*.

Como resultado del trabajo se logran identificar 441 fragmentos representativos de grados diferentes, de los 2239 presentes en las 330 moléculas del ensayo AID941, utilizando el método de *clustering* Basado en Grafo y el algoritmo PMC con la función de similitud Soergel, los cuales pueden ser los responsables de la actividad farmacológica del 79% de las moléculas activas de dicho ensayo.

Palabras Claves: *clustering* de fragmentos moleculares; métodos de agrupamiento; selección de compuestos; fragmentos representativos.

Índice

Índice de tabla

Índice de figuras

Introducción

El uso de medicamentos es parte de la vida de un ser humano, en 30 años se pasó de tener apenas una decena de medicamentos activos y fiables a disponer de un número considerable de ellos para una variedad importante de enfermedades y, por tanto, de una exposición limitada a una exposición masiva (1). Por otro lado, se conoce que aproximadamente el 4 % de las nuevas entidades químicas y biológicas que se introducen en el mercado deben ser retiradas por el descubrimiento de reacciones adversas no conocidas o no bien cuantificadas durante el desarrollo clínico (2). Es por eso que el proceso de investigación y fabricación de fármacos es meramente complejo, además del proceso de optimización de fármacos desarrollados, buscando compuestos más potentes y menos tóxicos que los existentes.

Uno de los objetivos esenciales de la química medicinal es el descubrimiento de nuevos agentes terapéuticos, haciendo uso de la información recopilada a través de la historia en lo que respecta a la producción de fármacos, revelando la evolución que ha tenido esta disciplina no solo en las metodologías de descubrimiento de fármacos guiadas por la intuición y el empirismo sino también en las técnicas computacionales que se emplean en la actualidad (3).

En tiempos pasados, las principales limitantes para la obtención de nuevos medicamentos consistían en los costos de tiempo y dinero en todo el proceso de descubrimiento y desarrollo, sin embargo, en la actualidad se busca racionalizar el desarrollo de nuevos agentes terapéuticos basándose en la relación que existe entre la estructura química del potencial fármaco con su actividad biológica (estudios *Structure-Activity Relationship* (SAR) y *Quantitative Structure-Activity Relationships* (QSAR)) (3).

El proceso moderno de descubrimiento y desarrollo de fármacos es largo y puede comenzar con la identificación de una molécula diana o blanco terapéutico. Este por lo general es una proteína con una función determinada en el organismo (4), que juega un papel fundamental en una determinada patología, y es necesario regular su actividad biológica mediante la unión de un compuesto químico (ligando) el cual se convierte en un potencial agente terapéutico.

Una vez identificado el blanco, la siguiente fase del proceso para la creación de un nuevo fármaco, es la búsqueda de compuestos líderes o cabezas de serie, o sea, un compuesto prototipo que tiene la actividad biológica o farmacológica deseada (inhibición o bloqueo del blanco) pero que puede presentar también ciertas características contraproducentes: alta toxicidad, insolubilidad, inestabilidad, problemas metabólicos, actividades biológicas secundarias, etc. Esta fase es la más compleja en todo el proceso, ya que las cantidades de moléculas disponibles para esta búsqueda superan los millones. Todo este

gran volumen de información se encuentra almacenada en múltiples bases de datos. Entre las más populares se encuentran: el banco de datos ZINC, creado por las compañías de síntesis química más importantes de mundo, con más de 13 millones de moléculas, la base de datos PubChem, (5); NCBI-USA base de datos que contiene alrededor de 20 millones compuestos evaluados como anticancerígenos, National Cancer Institute (EEUU) Database, (6); los repositorios de datos de la Sociedad de Quimioinformática y QSAR, (7) y los conjuntos de datos de la Academia Internacional de Química Matemática, (8) todas ellas disponibles gratuitamente; se pueden incluir también la base de datos MDDR (MDL Drug Data Report); las base de datos WDI (World Drug Index), (9); y la base de datos WOMBAT (World of Molecular Bioactivity Data), todas comerciales (10).

Para la búsqueda de un compuesto líder o cabeza de serie en las bases de datos existentes, se requiere de la existencia de uno o más ensayos biológicos, que permitan determinar con rapidez, la actividad biológica de los nuevos compuestos a desarrollar. Otra forma para desarrollar esta búsqueda, es a partir del aislamiento de las moléculas responsables de una acción biológica determinada, pertenecientes a algún producto de origen vegetal o animal (11).

Otra de las tareas importantes a realizar en esta búsqueda es la identificación del *farmacóforo.* Su primera definición fue dada por Ehrlich en 1909 cuando planteó que un farmacóforo consiste en "*una unidad central molecular que transporta (phoros) los rasgos esenciales responsables para la actividad biológica de una droga (=pharmacon)"* (12).

Otra definición fue la planteada por Escalona y cols., los cuales los definen como *el conjunto de grupos químicos, unidos o no entre sí, que todas las moléculas activas sobre un mismo receptor tienen en común, y que son esenciales para el reconocimiento por el mismo* (13).

Un elemento a tener en consideración en la búsqueda de moléculas con actividad biológica semejante, es el principio de similitud molecular, que afirma que las moléculas que son similares en su estructura pueden tener actividad biológica semejante (14), aunque este concepto es intuitivo y ambiguo, y está soportado por muchas observaciones (15), los químicos también han demostrado que pequeños cambios químico-estructurales en una molécula pueden modificar sus propiedades (16). Por lo que se puede afirmar que moléculas estructuralmente similares pueden tener actividad biológica diferente, así como moléculas estructuralmente diferentes puedan tener actividad biológica similares (17), dando paso a las paradojas estructurales.

La veracidad de este fenómeno ha conllevado a la adecuación de un nuevo campo de investigación llamado cotejo inexacto de grafos, en el cual existen varios métodos que se encargan de realizar búsquedas de subgrafos ponderados en colecciones de grafos (18), (19), que en este caso son grafos

moleculares. En la literatura se han reportado varios estudios comparativos entre técnicas de búsqueda de similitud resaltando sus méritos y deficiencias (20). Sin embargo, como Sheridan y Kearsley han señalado, es muy poco probable que un solo mecanismo de búsqueda pueda comportarse consistentemente superior a los demás en todos los problemas (21), en otras palabras, que no todos los métodos de búsqueda de fragmentos en colecciones de grafos funcionan eficientemente para todos los problemas; por lo que se puede afirmar que la búsqueda de fragmentos representativos en una colección de grafos sigue siendo un campo abierto dentro de la similitud molecular.

En un trabajo previo realizado dentro del proyecto *Visualización y Minería de Grafos Ponderados para la Quimioinformática,* que se ejecuta en la Universidad de las Ciencias Informáticas (UCI), se desarrolló un trabajo de diploma por Paneque y Govea (22), en la que los autores muestran, usando varias funciones de similitud molecular (23), la existencia de subgrafos en pares de moléculas, que aunque sean estructuralmente diferentes, poseen un valor máximo similar en las propiedades químico-físicas representadas por los índices Electrotopográfico (Si), Lipotopográfico (Λ_{3D}) y Refractotopográfico ($\Re_{3D}$) para átomos (24). No obstante, estos resultados no le permiten al especialista identificar la parte activa (fragmento molecular; farmacóforo) de la molécula a la cual se le atribuye la actividad biológica, además, no logra organizar toda la información adquirida enfocándola en la relación de todos los fragmentos encontrados, para su posterior análisis aplicando técnicas de inteligencia artificial. Por lo que se hace necesario identificar los fragmentos representativos en colecciones de moléculas (grafos moleculares).

Por lo anterior se plantea como **problema** a resolver: ¿Cómo identificar fragmentos representativos en colecciones de grafos moleculares utilizando descriptores híbridos?, definiéndose como **objeto de estudio**: Búsqueda de subgrafos moleculares representativos, centrándose en el **campo de acción**: Agrupamiento de subgrafos moleculares presentes en una colección. Para dar solución al problema de la investigación se define como **objetivo general**: Identificar fragmentos representativos en una colección de grafos moleculares a partir de métodos de *clustering* utilizando descriptores híbridos.

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

1. Selección de los métodos de *clustering* que mejor se adaptan al dominio de subgrafos moleculares y a los descriptores híbridos.
2. Definición de estructuras de datos a partir de la distancia y los índices topográficos de los subgrafos moleculares.
3. Definición de diferentes formas de expresión de rasgos como transformación de los datos.
4. Selección de diferentes medidas de similitud entre fragmentos moleculares.

5. Implementación de algoritmos para el cálculo de similitud entre fragmentos utilizando la medida de similitud seleccionada y las formas de expresión de rasgos definidas.
6. Implementación del algoritmo para el cálculo de similitud entre fragmentos utilizando *Propiedad Máxima Común*.
7. Reutilización de algoritmos de *clustering*, utilizando las medidas de similitud seleccionadas y las formas de expresión de rasgos definidas.
8. Reutilización de algoritmos de *clustering*, utilizando las medidas de similitud seleccionadas y la *Propiedad Máxima Común*.
9. Validación de los algoritmos en diferentes colecciones de grafos moleculares.

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 valoren los resultados obtenidos con los algoritmos de *clustering* propuestos.
- **Pruebas**: se utilizan para comprobar si los algoritmos de *clustering* propuestos obtienen resultados aceptables.

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 representativos a partir de los fragmentos moleculares similares encontrados en una colección de moléculas. 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.

Capítulo 1: Fundamentación teórica

En este capítulo se ofrece una perspectiva general del estudio molecular y su aplicación en la obtención de fármacos. Se define el concepto de grafo químico, así como su aplicación en la química molecular. Además, se presenta una recopilación de las funciones de similitud o diferencia más usadas en el cálculo de similitud molecular, el concepto de *Propiedad Máxima Común* como método de búsqueda de similitud y los conceptos que definen a una función como métrica. También se aborda sobre los principales métodos de *clustering* y algunas de las herramientas que existen para su aplicación. Finalmente se exponen los métodos de *clustering* que, según el estudio realizado, mejor se adaptan al dominio de subgrafos moleculares y a los descriptores híbridos, los cuales serán utilizados para realizar la investigación.

1.1.- Diseño y obtención de fármacos.

En el proceso de diseño y fabricación de fármacos, una fase crítica es la identificación de compuestos líderes (*leads*) o cabezas de series. Tales compuestos se obtienen a partir de diferentes enfoques que van, desde la idea del investigador, determinada en muchos casos por su experiencia personal, hasta el HTS (High Throughtput Screening), una técnica moderna para la obtención de hasta cientos de miles de compuestos, pasando por el cribado virtual en bases de datos o *Virtual Screening*. En los dos últimos enfoques se hace indispensable el uso de métodos y herramientas que logren una optimización en tiempo y recursos, así como la identificación farmacológica de los cabezas de series.

En cuanto se logran identificar los compuestos líderes, comienza un proceso denominado *lead optimization,* cuyo objetivo principal es introducir modificaciones estructurales para mejorar su eficacia terapéutica, bien sea incrementando su potencia y especificidad como minimizando su toxicidad y otros efectos secundarios (25).

La técnica de cribado virtual, consiste en el análisis computacional de bases de datos de compuestos químicos, dirigido a identificar y seleccionar un número limitado de candidatos que posean la actividad biológica deseada sobre un blanco terapéutico específico (26).

Existen también los métodos indirectos o basados en la estructura del ligando, los cuales se emplean cuando no se conoce la estructura del receptor biológico o molécula diana. Entre los más empleados están los que se conocen como estudios QSAR (25); estos métodos parten de la descripción de las moléculas por sus propiedades químico-físicas o mediante técnicas grafo-teóricas para establecer

modelos predictivos o explicativos de la acción de las mismas. Otro enfoque indirecto consiste en las búsquedas de similitud estructural entre los compuestos de interés y compuestos conocidos por su determinada actividad. Resulta también de amplio empleo, el proceso de búsqueda de potenciales grupos farmacofóricos (27).

Casi desde los mismos orígenes de los estudios QSAR, ha sido una preocupación para los investigadores, la gran cantidad de datos a analizar y cómo adaptar las técnicas matemáticas en la ayuda a la toma de decisiones. En 1973, Hansch (28) planteó el empleo del análisis clúster para la selección de sustituyentes en la estrategia del diseño de fármacos. El aplicó el método de *clustering* jerárquico para la selección de sustituyentes a emplear en las modificaciones químico-estructurales a realizar. Un planteamiento similar hizo Cramer en 1973 cuando propuso su método de análisis subestructural (29).

1.2.- Grafo químico.

Una forma de representación de las moléculas es como grafos químicos. Los especialistas de esta rama afirman que las moléculas están estrechamente relacionadas con los tipos de grafos tratados por los matemáticos en la teoría de grafos. Los grafos químicos proporcionan una metáfora potente e intuitiva para la comprensión de muchos aspectos de esta ciencia; sin embargo, tienen sus limitaciones, especialmente cuando se trata de cuestiones en el campo de la Quimiometría o la Quimioinformática (30).

La teoría de grafos químicos, es una potente herramienta matemática que en el campo de la Bioinformática y la Quimioinformática, permite la asociación de las estructuras químicas a los grafos matemáticos y las matrices de adyacencia o distancia correspondientes. A partir de las mismas, se pueden definir a su vez diversos descriptores para correlacionar la estructura química con diferentes propiedades químico-físicas o biológicas de las moléculas. En la teoría del grafo químico, una molécula se define generalmente como un grafo conexo y no dirigido, en el que los nodos son los átomos y las aristas son los enlaces entre estos (31).

1.3.- Medidas de similitud y de distancia.

La determinación de la similitud no es más que la estimación de un valor (medida de similitud), que caracteriza el grado de concordancia, de asociación, proximidad, alineamiento, porcentaje de identidad o semejanza entre pares de objetos, los cuales son descritos por diferentes rasgos. De manera

específica para las moléculas, son los descriptores estructurales o de propiedades moleculares los que se utilizan en la comparación entre las entidades químicas (30).

Por lo general los valores de similitud que se obtienen en el proceso de comparación de dos objetos moleculares se encuentran en el intervalo [0,1], además, a la similitud se le considera frecuentemente como una propiedad simétrica: “A” es tan similar a “B” como “B” a “A”, aunque se ha argumentado que ciertas similitudes son inherentemente asimétricas (30).

Las medidas de proximidad, similitud o semejanza, en el rango de valores de [0,1] muestran numéricamente cuán parecidos son dos objetos, por lo que el valor 1 será la similitud máxima entre pares de objetos y 0 el mínimo; sin embrago, también existen medidas de distancia o disimilitudes que operan de manera contraria a las medidas de similitud, donde dos objetos son similares si el coeficiente de distancia que los compara está próximo a 0 (30).

Por otra parte, cuando se aplican los conceptos de similitud y diversidad en química, es necesario definir similitudes globales y locales; las similitudes locales se centran en parte en un objeto (átomo, grupo funcional, las cadenas de proteínas, cadena de ADN, etc.), mientras que, para las similitudes globales, la semejanza se mide entre dos objetos enteros (moléculas, proteínas, etc.) (30).

Otro aspecto a tener en cuenta es la definición de *métrica*. Para que un determinado coeficiente sea considerado *métrica* debe satisfacer las siguientes condiciones: i) sus valores deben ser cero o positivos y la distancia de un objeto consigo mismo tiene que ser cero, ii) tiene que ser simétrico, iii) debe cumplir la desigualdad triangular y iv) la distancia entre dos objetos no idénticos tiene que ser superior a cero. Se denominan coeficientes pseudométricos a aquellos que presentan tres de estas propiedades, y coeficientes no-métricos a aquellos que no cumplen la tercera propiedad (25).

Existen varias medidas de proximidad usadas en la búsqueda de similitud. Las mismas han sido clasificadas acorde a su definición: coeficientes de distancias, de asociación y de correlación (30), entre las más usadas se encuentran la distancia Minkowski de la cual se derivan la distancia euclidiana y Manhattan, distancia de Mahalanobis, distancia euclidiana ponderada; aunque cuando se habla de medidas de similitud/distancia entre objetos moleculares es necesario hablar de búsqueda de similitud molecular, que 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. En el estudio de las diferentes funciones de similitud / diferencia presentes en la bibliografía (32), resaltan las funciones Sorensen, Tanimoto, Soergel, Czekanowski, Jaccard, Ruzicka, Dice-Sorensen, como aquellas que mejor se adaptan al cálculo de similitud molecular.

Uno de los métodos para la identificación de máxima similitud entre grafos químicos es la *Propiedad Máxima Común* desarrollada por Antelo y cols. (22)**.** Esta se define como: dados los grafos G_1 y G_2, se entiende por fragmentos con *Propiedad Máxima Común* (f_1, f_2), a los subgrafos de G_1 y G_2 que presentan la máxima similitud en las propiedades químico-físicas (representadas por los índices $Si,\ \Lambda_{3D}\ y\ \Re_{3D}$) entre sus centros descriptores (CD) y la distancia euclidiana entre sus centros de masa.

1.4.- Métodos de *Clustering*

El *clustering* es un proceso que divide un grupo de objetos en grupos o clústeres de objetos, de manera que éstos muestran un alto grado de similitud intra-clúster y de diferencia o disimilitud inter-clúster (25). De este modo, al seleccionar un compuesto perteneciente a cada clúster se obtiene una muestra representativa de todo el conjunto. Para medir la similitud entre objetos se suelen utilizar diferentes formas de distancia ya citadas en el epígrafe 1.3. Representar los datos por una serie de clústeres, conlleva la pérdida de información, pero consigue la simplificación de la muestra con lo cual se logra la facilitación del análisis. *Clustering* es una técnica más de *Machine Learning*, en la que el aprendizaje realizado es no supervisado (*unsupervised learning*). Desde un punto de vista práctico, el *clustering* juega un papel muy importante en aplicaciones de *data mining* o minería de datos tales como la exploración de datos científicos, la recuperación de la información y minería de texto; en aplicaciones sobre bases de datos espaciales (tales como GIS o datos procedentes de astronomía), aplicaciones Web, marketing, diagnóstico médico, análisis de ADN en biología computacional, cribado virtual de base datos de compuestos químicos y muchas otras (33).

Existe una gran variedad de algoritmos de *clustering* surgidos en los últimos años, los cuales se pueden clasificar en dos tipos: particionales y jerárquicos.

En los métodos particionales el objetivo principal es lograr, a partir de un conjunto de patrones D = $\{X_1,...,X_n\}$ y una medida de similitud entre patrones, identificar una partición $D_1,...,D_c$ que optimice una cierta función objetivo. Estos métodos permiten la optimización del criterio por búsqueda exhaustiva fuera de consideración, y utilizan métodos iterativos, aunque estos no siempre garanticen la convergencia a óptimos globales. Un aspecto importante es la elección de la función objetivo, porque esta permitirá evaluar la calidad de la partición; entre ellas se pueden señalar el criterio de SSE (*sum of*

squared error), y el criterio de mínima varianza. Entre los principales algoritmos que se clasifican como particionales destacan el clásico *k*-means, *Fuzzy k*-means, DBSCAN (*Clustering* basado en funciones de densidad), EM (33). Los métodos particionales presentan algunos inconvenientes los cuales hay que considerar a la hora de utilizar los algoritmos desarrollados, pues estos son sensibles a la inicialización. Como son algoritmos rápidos, eso se resuelve ejecutando el algoritmo para varias inicializaciones y tomando la mejor solución (valor de función criterio mínimo). Otro problema es que la cantidad de clústeres es una entrada en varios algoritmos, y eso generalmente es difícil de establecer.

Otro grupo de algoritmos de *clustering* son los clasificados como jerárquicos. Estos no necesariamente compiten con los métodos particionales, ya que representan a los datos de manera diferente. Su objetivo principal es crear jerarquía de particiones "anidadas"; cada nivel de la jerarquía es en sí mismo una partición, obtenida por unión de clústeres de la jerarquía inferior. En comparación con los métodos particionales, son más versátiles y pueden lidiar con clústeres de formas variadas, pero son computacionalmente más complejos (típicamente $O(n^2 \log n)$ $vs. O(\mathrm{n})$ para el k-means) (33). Los métodos jerárquicos se pueden clasificar en dos grupos: aglomerativos, en los que se comienza con tantos clústeres como individuos y consiste en ir formando (aglomerando) grupos según su similitud, y los divisivos, en que se comienza con un único clúster y consiste en ir dividiendo clústeres según la disimilitud entre sus componentes (34).

Se conoce que se han desarrollado varias herramientas para el agrupamiento de datos por varios métodos de *clustering* y por diferentes funciones de similitud intra-cluster; entre las más usadas y populares se encuentra *Weka* (35), software para la minería de datos, el cual contiene varios algoritmos de *clustering* implementados y muestra en su interfaz los resultados que permiten al usuario sacar conclusiones importantes de la información obtenida, pero debido a la forma en que esta herramienta muestra sus resultados, su comprensión no se facilita, ya que no permite visualizar las moléculas asociadas a los patrones agrupados, y de esta manera el químico no pude identificar el clúster(es) que contiene los fragmentos representativos a los cuales se le atribuye la actividad farmacológica. Otro software muy usado es *R,* en el cual se ha desarrollado el paquete *WGCNA* (*Weighted Correlation Network Analysis*) (36) que permite aplicar técnicas de *clustering*, pero a pesar de su eficiencia a la hora de aplicar los algoritmos y mostrar resultados beneficiosos, tiene la misma limitación.

En la literatura estudiada también se hace referencia a la aplicación de métodos de *clustering* en el campo de la bioinformática. Muchos de estos métodos son adaptaciones de los ya existentes, como por ejemplo, el algoritmo de *clustering* difuso *c-means*, una variante difusa del algoritmo clásico *k-means*,

empleado para la clasificación de datos de bioinformática (37), otro es *Chameleoclust* (38), un algoritmo evolutivo que logra adaptarse a diferentes conjuntos de datos y sus resultados permiten un análisis detallado de moléculas químicas; otros destacados son los algoritmos jerárquicos *single linkage*, *complete linkage*, *average linkage*; también se hace mención de los métodos *single-pass*, los de *relocation* y los de *nearest-neighbour,* sobre los cuales se han desarrollado otros algoritmos, destacando en la literatura al k-*means* y *Jarvis-Patrick* (25). La gran limitación que tienen los algoritmos desarrollados en esta rama, es el criterio de similitud que siguen para comparar pares de grafos moleculares. A partir de esos criterios se agrupan las moléculas por sus similitudes, sin embargo, pares de moléculas pueden que no sean similares en su totalidad, pero sí lo sean en una parte de ella, y si además se logra confirmar que esa parte común es la responsable de la actividad de su correspondiente molécula, se estaría hablando de dos moléculas con actividad similar y por ende con una misma aplicación farmacéutica.

1.5.- Algoritmos de *clustering* seleccionados.

Los algoritmos seleccionados por el estudio realizado fueron: algoritmo Basado en Grafo y el algoritmo Jarvis Patrick. A continuación, se explican los algoritmos de *clustering* seleccionados que mejor se adaptan al dominio de grafos moleculares y a los descriptores a utilizar (índices de Estado $Si,\ \Lambda_{3D}\ y\ \Re_{3D}$ para átomos).

1.5.1.- Método Basado en Grafo (Algoritmo basado en matriz de similitud).

Uno de los principales inconvenientes en los métodos de *clustering* explicados en secciones anteriores es que muchos trabajan con funciones heurísticas buscando una mejor aproximación a óptimos globales, pero esto los hace dependientes del orden en que se le presentan los patrones. Los métodos basados en grafos, intentan evitar este hecho, pero su costo computacional los hace inaplicables en muchas ocasiones.

La matriz de similitud se emplea para mostrar el grado de similitud entre un conjunto de patrones. Se construye una matriz *S* simétrica de tamaño NxN, siendo N el número de patrones del conjunto de entrenamiento. *S[i,j]* toma el valor 1 si la distancia entre los patrones *i* y *j* queda por debajo de un umbral preestablecido θ. En caso contrario, *S[i,j]* vale 0. Por lo tanto, con un bit por celda podemos almacenar la matriz de similitud (39). A continuación, se muestran los pasos que describen al algoritmo:

Paso 1: Mientras queden patrones en la matriz de similitud S ir al paso 2, sino ir al paso 6**.**

Paso 2: Seleccionar la fila i de la matriz de similitud S que contenga más unos. Si hay varias, escoger una al azar.
Paso 3: Crear un agrupamiento con los patrones j tales que S[i,j] = 1
Paso 4: Añadir al agrupamiento todos aquellos patrones k tales que S[j,k] = 1, siendo j un patrón incluido en el nuevo agrupamiento hasta que no se puedan añadir más patrones a dicho agrupamiento.
Paso 5: Reducir la matriz de similitud: Eliminar de S todas las filas y columnas correspondientes a patrones incluidos en el agrupamiento recién creado e ir al **paso 1**.
Paso 6: Fin del algoritmo

1.5.2.- Algoritmo Jarvis Patrick

Este algoritmo identifica los K compuestos más próximos para cada compuesto N de la base de datos. Una vez que se ha construido esta lista para todos los compuestos, dos fragmentos se agrupan en un clúster si ellas son vecinas recíprocamente y adicionalmente, si comparten en común un número mínimo de vecinos K*min* (s*imilarity threshold*). Este valor de K*min* es el que determina principalmente la partición. El proceso de agrupar los pares se repite hasta que no se identifica un nuevo par a agrupar. Este algoritmo presenta la desventaja de que identifica un gran número de clústeres compuestos de muy pocos fragmentos o *singletons* y también la imposibilidad de especificar a priori el número de clústeres finales requeridos (25).

La condición de vecino está acotada por la métrica que se usa en la similitud/distancia entre pares de objetos moleculares.

Este algoritmo cuenta con dos parámetros fundamentales: 1) número de vecinos para examinar (***Neighbors to Examine***), 2) número mínimo requerido de vecinos en común ***(Neighbors in Common***). El primer parámetro, vecinos para examinar, especifica el número de vecinos de cada elemento a tener en cuenta al contar el número de vecinos mutuos compartidos con otro elemento. Este valor debe ser de al menos 2. Valores bajos provoca que el algoritmo termine rápido y que se generen muchas agrupaciones con pocos elementos, de manera contraria cuando este parámetro toma valores más altos, provocando que el algoritmo tome más tiempo para terminar, pero logrando menos agrupaciones con mayor cantidad de elementos. El segundo parámetro, vecinos en común, especifica el número mínimo de vecinos más cercanos en común que dos elementos deben tener para que puedan estar en el mismo grupo. Este valor debe ser de al menos 1, y no debe exceder el valor del primer parámetro (40).

1.6.- Conclusiones del capítulo.

En este capítulo se presentaron las tendencias actuales en el diseño y obtención de fármacos, se abordó sobre los principales métodos de selección de compuestos, seleccionando los métodos de *clustering* para dar solución al problema planteado. También se profundizó en las diferentes medidas de similitud/distancia que permitirán comparar los objetos moleculares como parte del proceso de agrupamiento, destacando sobre todo el concepto de *Propiedad Máxima Común*. Entre todos los algoritmos de *clustering* estudiados se seleccionaron aquellos que mejor se adaptan al dominio de grafos moleculares y a los descriptores a utilizar (índices de Estado Si, Λ_{3D} y $\Re_{3D}$ para átomos).

Capítulo 2: Materiales y métodos

En este capítulo se exponen las dos variantes de estructura de datos utilizados para que funcionen correctamente los algoritmos de agrupamiento, así como los distintos métodos de rasgos definidos como transformación de los datos moleculares. Igualmente, se muestran las funciones de similitud/distancia empleadas y los algoritmos implementados. También se explican los procedimientos de cálculo de similitud por rasgos y por *Propiedad Máxima Común*, 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 la biblioteca de funciones utilizada. En el desarrollo del capítulo se presentan los algoritmos implementados para dar solución al problema planteado.

2.1.- Preprocesamiento de datos

Este trabajo va enfocado a aplicar métodos de *clustering* al conjunto de fragmentos similares encontrados en investigaciones anteriores. Estos fragmentos se obtienen a partir de búsquedas de subgrafos en colecciones de grafos moleculares utilizando descriptores híbridos. A este conjunto de datos se le aplicó como preprocesamiento una limpieza de datos, eliminando fragmentos repetidos pertenecientes a un mismo grafo molecular que se encontraban en el conjunto de fragmentos.

2.2.- Estructura de datos

Un elemento crucial en las técnicas de *clustering* es la estructura de datos, por lo general los datos pueden ser expresados como se muestran en la Figura 1.

Matriz de datos

- Matriz de dos modos: representa *n* objetos descritos por *p* variables (*n* objetos x *p* variables).
- Entrada de datos clásica, modelo relacional.

$$\begin{pmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{pmatrix}$$

Matriz disimilitud

- Matriz de un modo: representa la proximidad para todos los pares de los *n* objetos.
- $d(i,j)$ representa medida de proximidad entre los objetos *i* y *j*. Deseable para determinados algoritmos.

$$\begin{pmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{pmatrix}$$

Figura 1. Estructura de datos utilizada. Fuente: elaboración propia

En el presente trabajo se utilizaron dos enfoques buscando los mejores resultados, el primero fue adaptar los datos que describen a los fragmentos moleculares al paradigma clásico de estructura de datos antes mencionado, de manera que puedan funcionar en su forma original los algoritmos de *clustering*, y el segundo enfoque fue modificar la forma de representar los datos según la estructura de datos mencionada, para poder usar la *Propiedad Máxima Común* como criterio de similitud molecular.

2.2.1.- Estructura de datos por rasgos

Los grafos químicos o moleculares, son descritos por descriptores moleculares que son 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 (41).

A partir de la fragmentación de un grafo químico se obtienen los fragmentos reducidos, estos fragmentos se denominan centros descriptores (CD), se entiende por fragmento molecular de orden *n* a la combinación entre *n* centros descriptores relacionados entre sí por la distancia euclidiana entre sus respectivos centros de masas. Estos CD no son más que agrupaciones de átomos y poseen propiedades químico-físicas las cuales fueron calculadas a partir del uso de descriptores híbridos tales como los índices de Estado $Si,\ \Lambda_{3D}\ y\ \Re_{3D}$. El valor de estos índices fue calculado a nivel atómico y luego fue calculado el valor total del índice de cada CD para cada índice. Adaptando estos tipos de datos a la estructura de datos clásica de los algoritmos de *clustering*, se definieron tres formas de expresión de rasgos.

2.2.1.1.- Primera forma de expresión de rasgos.

Se calcula el valor total del índice de cada fragmento para cada índice, por lo tanto, el índice total de Estado Si para un fragmento molecular se calcula como la suma de los valores de dicho índice por cada CD que conforma al fragmento, de igual manera se calcula el valor total de los restantes índices. Estos valores topográficos totales van a constituir las *p* variables mencionadas en la sección 2.2, que caracterizan a un fragmento. Por tanto, un fragmento molecular será descrito por 4 variables, las tres primeras variables serán los índices totales topográficos del fragmento, el cuarto rasgo será la distancia euclidiana entre los CD del fragmento molecular.

Esta variante permitirá calcular la similitud entre dos fragmentos de orden n (misma cantidad de CD), utilizando una métrica seleccionada y comparando los vectores $v1$ y $v2$ para obtener la similitud por propiedades químico-físicas y los vectores $v3$ y $v4$ para obtener la similitud por distancia de los fragmentos *i* y *j*.

$$v1 = \left\{\sum_{i=1}^{n} Si_{CD\,i}, \sum_{i=1}^{n} \Re_{3D\,CD\,i}, \sum_{i=1}^{n} \Lambda_{3D\,CD\,i}\right\}$$

$$v2 = \left\{\sum_{j=1}^{n} Si_{CD\,j}, \sum_{j=1}^{n} \Re_{3D\,CD\,j}, \sum_{j=1}^{n} \Lambda_{3D\,CD\,j}\right\}$$

$$v3 = d_E\left(\sum_{i=1}^{n-1} d_{E\,(\mathrm{CD}_i, CD_{i+1})}\right)$$

$$v4 = d_E\left(\sum_{j=1}^{n-1} d_{E\,(\mathrm{CD}_j, CD_{j+1})}\right)$$

2.2.1.2.- Segunda forma de expresión de rasgos.

En esta variante se calculan las propiedades totales de cada CD que conforman a un fragmento, estas propiedades totales se obtienen a partir de la suma de los valores de los tres índices del CD, por lo cual un fragmento de orden *n* será descrito por *n* variables que constituirán las propiedades totales de cada CD, más una variable adicional que representará la distancia euclidiana entre los CD del fragmento molecular.

Esta forma permitirá calcular la similitud entre dos fragmentos de un mismo orden, utilizando una métrica seleccionada y comparando los vectores $v1$ y $v2$ para obtener la similitud por propiedades químico-físicas y los vectores $v3$ y $v4$ (sección 2.2.1.1) para obtener la similitud por distancia de los fragmentos *i* y *j*.

$$v1 = \left\{(Si + \Re_{3D} + \Lambda_{3D})_{CD_{1i}}, (Si + \Re_{3D} + \Lambda_{3D})_{CD_{2i}}, \dots, (Si + \Re_{3D} + \Lambda_{3D})_{CD_{ni}}\right\}$$

$$v2 = \left\{(Si + \Re_{3D} + \Lambda_{3D})_{CD_{1j}}, (Si + \Re_{3D} + \Lambda_{3D})_{CD_{2j}}, \dots, (Si + \Re_{3D} + \Lambda_{3D})_{CD_{nj}}\right\}$$

2.2.1.3.- Tercera forma de expresión de rasgos

En esta tercera forma un fragmento de orden *n* será descrito por *3*n* variables, cada variable constituirá el valor de un índice de un CD determinado en el fragmento, más una variable adicional que representará la distancia euclidiana entre los CD del fragmento molecular, de manera que permitirá calcular la similitud entre dos fragmentos de un mismo orden, utilizando una métrica seleccionada y comparando los

vectores $v1$ y $v2$ para obtener la similitud por propiedades químico-físicas y los vectores $v3$ y $v4$ (sección 2.2.1.1) para obtener la similitud por distancia de los fragmentos *i* y *j*.

$$v1 = \left\{(Si, \Re_{3D}, \Lambda_{3D})_{CD_{1i}}, (Si, \Re_{3D}, \Lambda_{3D})_{CD_{2i}}, \dots, (Si, \Re_{3D}, \Lambda_{3D})_{CD_{ni}}\right\}$$

$$v2 = \left\{(Si, \Re_{3D}, \Lambda_{3D})_{CD_{1j}}, (Si, \Re_{3D}, \Lambda_{3D})_{CD_{2j}}, \dots, (Si, \Re_{3D}, \Lambda_{3D})_{CD_{nj}}\right\}$$

2.2.2.- Estructura de datos utilizando *Propiedad Máxima Común*

Los algoritmos de *clustering* trabajan con una matriz de similitud o de un modo explicado en la sección 2.2, donde *d(i,j)* como bien se explica, muestra la distancia o similitud entre los objetos *i* y *j*. La *Propiedad Máxima Común* permite encontrar los subgrafos máximos comunes entre dos moléculas, pero la misma no devuelve un valor que cuantifique el grado de similitud de los subgrafos obtenidos, esta propiedad se utiliza en esta investigación para confirmar que dos fragmentos moleculares de un mismo orden son similares, ya que, si la *Propiedad Máxima Común* entre dos fragmentos moleculares de un mismo orden son los mismos fragmentos comparados, significa que dichos fragmentos son similares. Esta propiedad fue adaptada para que devuelva valores binarios: 1 para cuando dos fragmentos sean comunes, y 0 para en caso contrario, estos valores permitirán llenar la matriz de similitud, por lo cual si *d(i,j)*=1 significa que los fragmentos *i* y *j* son similares.

2.3.- Funciones de similitud utilizadas

Todos los algoritmos de *clustering* usan métricas para calcular la similitud/distancia entre objetos y estos valores son almacenados en la matriz de similitud/distancia para su posterior análisis a la hora de conformar los clústeres. Como los patrones que se trabajan son objetos moleculares, se seleccionaron aquellas medidas de similitud que mejor se adaptan al cálculo de similitud molecular. En la Tabla 1 se presentan los coeficientes de similitud/distancia encontrados en la bibliografía (32), y que serán empleados en esta investigación. Estas funciones fueron utilizadas para encontrar la colección de fragmentos explicado en la sección 2.1 y es por ello que se decide emplear las mismas medidas de similitud como métricas para el cálculo de similitud intra-cluster, tales medidas han sido probadas y validadas para el cálculo de similitud molecular y su eficiencia estará condicionada por los umbrales a utilizar. En este trabajo se utilizaron los umbrales encontrados en la bibliografía (23).

Tabla 1. Funciones de similitud y distancia

No	NOMBRE	FÓRMULA	INTERVALO
1	Sorensen	$\frac{\sum \lvert a_i - b_i \rvert}{\sum (a_i + b_i)}$	[1, 0]

2	Tanimoto	$\frac{\sum a_i + \sum b_i - 2\sum \min(a_i,\ b_i)}{\sum a_i + \sum b_i - \sum \min(a_i,\ b_i)}$	[1, 0]
3	Soergel	$\frac{\sum \lvert a_i - b_i \rvert}{\sum \max(a_i, b_i)}$	[1, 0]
4	Jaccard	$\frac{\sum a_i * b_i}{\sum a_i^2 + \sum b_i^2 - \sum a_i * b_i}$	[0, 1]
5	Ruzicka	$\frac{\sum \min(a_i, b_i)}{\sum \max(a_i, b_i)}$	[0, 1]
6	Dice-Sorensen	$\frac{2 * \sum a_i * b_i}{\sum a_i^2 + \sum b_i^2}$	[0, 1]

2.4.- Algoritmos implementados

2.4.1.- Cálculo de similitud usando primera forma de expresión de rasgos.

El objetivo de este algoritmo es comparar pares de fragmentos, y devolver un valor que cuantifique el grado de similitud entre los fragmentos comparados usando la función de similitud seleccionada y la primera forma de expresión de rasgos definida en la sección 2.2.1.1. Este algoritmo permitirá llenar la matriz de similitud de los algoritmos de *clustering* utilizados y el mismo puede describirse como un conjunto de pasos sencillos:

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

Segundo Paso: Se calcula el valor total del índice de cada fragmento para cada índice.

Tercer paso: Se construyen los vectores de distancia y de propiedad de cada fragmento, este ultimó estará formado por los índices totales $Si,\ \Lambda_{3D}\ y\ \Re_{3D}$ de cada fragmento.

Cuarto paso: Se comparan los vectores de propiedad usando la métrica seleccionada y se obtiene el valor de similitud por propiedad entre ambos fragmentos.

Quinto paso: Se comparan los vectores de distancia usando la métrica seleccionada y se obtiene el valor de similitud por distancia entre ambos fragmentos, si este valor supera la cota prefijada se devuelve la similitud por propiedad calculada.

Tabla 2. Algoritmo para el cálculo de similitud por rasgo1

```
Algoritmo de Similitud utilizando Rasgos1 (F1,F2,u,m)
F1← fragmento molecular 1
F2← fragmento molecular 2
u←  umbral de similitud
m← medida de similitud seleccionada
      Inicio
     1. V1 ← {vector topográfico por rasgo1 del fragmento F1}
     2. V2 ← {vector topográfico por rasgo1 del fragmento F2}
     3. CSP← CalcularSemejanza(V1,V2,m)
     4. V3 ← {vector de distancia euclideana del fragmento F1}
     5. V4 ← {vector de distancia euclideana del fragmento F2}
     6. CSD← CalcularSemejanza(V3,V4,m)
     7. si CSD ≥ u
     8.    retornar CSP
     9. sino
     10.       retornar -1;
        Fin
```

En la

Tabla **2** se muestra el pseudocódigo del algoritmo para el cálculo de similitud utilizando como descripción de los datos la definición de la primera forma de expresión de rasgos. En las dos primeras líneas se obtienen los vectores topográficos totales de los dos fragmentos a comparar, la conformación de estos vectores se encuentra explicado en la sección 2.2.1.1. Posteriormente en la línea 3 se calcula la similitud entre los vectores de propiedad usando la medida de similitud seleccionada. A continuación en las líneas 4 y 5, se elaboran los vectores de distancia de cada fragmento y se calcula la similitud usando la medida de similitud seleccionada, y si este valor supera el umbral especificado, se devuelve la similitud por propiedad calculada como valor cuantitativo de semejanza entre los fragmentos comparados.

2.4.2.- Cálculo de similitud usando segunda forma de expresión de rasgos

El objetivo de este algoritmo es comparar pares de fragmentos, y devolver un valor que cuantifique el grado de similitud entre los fragmentos comparados usando la función de similitud seleccionada y la segunda forma de expresión de rasgos definida en la sección 2.2.1.2. A diferencia del anterior este algoritmo trabaja un emparejamiento por propiedades entre los CD de los fragmentos en cuestión, buscando establecer los rasgos de propiedades totales en el orden correcto para luego poder aplicar el cálculo de similitud. Este algoritmo permitirá llenar la matriz de similitud de los algoritmos de *clustering* utilizados y el mismo puede describirse como un conjunto de pasos sencillos:

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

Segundo Paso: se construye la matriz de semejanza por propiedades químico-físicas entre los pares de fragmentos en cuestión, donde la i-ésima fila contiene el centro descriptor i-ésimo del primer fragmento molecular, la j-ésima columna representa el j-ésimo centro descriptor del segundo fragmento molecular, y la intercepción fila-columna muestra el índice de similitud entre los centros descriptores de ambos fragmentos, calculado a partir del vector de propiedades del CD, utilizando la función de similitud seleccionada.

Tercer paso: se identifican las celdas que posean el mayor valor de similitud y se obtiene el emparejamiento o *matching* entre CD por propiedad (alineamiento de fragmentos por propiedad).

Cuarto paso: Una vez alineados los fragmentos se calculan las propiedades totales de sus CD y se obtienen los vectores de propiedad y de distancia de cada fragmento.

Quinto paso: Se comparan los vectores de propiedad usando la medida de similitud seleccionada y se obtiene el valor de similitud por propiedad.

Sexto paso: Se comparan los vectores de distancia usando la métrica seleccionada y se obtiene el valor de similitud por distancia entre ambos fragmentos, si este valor supera la cota prefijada se devuelve la similitud por propiedad calculada.

Tabla 3. Algoritmo para el cálculo de similitud por rasgo2.

Algoritmo de Similitud utilizando Rasgos2 (F1,F2,u,m)
F1←fragmento molecular 1
F2←fragmento molecular 2
u← umbral de similitud
m← medida de similitud seleccionada
F11←fragmento molecular 1 alineado
F21←fragmento molecular 2 alineado
Inicio 1. CD1 ←{centros descriptores del fragmento molecular F1} 2. CD2 ←{centros descriptores del fragmento molecular F2} 3. (F11,F21)←**getMatchingCD(CD1,CD2,m)**

```
4. V1←{vector topográfico por rasgo2 del fragmento F11}
5. V2←{vector topográfico por rasgo2 del fragmento F21}
6. CSP←CalcularSemejanza(V1,V2,m)
7. V3 ←{vector de distancia euclideana del fragmento F11}
8. V4 ←{vector de distancia euclideana del fragmento F21}
9. CSD←CalcularSemejanza(V3,V4,m)
10.   si CSD ≥ u
11.      retornar CSP
12.   sino
13.       retornar -1;
   Fin
```

En la Tabla 3 se muestra el pseudocódigo del algoritmo para el cálculo de similitud utilizando como descripción de los datos la definición de la segunda forma de expresión de rasgos. En las dos primeras líneas se obtiene el conjunto de centros descriptores de cada fragmento, posteriormente se aplica un método de *matching* entre los conjuntos de centros descriptores obteniéndose así, una alienación por centros descriptores entre los fragmentos. En la línea 3 se obtienen los fragmentos alineados. Luego en las líneas 4 y 5 se obtienen los vectores topográficos del segundo tipo de rasgo, la conformación de estos vectores se explicó en la sección 2.2.1.2. Posteriormente y de manera similar al algoritmo anterior en la línea 6 se calcula la similitud por propiedad utilizando la medida de similitud seleccionada, y en las líneas 7 y 8 se conforman los vectores de distancia de cada fragmento. Finalmente se calcula la similitud usando la medida de similitud seleccionada, y si este valor supera el umbral especificado, se devuelve la similitud por propiedad calculada como valor cuantitativo de semejanza entre los fragmentos comparados.

2.4.3.- Cálculo de similitud usando tercera forma de expresión de rasgos

El objetivo de este algoritmo es comparar pares de fragmentos, y devolver un valor que cuantifique el grado de similitud entre los fragmentos comparados usando la función de similitud seleccionada y la tercera forma de expresión de rasgos definida en la sección 2.2.1.3. Este algoritmo también trabaja el emparejamiento entre fragmentos buscando la mejor alineación entre sus centros descriptores (CD), una vez que estén relacionados los CD que mejor se asemejan, se crean los vectores de propiedad con los valores de los índices de cada CD y los vectores de distancia para luego efectuar el cálculo de similitud entre pares de fragmentos. Este algoritmo permitirá llenar la matriz de similitud de los algoritmos de clustering utilizados y el mismo puede describirse como un conjunto de pasos sencillos:

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

Segundo Paso: se construye la matriz de semejanza por propiedades químico-físicas entre los pares de fragmentos en cuestión, donde la i-ésima fila contiene el centro descriptor i-ésimo del primer fragmento molecular, la j-ésima columna representa el j-ésimo centro descriptor del segundo fragmento molecular y la intercepción fila-columna muestra el índice de similitud entre CDs de ambos fragmentos, calculado a partir del vector de propiedades del CD, utilizando la función de similitud seleccionada.

Tercer paso: se identifican las celdas que posean el mayor valor de similitud y se obtiene el emparejamiento o *matching* entre CDs por propiedad (alineamiento de fragmentos por propiedad).

Cuarto paso: se conforman los vectores de propiedad de cada fragmento alineado y los vectores de distancia.

Quinto paso: Se comparan los vectores de propiedad usando la medida de similitud seleccionada y se obtiene el valor de similitud por propiedad.

Sexto paso: Se comparan los vectores de distancia usando la medida de similitud seleccionada y se obtiene el valor de similitud por distancia entre ambos fragmentos, si este valor supera la cota prefijada se devuelve la similitud por propiedad calculada.

Tabla 4. Algoritmo para el cálculo de similitud por rasgo3.

Algoritmo de Similitud utilizando Rasgos3 (F1,F2,u,m)
F1←fragmento molecular 1
F2←fragmento molecular 2
u← umbral de similitud
m← medida de similitud seleccionada
F11←fragmento molecular 1 alineado
F21←fragmento molecular 2 alineado

```
 Inicio
1. CD1 ←{centros descriptores del fragmento molecular F1}
2. CD2 ←{centros descriptores del fragmento molecular F2}
3. (F11,F21)←getMatchingCD(CD1,CD2,m)
4. V1←{ vector topográfico por rasgo3 del fragmento F11}
5. V2←{ vector topográfico por rasgo3 del fragmento F21}
6. CSP←CalcularSemejanza(V1,V2,m)
7. V3 ←{vector de distancia euclideana del fragmento F1}
8. V4 ←{vector de distancia euclideana del fragmento F2}
9. CSD←CalcularSemejanza(V3,V4,m)
10.   si CSD ≥ u
11.      retornar CSP
12.   sino
13.       retornar -1;
   Fin
```

En la Tabla 4 se muestra el pseudocódigo del algoritmo para el cálculo de similitud utilizando como descripción de los datos la definición de la tercera forma de expresión de rasgos. En las dos primeras líneas se obtienen el conjunto de centros descriptores (CD) de cada fragmento. Posteriormente se obtienen los fragmentos alineados por los CD de los fragmentos que más comunes son, aplicando el método de *matching* en la línea 3. Luego en las líneas 4 y 5 se obtienen los vectores topográficos del tercer tipo de rasgo, la conformación de estos vectores se explicó en la sección 2.2.1.3. Posteriormente y de manera similar al algoritmo anterior en la línea 6 se calcula la similitud por propiedad utilizando la medida de similitud seleccionada. Luego en las líneas 7 y 8 se conforman los vectores de distancia de cada fragmento y se calcula la similitud usando la medida de similitud seleccionada. Finalmente se compara este valor con el umbral especificado, y de superarlo, se devuelve la similitud por propiedad calculada como valor cuantitativo de semejanza entre los fragmentos comparados.

2.4.4.- Cálculo de similitud utilizando *Propiedad Máxima Común*

El objetivo de este algoritmo es utilizar el concepto de *Propiedad Máxima Común*, para conocer si dos fragmentos moleculares son similares por los descriptores utilizados, este algoritmo devolverá valor 1 en caso de que la *Propiedad Máxima Común* entre los fragmentos comparados sean los mismos fragmentos y 0 en caso contrario. Este algoritmo permitirá llenar con valores binarios la matriz de similitud de los algoritmos de *clustering* a utilizar y el mismo puede describirse como un conjunto de pasos sencillos:

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

Segundo paso: Se aplica el algoritmo de *Propiedad Máxima Común* pasándole como parámetro los fragmentos a comparar y obteniéndose los subgrafos máximos comunes por propiedades químico-físicas.

Tercer paso: Comparar los subgrafos encontrados con sus respectivos fragmentos, si son iguales, entonces se construye los vectores de distancia y se comparan usando la métrica seleccionada, si no son iguales se retorna 0.

Cuarto paso: Una vez que se obtiene el valor de similitud por distancia entre ambos fragmentos, si este supera la cota prefijada se devuelve valor 1 indicando que los fragmentos moleculares son similares, se devuelve 0 en caso contrario.

Tabla 5. Algoritmo para el cálculo de similitud por *Propiedad Máxima Común*.

```
Algoritmo de Similitud utilizando PMC (F1,F2,u,m)
F1←fragmento molecular 1
F2←fragmento molecular 2
u← umbral de similitud
m← medida de similitud seleccionada
       Inicio
      1. CD1← {centros descriptores del fragmento molecular F1}
      2. CD2← {centros descriptores del fragmento molecular F2}
      3. (SF1,SF2)← BuscarFragmentosPQF(CD1, CD2, u)
      4. si SF1==F1 y SF2==F2
      5.         V1←{vector de distancia euclideana del fragmento F1}
      6.         V2←{vector de distancia euclideana del fragmento F2}
      7.          CS←CalcularSemejanza(V1,V2)
      8.          si CS ≥ u
      9.                retornar 1
      10.            fin si
      11.   sino
      12.        retornar 0
         Fin
```

En la tabla 5 se muestra el pseudocódigo del algoritmo para el cálculo de similitud por *Propiedad Máxima Común*. En las dos primeras líneas se obtienen el conjunto de CD de cada fragmento. Luego se aplica *Propiedad Máxima Común* con la ejecución del método que se encuentra en la línea 3 (22), obteniéndose los subgrafos máximos comunes por propiedad entre los fragmentos comparados. A continuación se comparan los subgrafos obtenidos con sus respectivos grafos y si estos son iguales se conforman los vectores de distancia y se calcula la similitud estructural entre los fragmentos. Si este valor de similitud

supera el umbral especificado se devuelve valor 1 como valor binario que confirma que los fragmentos comparados son similares, en caso contrario se retorna valor 0.

2.5.- Algoritmos de *clustering* utilizados

Todos los algoritmos van a partir de una colección de fragmentos moleculares previamente encontrados, los cuales ya presentan algunas características similares, como es el orden del fragmento (cantidad de centros descriptores que conforman al fragmento). Por lo cual la primera adaptación en los algoritmos es que se conforma la matriz de similitud con los fragmentos de un mismo orden (misma cantidad de centros descriptores), realizándose de manera iterativa el proceso de agrupamiento de fragmentos moleculares de igual orden.

2.5.1.- Algoritmo Basado en Grafo

A continuación, se muestra el pseudocódigo del algoritmo de *clustering* Basado en Grafo, los pasos que se siguieron fueron los explicados en la sección 1.5.2.

Tabla 6. Algoritmo de *clustering* Basado en Grafo.

Algoritmo *Clustering* Basado en Grafo (MS)

```
Inicio
  1. Mientras ExistPattern(MS)
  2.       fila_i ← BuscarMaxFila1(MS)
  3.       Cluster ← CreateCluster(fila_i , MS)
  4.       Para cada j en Cluster  hacer
  5.              Para cada k en MS hacer
  6.                       si (MS[j][k]==1) entonces
  7.                          si(NotExist(k,Cluster))
  8.                              AddToCluster(k, Cluster)
  9.                          fin si
  10.                      fin si
  11.             fin Para
  12.      fin Para
  13.      ReducirMatriz(MS,Cluster)
  14.    fin Mientras
Fin
```

En este algoritmo la matriz de similitud explicada en la sección 1.5.1, se construye utilizando una de las dos variantes de estructura de datos explicada en la sección 2.2, Para el caso en que se utilice la estructura de datos por rasgos, la matriz será almacenada con valores binarios, mostrando valor 1 en la intercepción fila *i* - columna *j*, si el valor de similitud calculado utilizando el algoritmo para el cálculo de similitud por rasgo seleccionado, del fragmento i-ésimo y j-ésimos supera el umbral de semejanza; de lo

contrario, el valor almacenado en la celda será 0, indicando que los fragmentos *i* y *j* no son semejantes al no superar el umbral de similitud. En caso de que se utilice la estructura de datos utilizando *Propiedad Máxima Común* para llenar la matriz de similitud, se utiliza el algoritmo explicado en la sección 2.4.4, el cual devolverá valor 1 cuando los fragmentos comparados sean similares y 0 cuando no, tales valores serán almacenados en la matriz de similitud en la celda correspondiente a los patrones *i* y *j* comparados.

2.5.2.- Algoritmo Jarvis Patrick

A continuación, se muestra el pseudocódigo del algoritmo de agrupamiento Jarvis Patrick

Tabla 7. Algoritmo de agrupamiento Jarvis Patrick.

```
Algoritmo Clustering utilizando JarvisPatrick (fragmentos, MS, umbral)
umbral← cota mínima de vecinos en común como criterio de agrupamiento
   Inicio
     1. Para cada i en fragmentos hacer
     2.     icluster←getCluster(i)
     3.     ivecinos ← { lista de vecinos más cercanos a i }
     4.     Para cada j > i en fragmentos hacer
     5.          jcluster ← getCluster(j)
     6.          jvecinos ← { lista de vecinos más cercanos a j }
     7.          si SonVecinos(i,j)
     8.               si CompartenMinVecinos(ivecinos,jvecinos,umbral)
     9.                    si (icluster>=0 && jcluster>=0)
     10.                        icluster=mergeCluster(icluster,jcluster)
     11.                   sino  si icluster>=0
     12.                         assigntoCluster(j, icluster)
     13.                         jcluster ← icluster
     14.                    sino si jcluster >= 0
     15.                         assigntoCluster(i, jcluster)
     16.                         icluster ← jcluster
     17.                    sino
     18.                         icluster ← assignToNextCluster(i)
     19.                         assigntoCluster(j, icluster)
     20.                         jcluster ← icluster
     21.                   fin si
     22.              fin si
     23.         fin si
     24.       fin para
     25.   fin para
       Fin
```

En este algoritmo también se hace uso de una matriz de similitud con valores binarios que se construye de manera idéntica a la explicada en la sección anterior, una de las adaptaciones que se le hizo a este algoritmo es que no se tiene en cuenta el parámetro de vecinos a examinar cuando se explicó en la sección 1.5.2. Cuando se comparan dos fragmentos para decidir si agruparlos o no, se examinan todos los vecinos de los fragmentos que se están comparado, buscando el número de vecinos en común, y

así comprobar si superan el umbral mínimo de vecinos en común que deben de tener los fragmentos para que se encuentren en un mismo grupo. Una vez finalizado este algoritmo aquellos fragmentos que no se agruparon en ningún clúster, se consideran agrupaciones individuales.

2.6.- Pasos a seguir para desarrollar la investigación

Para encontrar los fragmentos representativos en una colección de grafos moleculares se siguieron una serie de pasos:

Paso 1: se determina si existe diferencia significativa en los resultados obtenidos por los dos métodos de *clustering* aplicando la prueba estadística no paramétrica de Wilcoxon, en caso de que exista diferencia entre los dos métodos se selecciona el que ofrezca mejores resultados, de lo contrario se selecciona cualquiera de los dos métodos.

Paso 2: se verifica si los resultados de los algoritmos para el cálculo de similitud se comportan significativamente diferentes a través de la aplicación de la prueba estadística no paramétrica de Friedman. En caso de que no exista diferencia significativa entre los algoritmos de similitud se utiliza cualquiera de estos, en caso de que exista diferencia significativa se aplica la prueba de Wilcoxon, para determinar el algoritmo que mejor se comporta.

Paso 3: una vez seleccionado el método de *clustering* y el algoritmo para el cálculo de similitud, se obtienen los clústeres representativos a partir de la proporción que se defina entre la cantidad de fragmentos que pertenecen a moléculas activas y la cantidad de fragmentos que pertenecen a moléculas inactivas agrupados en un mismo clúster.

Paso 4: se valida una muestra de los clústeres representativos obtenidos aplicando la prueba de correlación y la prueba de distancia entre los centroides (fragmento que presenta las propiedades químico-físicas medias de todos los fragmentos agrupados en un clúster) de los clústeres seleccionados.

Paso 5: se obtienen los fragmentos representativos de una colección de grafos moleculares.

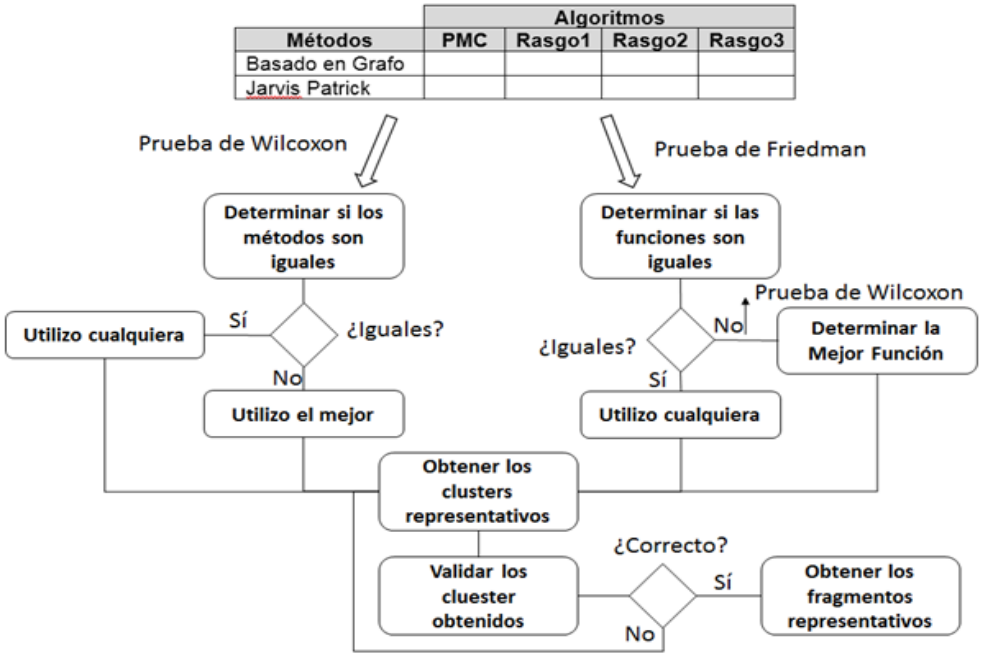


Figura 2. Pasos a seguir para obtener los fragmentos representativos. Fuente: elaboración propia

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.

2.8.- Entorno de Desarrollo Integrado: Eclipse

Se utilizó como entorno de desarrollo el Eclipse, plataforma extensible, basada en Java y liberada bajo Licencia Publica Eclipse (EPL). La misma es una potente herramienta universal de entorno de desarrollo de software hecha en Java y lo usa como lenguaje de programación principal, aunque permite plugins para varios lenguajes. Eclipse fue desarrollado inicialmente por Alphaworks, laboratorio de desarrollo de IBM y actualmente es desarrollado por la Fundación Eclipse (42)**,** 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.

2.9.- Biblioteca utilizada

A continuación, se muestra la biblioteca utilizada en el desarrollo práctico de la investigación. La misma está relacionada con el manejo visual 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 (43)**.**

2.10.- Conclusiones del capítulo

En este capítulo se expusieron las dos variantes de estructura de datos utilizada, y se definieron los diferentes tipos de rasgos como transformación de los datos moleculares. También se mostraron las funciones de similitud/distancia seleccionadas, dentro de la gran variedad de funciones presentes en la bibliografía. Se hizo uso de la *Propiedad Máxima Común* como método de búsqueda de similitud molecular. Se presentaron además los algoritmos implementados para el cálculo de similitud entre fragmentos moleculares de un mismo orden y los algoritmos de *clustering*. Finalmente se presentaron las características principales que motivaron a la selección del lenguaje de programación y el entorno de desarrollo, así como la biblioteca a utilizar en la implementación de los algoritmos.

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 *clustering*, y los algoritmos para el cálculo de similitud desarrollados. Se muestran los cálculos realizados para definir el método de *clustering* y las funciones de similitud que se utilizaran en la investigación. Se presentan los resultados alcanzados mediante el empleo de los métodos de *clustering* utilizando descriptores híbridos ponderados por propiedades químico-físicas, validando la utilización de estos, independientemente de las diferencias estructurales que se detecten. Finalmente se exhiben los fragmentos representativos del ensayo AID941 como resultado de aplicar el método Basado en Grafo con el empleo del algoritmo para el cálculo de similitud por *Propiedad Máxima Común* y la función Soergel.

3.1.- Aplicación de los algoritmos de *clustering* con las distintas funciones de similitud.

Un elemento necesario para aplicar los métodos de *clustering* abordados en el epígrafe 2.5 es el empleo de funciones de similitud o de distancia, ya que las mismas constituyen el grado de error permisible que se estará utilizando. En el presente trabajo se utilizaron las funciones de similitud o de distancia con las cotas de mínima similitud o los umbrales de máxima diferencia (distancia) publicadas por Carrasco y cols (23). Otro elemento importante es el conjunto de datos que se emplearan para aplicar los métodos de *clustering* (Basado en Grafo y Jarvis Patrick), se decidió utilizar 3 ensayos pertenecientes a la base de datos NCBI de compuestos evaluados como anticancerígenos (5): AID941, con 330 moléculas de las cuales 166 son activas; AID488886, con 73 moléculas con 67 activas; y AID88906, con un total de 174 moléculas de las cuales 100 son activas.

Con la combinación de las funciones de similitud, los ensayos seleccionados y los cuatros algoritmos implementados para el cálculo de similitud descritos en el epígrafe 2.4, se aplicaron los dos métodos de *clustering* obteniéndose los resultados que se muestran en las tablas de 8 a la 19.

Es importante señalar dos aspectos fundamentales en los resultados que se muestran: uno es el criterio que se tomó para determinar que un clúster fuera representativo, en este caso se utilizó como variable la proporción de moléculas activas, definiendo que la misma fuera tres veces mayor sobre las inactivas; y el segundo aspecto como determinar el por ciento de moléculas activas asociadas a los fragmentos presentes en los clústeres representativos obtenidos, para ello se obtuvieron las moléculas no duplicadas que contienen a los fragmentos agrupados en cado uno de los clústeres y se calculó el por ciento con respecto al total de moléculas activas de cada ensayo.

Tabla 8. Resultados con el algoritmo Basado en Grafo aplicando Dice-Sorensen con umbral 0,998.

Ensayos	Algoritmo	Características			
		Clúster	Clústeres repres.	Moléculas activas / fragm. repres.	% de moléculas activas
AID941	**PMC**	465	149	132	79,51
Cant. de moléculas: 330	**Rasgo 1**	410	124	130	78,31
Cant. activas: 166	**Rasgo 2**	357	118	127	76,50
Cant de fragmentos:1516	**Rasgo 3**	378	126	132	79,51
AID488886	**PMC**	50	40	49	73,13
Cant. de moléculas: 73	**Rasgo 1**	50	36	49	73,13
Cant. activas: 67	**Rasgo 2**	44	35	49	73,13
Cant de fragmentos: 119	**Rasgo 3**	45	38	49	73,13
AID488906	**PMC**	263	82	74	74,00
Cant. de moléculas: 174	**Rasgo 1**	249	73	74	74,00
Cant. activas: 100	**Rasgo 2**	227	71	73	73,00
Cant de fragmentos: 848	**Rasgo 3**	225	73	74	74,00

Tabla 9. Resultados con el algoritmo Jarvis Patrick mutual neighbors =0 aplicando Dice-Sorensen con umbral 0,998.

Ensayos	Algoritmo	Características			
		Clúster	Clústeres repres.	Moléculas activas / fragm. repres.	% de moléculas activas
AID941	**PMC**	438	143	131	78,91
Cant. de moléculas: 330	**Rasgo 1**	388	123	130	78,31
Cant. activas: 166	**Rasgo 2**	335	113	126	75,90
Cant de fragmentos: 1516	**Rasgo 3**	353	122	131	78,91
AID488886	**PMC**	50	40	49	73,13
Cant. de moléculas: 73	**Rasgo 1**	50	36	49	73,13
Cant. activas: 67	**Rasgo 2**	44	35	49	73,13
Cant de fragmentos: 119	**Rasgo 3**	45	38	49	73,13
AID488906	**PMC**	257	82	74	74,00
Cant. de moléculas: 174	**Rasgo 1**	243	73	74	74,00
Cant. activas: 100	**Rasgo 2**	216	69	73	73,00
Cant de fragmentos: 848	**Rasgo 3**	218	72	74	74,00

Tabla 10. Resultados con el algoritmo Basado en Grafo aplicando Jaccard con umbral 0,995.

Ensayos	Algoritmo	Características			
		Clúster	Clústeres repres.	Moléculas activas / fragm. repres.	% de moléculas activas
AID941	**PMC**	544	161	139	83,73
Cant. de moléculas: 330	**Rasgo 1**	434	120	133	80,12
Cant. activas: 166	**Rasgo 2**	404	121	131	78,91
Cant de fragmentos: 1776	**Rasgo 3**	428	136	134	80,72
AID488886	**PMC**	58	47	52	77,61
Cant. de moléculas: 73	**Rasgo 1**	58	41	52	77,61
Cant. activas: 67	**Rasgo 2**	52	43	52	77,61
Cant de fragmentos: 138	**Rasgo 3**	55	47	52	77,61
AID488906	**PMC**	285	85	75	75,00
Cant. de moléculas: 174	**Rasgo 1**	256	76	73	73,00
Cant. activas: 100	**Rasgo 2**	233	68	74	74,00
Cant de fragmentos: 918	**Rasgo 3**	235	70	74	74,00

Tabla 11. Resultados con el algoritmo Jarvis Patrick mutual neighbors =0 aplicando Jaccard con umbral 0,995.

Ensayos	Algoritmo	Características			
		Clúster	Clústeres repres.	Moléculas activas / fragm. repres.	% de moléculas activas
AID941	**PMC**	507	157	138	83,13
Cant. de moléculas: 330	**Rasgo 1**	400	116	134	80,72
Cant. activas: 166	**Rasgo 2**	366	116	131	78,91
Cant de fragmentos: 1776	**Rasgo 3**	398	132	133	80,12
AID488886	**PMC**	58	47	52	77,61
Cant. de moléculas: 73	**Rasgo 1**	58	41	52	77,61
Cant. activas: 67	**Rasgo 2**	52	43	52	77,61
Cant de fragmentos: 138	**Rasgo 3**	54	47	52	77,61
AID488906	**PMC**	276	84	75	75,00
Cant. de moléculas: 174	**Rasgo 1**	250	76	73	73,00
Cant. activas: 100	**Rasgo 2**	222	67	75	75,00
Cant de fragmentos: 918	**Rasgo 3**	224	70	74	74,00

Tabla 12. Resultados con el algoritmo Basado en Grafo aplicando Tanimoto con umbral 0,078.

Ensayos	Algoritmo	Características			
		Clúster	Clústeres repres.	Moléculas activas / fragm. repres.	% de moléculas activas
AID941	**PMC**	528	152	131	78,91
Cant. de moléculas: 330	**Rasgo 1**	675	115	124	74,69
Cant. activas: 166	**Rasgo 2**	749	109	125	75,30
Cant de fragmentos: 1927	**Rasgo 3**	932	123	122	73,49
AID488886	**PMC**	74	63	53	79,10
Cant. de moléculas: 73	**Rasgo 1**	81	52	51	76,11
Cant. activas: 67	**Rasgo 2**	84	54	51	76,11
Cant de fragmentos: 173	**Rasgo 3**	93	55	51	76,11
AID488906	**PMC**	227	78	73	73,00
Cant. de moléculas: 174	**Rasgo 1**	331	62	65	65,00
Cant. activas: 100	**Rasgo 2**	314	65	63	63,00
Cant de fragmentos: 889	**Rasgo 3**	362	59	60	60,00

Tabla 13. Resultados con el algoritmo Jarvis Patrick mutual neighbors =0 aplicando Tanimoto con umbral 0,078.

Ensayos	Algoritmo	Características			
		Clúster	Clústeres repres.	Moléculas activas / fragm. repres.	% de moléculas activas
AID941	**PMC**	495	147	131	78,91
Cant. de moléculas: 330	**Rasgo 1**	641	109	120	72,28
Cant. activas: 166	**Rasgo 2**	714	101	121	72,89
Cant de fragmentos: 1927	**Rasgo 3**	913	120	122	73,49
AID488886	**PMC**	74	63	53	79,10
Cant. de moléculas: 73	**Rasgo 1**	81	52	51	76,11
Cant. activas: 67	**Rasgo 2**	84	54	51	76,11
Cant de fragmentos: 173	**Rasgo 3**	93	55	51	76,11
AID488906	**PMC**	274	78	73	73,00
Cant. de moléculas: 174	**Rasgo 1**	328	61	63	63,00
Cant. activas: 100	**Rasgo 2**	308	63	63	63,00
Cant de fragmentos: 889	**Rasgo 3**	362	59	60	60,00

Tabla 14. Resultados con el algoritmo Basado en Grafo aplicando Ruzicka con umbral 0,910.

Ensayos	Algoritmo	Características			
		Clúster	Clústeres repres.	Moléculas activas / fragm. repres.	% de moléculas activas
AID941	**PMC**	588	158	131	78,91
Cant. de moléculas: 330	**Rasgo 1**	659	103	128	77,10
Cant. activas: 166	**Rasgo 2**	778	102	118	71,08
Cant de fragmentos: 2239	**Rasgo 3**	987	129	124	74,69
AID488886	**PMC**	89	75	55	82,02
Cant. de moléculas: 73	**Rasgo 1**	94	60	53	79,10
Cant. activas: 67	**Rasgo 2**	95	62	53	79,10
Cant de fragmentos: 220	**Rasgo 3**	110	63	53	79,10
AID488906	**PMC**	317	87	80	80,00
Cant. de moléculas: 174	**Rasgo 1**	348	73	77	77,00
Cant. activas: 100	**Rasgo 2**	352	67	75	75,00
Cant de fragmentos: 1032	**Rasgo 3**	408	64	69	69,00

Tabla 15. Resultados con el algoritmo Jarvis Patrick mutual neighbors =0 aplicando Ruzicka con umbral 0,910.

Ensayos	Algoritmo	Características			
		Clúster	Clústeres repres.	Moléculas activas / fragm. repres.	% de moléculas activas
AID941	**PMC**	545	153	131	78,91
Cant. de moléculas: 330	**Rasgo 1**	575	88	118	71,08
Cant. activas: 166	**Rasgo 2**	712	97	118	71,08
Cant de fragmentos: 2239	**Rasgo 3**	951	122	125	75,30
AID488886	**PMC**	89	75	55	82,08
Cant. de moléculas: 73	**Rasgo 1**	94	60	53	79,10
Cant. activas: 67	**Rasgo 2**	95	62	53	79,10
Cant de fragmentos: 220	**Rasgo 3**	110	63	53	79,10
AID488906	**PMC**	310	87	80	80,00
Cant. de moléculas: 174	**Rasgo 1**	341	72	77	77,00
Cant. activas: 100	**Rasgo 2**	343	65	73	73,00
Cant de fragmentos: 1032	**Rasgo 3**	406	64	69	69,00

Tabla 16. Resultados con el algoritmo Basado en Grafo aplicando Sorensen con umbral 0,047.

Ensayos	Algoritmo	Características			
		Clúster	Clústeres repres.	Moléculas activas / fragm. repres.	% de moléculas activas
AID941	**PMC**	695	161	137	82,53
Cant. de moléculas: 330	**Rasgo 1**	1021	111	125	75,30
Cant. activas: 166	**Rasgo 2**	1415	90	106	63,85
Cant de fragmentos: 2963	**Rasgo 3**	1791	126	108	65,06
AID488886	**PMC**	98	86	59	88,05
Cant. de moléculas: 73	**Rasgo 1**	129	50	54	80,59
Cant. activas: 67	**Rasgo 2**	143	48	53	79,10
Cant de fragmentos: 235	**Rasgo 3**	151	48	50	74,62
AID488906	**PMC**	285	85	75	75,00
Cant. de moléculas: 174	**Rasgo 1**	256	76	73	73,00
Cant. activas: 100	**Rasgo 2**	233	68	74	74,00
Cant de fragmentos: 1262	**Rasgo 3**	235	70	74	74,00

Tabla 17. Resultados con el algoritmo Jarvis Patrick mutual neighbors =0 aplicando Sorensen con umbral 0,047.

Ensayos	Algoritmo	Características			
		Clúster	Clústeres repres.	Moléculas activas / fragm. repres.	% de moléculas activas
AID941	**PMC**	548	146	133	80,12
Cant. de moléculas: 330	**Rasgo 1**	881	91	111	66,86
Cant. activas: 166	**Rasgo 2**	1282	77	97	58,43
Cant de fragmentos: 2963	**Rasgo 3**	1726	108	133	65,06
AID488886	**PMC**	98	86	59	88,05
Cant. de moléculas: 73	**Rasgo 1**	128	49	54	80,59
Cant. activas: 67	**Rasgo 2**	143	48	53	79,10
Cant de fragmentos: 235	**Rasgo 3**	151	48	50	74,62
AID488906	**PMC**	344	96	79	79,00
Cant. de moléculas: 174	**Rasgo 1**	485	71	72	72,00
Cant. activas: 100	**Rasgo 2**	564	60	65	65,00
Cant de fragmentos: 1262	**Rasgo 3**	659	68	71	71,00

Tabla 18. Resultados con el algoritmo Basado en Grafo aplicando Soergel con umbral 0,090.

Ensayos	Algoritmo	Características			
		Clúster	Clústeres repres.	Moléculas activas / fragm. repres.	% de moléculas activas
AID941	**PMC**	588	158	131	78,91
Cant. de moléculas: 330	**Rasgo 1**	659	103	128	77,10
Cant. activas: 166	**Rasgo 2**	749	100	120	72,28
Cant de fragmentos: 2239	**Rasgo 3**	977	128	124	74,69
AID488886	**PMC**	89	75	55	82,08
Cant. de moléculas: 73	**Rasgo 1**	94	60	53	79,10
Cant. activas: 67	**Rasgo 2**	95	62	53	79,10
Cant de fragmentos: 220	**Rasgo 3**	110	63	53	79,10
AID488906	**PMC**	317	87	80	80,00
Cant. de moléculas: 174	**Rasgo 1**	348	73	77	77,00
Cant. activas: 100	**Rasgo 2**	342	64	73	73,00
Cant de fragmentos: 1032	**Rasgo 3**	406	64	69	69,00

Tabla 19. Resultados con el algoritmo Jarvis Patrick mutual neighbors =0 aplicando Soergel con umbral 0,090.

Ensayos	Algoritmo	Características			
		Clúster	Clústeres repres.	Moléculas activas / fragm. repres.	% de moléculas activas
AID941	**PMC**	545	153	131	78,91
Cant. de moléculas: 330	**Rasgo 1**	575	88	118	71,08
Cant. activas: 166	**Rasgo 2**	673	93	118	71,08
Cant de fragmentos: 2239	**Rasgo 3**	940	121	125	75,30
AID488886	**PMC**	89	75	55	82,08
Cant. de moléculas: 73	**Rasgo 1**	94	60	53	79,10
Cant. activas: 67	**Rasgo 2**	94	62	54	80,59
Cant de fragmentos: 220	**Rasgo 3**	110	63	53	79,10
AID488906	**PMC**	310	87	80	80,00
Cant. de moléculas: 174	**Rasgo 1**	341	72	77	77,00
Cant. activas: 100	**Rasgo 2**	329	63	72	72,00
Cant de fragmentos: 1032	**Rasgo 3**	404	64	69	69,00

A simple vista los resultados obtenidos muestran dos supuestos: el primero, que el por ciento de moléculas activas asociadas a los fragmentos representativos encontrados por cada método de *clustering* no son diferentes, ya que, se encuentran por encima del 70 en la mayoría de los casos; y el segundo, que el algoritmo del cálculo de similitud por *Propiedad Máxima Común* muestra los mejores

resultados. Pero para probar estos supuestos se deben aplicar pruebas estadísticas que lo afirme, es por ello que en los siguientes epígrafes de este capítulo se abordan un conjunto de pruebas no paramétricas.

3.2.- Selección del método de *clustering* a emplear en la investigación.

Con el objetivo de probar el primer supuesto planteado en el epígrafe anterior, se prepararon los resultados obtenidos en dos variables por cada una de las seis funciones de similitud utilizadas, cada variable representa uno de los dos métodos de *clustering* empleado. Estas variables generan varias tuplas por la combinación de los ensayos utilizados con los algoritmos de similitud y con el valor de la proporción de moléculas activas asociadas a los fragmentos agrupados en los clústeres representativos.

Para comparar si existe diferencia entre las dos variables generadas por cada función de similitud se aplicó la prueba no paramétrica Wilcoxon para dos muestras relacionadas con un nivel de confiabilidad del 95%. La misma plantea como hipótesis nula (H_o): no existen diferencias significativas en los resultados al utilizar los dos métodos de *clustering* y como hipótesis de investigación (H_1): existen diferencias significativas en los resultados al utilizar los dos métodos de *clustering*. Los resultados de la aplicación de las pruebas se muestran en la Tabla 20.

Tabla 20. Resultados de las pruebas de Wilcoxon con confiabilidad 0,95 para comparar los métodos basado en grafo y Jarvis Patrick.

	Soergel		**DiceSorensen**		**Jaccard**	
Ensayos con algoritmos	**Basado en Grafo**	**Jarvis Patrick**	**Basado en Grafo**	**Jarvis Patrick**	**Basado en Grafo**	**Jarvis Patrick**
AID941-PMC	0,7891	0,7891	0,7951	0,7891	0,8373	0,8313
AID941-Rasgo1	0,7710	0,7108	0,7831	0,7831	0,8012	0,8072
AID941- Rasgo2	0,7228	0,7108	0,7650	0,7590	0,7891	0,7891
AID941- Rasgo3	0,7469	0,7530	0,7951	0,7891	0,8072	0,8012
AID488886- PMC	0,8208	0,8208	0,7313	0,7313	0,7761	0,7761
AID488886- Rasgo1	0,7910	0,7910	0,7313	0,7313	0,7761	0,7761
AID488886- Rasgo2	0,7910	0,8059	0,7313	0,7313	0,7761	0,7761
AID488886- Rasgo3	0,7910	0,7910	0,7313	0,7313	0,7761	0,7761
AID488906- PMC	0,8000	0,8000	0,7400	0,7400	0,7500	0,7500
AID488906- Rasgo1	0,7700	0,7700	0,7400	0,7400	0,7300	0,7300
AID488906- Rasgo2	0,7300	0,7200	0,7300	0,7300	0,7400	0,7500
AID488906- Rasgo3	0,6900	0,6900	0,7400	0,7400	0,7400	0,7400
$\bar{X}$	**0,7678**	**0, 7627**	**0,7511**	**0,7496**	**0,7749**	**0,7752**
S	**0,0379**	**0,0442**	**0,0260**	**0,0239**	**0,0314**	**0,0295**
	Wilcoxon Test **Z** = -0,674 **Asymp. Sig. (2-tailed)** = 0,500		**Wilcoxon Test** **Z** = -1,732 **Asymp. Sig. (2-tailed)** = 0,083		**Wilcoxon Test** **Z** = -0,378 **Asymp. Sig. (2-tailed)** = 0,705	
	Tanimoto		**Ruzicka**		**Sorensen**	

Ensayos con algoritmos	Basado en Grafo	Jarvis Patrick	Basado en Grafo	Jarvis Patrick	Basado en Grafo	Jarvis Patrick
AID941-PMC	0,7891	0,7891	0,7891	0,7891	0,8253	0,8012
AID941-Rasgo1	0,7469	0,7228	0,7710	0,7108	0,7530	0,6686
AID941- Rasgo2	0,7530	0,7289	0,7108	0,7108	0,6385	0,5843
AID941- Rasgo3	0,7349	0,7349	0,7469	0,7530	0,6506	0,6506
AID488886- PMC	0,7910	0,7910	0,8202	0,8208	0,8805	0,8805
AID488886- Rasgo1	0,7611	0,7611	0,7910	0,7910	0,8059	0,8059
AID488886- Rasgo2	0,7611	0,7611	0,7910	0,7910	0,7910	0,7910
AID488886- Rasgo3	0,7611	0,7611	0,7910	0,7910	0,7462	0,7462
AID488906- PMC	0,7300	0,7300	0,8000	0,8000	0,7500	0,7900
AID488906- Rasgo1	0,6500	0,6300	0,7700	0,7700	0,7300	0,7200
AID488906- Rasgo2	0,6300	0,6300	0,7500	0,7300	0,7400	0,6500
AID488906- Rasgo3	0,6000	0,6000	0,6900	0,6900	0,7400	0,7100
$\bar{X}$	**0,7256**	**0,7200**	**0,7684**	**0,7622**	**0,7542**	**0,7331**
S	**0,0632**	**0,0645**	**0,0380**	**0,0423**	**0,0675**	**0,0848**
	Wilcoxon Test **Z** = -1,633 **Asymp. Sig. (2-tailed)** = 0,102		**Wilcoxon Test** **Z** = -0,730 **Asymp. Sig. (2-tailed)** = 0,465		**Wilcoxon Test** **Z** = -1,690 **Asymp. Sig. (2-tailed)** = 0,090	

Analizando los resultados de las pruebas de Wilcoxon se puede observar que todos los valores del parámetro ***Asymp. Sig. (2-tailed)*** son mayores que el nivel de significación utilizado en la prueba (0,05) por lo que se acepta la hipótesis nula y se puede afirmar que no existen diferencias significativas en los valores obtenidos al aplicar los métodos de *clustering* (Basado en Grafo y Jarvis Patrick), demostrándose el primer supuesto expresado en el epígrafe 3.1.

3.3.- Selección del algoritmo de similitud a emplear en la investigación.

El paso siguiente es probar el segundo supuesto, para ello es necesario comprobar si existe diferencia en los resultados obtenidos por los cuatro algoritmos de similitud implementados. Primeramente, se debe preparar los datos obtenidos en el epígrafe 3.1 en cuatro variables, las cuales representan cada uno de los algoritmos. Como se explicó en el epígrafe anterior cada variable generan varias tuplas dadas por la combinación de los métodos de *clustering*, los ensayos utilizados y las 6 funciones de similitud empleadas y con el valor de la proporción de moléculas activas asociadas a los fragmentos agrupados en los clústeres representativos.

Seguidamente para determinar si existen diferencias significativas en los resultados obtenidos en las cuatro variables se utiliza la prueba no paramétrica de comparación de k variables relacionadas Friedman con un nivel de confiabilidad de un 95%. Esta prueba plantea como hipótesis nula (H_o): no existen diferencias significativas en los resultados al utilizar los cuatro algoritmos de similitud y como

hipótesis de investigación (H_1): existen diferencias significativas en los resultados al utilizar los cuatro algoritmos de similitud. Los resultados de la aplicación de la prueba se muestran en la Tabla 21.

Tabla 21. Resultados de la prueba de Friedman con confiabilidad 0,95 para comparar las funciones utilizadas.

	Algoritmo			
Método, ensayo y función de similitud	**PMC**	**Rasgo1**	**Rasgo2**	**Rasgo3**
Basado en Grafo - AID941 - Soergel	0,7891	0,7710	0,7228	0,7469
Basado en Grafo - AID488886 - Soergel	0,8208	0,7910	0,7910	0,7910
Basado en Grafo - AID488906 - Soergel	0,8000	0,7700	0,7300	0,6900
Jarvis Patrick - AID941 - Soergel	0,7891	0,7108	0,7108	0,7530
Jarvis Patrick - AID488886 - Soergel	0,8208	0,7910	0,8059	0,7910
Jarvis Patrick - AID488906 - Soergel	0,8000	0,7700	0,7200	0,6900
Basado en Grafo - AID941 - Dice-Sorensen	0,7951	0,7831	0,7650	0,7951
Basado en Grafo - AID488886 - Dice-Sorensen	0,7313	0,7313	0,7313	0,7313
Basado en Grafo - AID488906 - Dice-Sorensen	0,7400	0,7400	0,7300	0,7400
Jarvis Patrick - AID941 - Dice-Sorensen	0,7891	0,7831	0,7590	0,7891
Jarvis Patrick - AID488886 - Dice-Sorensen	0,7313	0,7313	0,7313	0,7313
Jarvis Patrick - AID488906 - Dice-Sorensen	0,7400	0,7400	0,7300	0,7400
Basado en Grafo - AID941- Jaccard	0,8373	0,8012	0,7891	0,8072
Basado en Grafo - AID488886 - Jaccard	0,7761	0,7761	0,7761	0,7761
Basado en Grafo - AID488906 - Jaccard	0,7500	0,7300	0,7400	0,7400
Jarvis Patrick - AID941 - Jaccard	0,8313	0,8072	0,7891	0,8012
Jarvis Patrick - AID488886 - Jaccard	0,7761	0,7761	0,7761	0,7761
Jarvis Patrick - AID488906 - Jaccard	0,7500	0,7300	0,7500	0,7400
Basado en Grafo - AID941 - Tanimoto	0,7891	0,7469	0,7530	0,7349
Basado en Grafo - AID488886 - Tanimoto	0,7910	0,7611	0,7611	0,7611
Basado en Grafo - AID488906 - Tanimoto	0,7300	0,6500	0,6300	0,6000
Jarvis Patrick - AID941 - Tanimoto	0,7891	0,7228	0,7289	0,7349
Jarvis Patrick - AID488886 - Tanimoto	0,7910	0,7611	0,7611	0,7611
Jarvis Patrick - AID488906 - Tanimoto	0,7300	0,6300	0,6300	0,6000
Basado en Grafo - AID941 - Ruzicka	0,7891	0,7710	0,7108	0,7469
Basado en Grafo - AID488886 - Ruzicka	0,8202	0,7910	0,7910	0,7910
Basado en Grafo - AID488906 - Ruzicka	0,8000	0,7700	0,7500	0,6900
Jarvis Patrick - AID941 - Ruzicka	0,7891	0,7108	0,7108	0,7530
Jarvis Patrick - AID488886 - Ruzicka	0,8208	0,7910	0,7910	0,7910
Jarvis Patrick - AID488906 - Ruzicka	0,8000	0,7700	0,7300	0,6900
Basado en Grafo - AID941 - Sorensen	0,8253	0,7530	0,6385	0,6506
Basado en Grafo - AID488886 - Sorensen	0,8805	0,8059	0,7910	0,7462
Basado en Grafo - AID488906 - Sorensen	0,7500	0,7300	0,7400	0,7400
Jarvis Patrick - AID941 - Sorensen	0,8012	0,6686	0,5843	0,6506
Jarvis Patrick - AID488886 - Sorensen	0,8805	0,8059	0,7910	0,7462
Jarvis Patrick - AID488906 - Sorensen	0,7900	0,7200	0,6500	0,7100
$\bar{X}$	**0,7898**	**0,7525**	**0,7358**	**0,7368**
S	**0,0378**	**0,0423**	**0,0526**	**0, 0518**
Rango medio	**3,74**	**2,33**	**1,86**	**2,07**
	Friedman Test			
	Chi-Square = 59,057			

	df = 3 **Asymp. Sig.** = 0,000

Analizando los resultados de la prueba de Friedman se puede observar que el valor del parámetro ***Asymp. Sig. (2-tailed)*** es menor que el nivel de significación utilizado en la prueba (0,05) por lo que se rechaza la hipótesis nula, afirmándose que existen diferencias significativas en los valores obtenidos al aplicar los cuatro algoritmos de similitud (PMC, Rasgo1, Rasgo2 y Rasgo3).

Como los valores obtenidos son diferentes es necesario determinar cuál de los algoritmos de similitud obtiene los mejores resultados, de manera que se utilizaron los datos preparados anteriormente y se comparan las variables Rasgo1 vs Rasgo2 y Rasgo1 vs Rasgo3 para determinar si hay diferencia significativa a partir de la prueba no paramétrica Wilcoxon. Los resultados de la aplicación de las pruebas se muestran en la Tabla 22.

Tabla 22. Resultados de las pruebas de Wilcoxon con confiabilidad 0,95 para comparar las funciones Rasgo1, Rasgo2 y Rasgo3.

	Algoritmo			
Método, ensayo y función de similitud	**Rasgo1**	**Rasgo2**	**Rasgo1**	**Rasgo3**
Basado en Grafo - AID941 - Soergel	0,7710	0,7228	0,7710	0,7469
Basado en Grafo - AID488886 - Soergel	0,7910	0,7910	0,7910	0,7910
Basado en Grafo - AID488906 - Soergel	0,7700	0,7300	0,7700	0,6900
Jarvis Patrick - AID941 - Soergel	0,7108	0,7108	0,7108	0,7530
Jarvis Patrick - AID488886 - Soergel	0,7910	0,8059	0,7910	0,7910
Jarvis Patrick - AID488906 - Soergel	0,7700	0,7200	0,7200	0,6900
Basado en Grafo - AID941 - Dice-Sorensen	0,7831	0,7650	0,7831	0,7951
Basado en Grafo - AID488886 - Dice-Sorensen	0,7313	0,7313	0,7313	0,7313
Basado en Grafo - AID488906 - Dice-Sorensen	0,7400	0,7300	0,7400	0,7400
Jarvis Patrick - AID941 - Dice-Sorensen	0,7831	0,7590	0,7831	0,7891
Jarvis Patrick - AID488886 - Dice-Sorensen	0,7313	0,7313	0,7313	0,7313
Jarvis Patrick - AID488906 - Dice-Sorensen	0,7400	0,7300	0,7400	0,7400
Basado en Grafo - AID941- Jaccard	0,8012	0,7891	0,8012	0,8072
Basado en Grafo - AID488886 - Jaccard	0,7761	0,7761	0,7761	0,7761
Basado en Grafo - AID488906 - Jaccard	0,7300	0,7400	0,7300	0,7400
Jarvis Patrick - AID941 - Jaccard	0,8072	0,7891	0,8072	0,8012
Jarvis Patrick - AID488886 - Jaccard	0,7761	0,7761	0,7761	0,7761
Jarvis Patrick - AID488906 - Jaccard	0,7300	0,7500	0,7300	0,7400
Basado en Grafo - AID941 - Tanimoto	0,7469	0,7530	0,7469	0,7349
Basado en Grafo - AID488886 - Tanimoto	0,7611	0,7611	0,7611	0,7611
Basado en Grafo - AID488906 - Tanimoto	0,6500	0,6300	0,6500	0,6000
Jarvis Patrick - AID941 - Tanimoto	0,7228	0,7289	0,7228	0,7349
Jarvis Patrick - AID488886 - Tanimoto	0,7611	0,7611	0,7611	0,7611
Jarvis Patrick - AID488906 - Tanimoto	0,6300	0,6300	0,6300	0,6000
Basado en Grafo - AID941 - Ruzicka	0,7710	0,7108	0,7710	0,7469
Basado en Grafo - AID488886 - Ruzicka	0,7910	0,7910	0,7910	0,7910

Basado en Grafo - AID488906 - Ruzicka	0,7700	0,7500	0,7700	0,6900
Jarvis Patrick - AID941 - Ruzicka	0,7108	0,7108	0,7108	0,7530
Jarvis Patrick - AID488886 - Ruzicka	0,7910	0,7910	0,7910	0,7910
Jarvis Patrick - AID488906 - Ruzicka	0,7700	0,7300	0,7700	0,6900
Basado en Grafo - AID941 - Sorensen	0,7530	0,6385	0,7530	0,6506
Basado en Grafo - AID488886 - Sorensen	0,8059	0,7910	0,8059	0,7462
Basado en Grafo - AID488906 - Sorensen	0,7300	0,7400	0,7300	0,7400
Jarvis Patrick - AID941 - Sorensen	0,6686	0,5843	0,6686	0,6506
Jarvis Patrick - AID488886 - Sorensen	0,8059	0,7910	0,8059	0,7462
Jarvis Patrick - AID488906 - Sorensen	0,7200	0,6500	0,7200	0,7100
$\bar{X}$	**0,7525**	**0,7358**	**0,7525**	**0,7368**
S	**0,0423**	**0,0526**	**0,0423**	**0, 0518**
Rango medio negativo	**14,72**		**15,33**	
Rango medio positivo	**5,83**		**7,78**	
	Wilcoxon Test **Z** = -3,289 **Asymp. Sig. (2-tailed)** = 0,001		**Wilcoxon Test** **Z** = -2,289 **Asymp. Sig. (2-tailed)** = 0,022	

Como muestran los resultados al aplicar la prueba, el valor del parámetro **Asymp. Sig.(2-tailed)** en ambas comparaciones es menor que el nivel de significancia (0.05) rechazándose H_o en ambas comparaciones, por lo cual se infiere que el algoritmo Rasgo1 se comporta diferente a los algoritmos Rasgo2 y Rasgo3, tomándose Rasgo1 como el algoritmo que mejor se comporta en la actual comparación a partir de los valores de la media. Posteriormente se decide comparar los algoritmos de similitud Rasgo2 vs Rasgo3 y PMC vs Rasgo1 a partir de la prueba no paramétrica Wilcoxon. Los resultados al aplicar las pruebas se muestran en la Tabla 23.

Tabla 23. Resultados de las pruebas de Wilcoxon con confiabilidad 0,95 para comparar las funciones Rango2 con Rango3 y PMC con Rango1.

	Algoritmo			
Método, ensayo y función de similitud	**Rasgo 2**	**Rasgo 3**	**PMC**	**Rasgo 1**
Basado en Grafo - AID941 - Soergel	0,7228	0,7469	0,7891	0,7710
Basado en Grafo - AID488886 - Soergel	0,7910	0,7910	0,8208	0,7910
Basado en Grafo - AID488906 - Soergel	0,7300	0,6900	0,8000	0,7700
Jarvis Patrick - AID941 - Soergel	0,7108	0,7530	0,7891	0,7108
Jarvis Patrick - AID488886 - Soergel	0,8059	0,7910	0,8208	0,7910
Jarvis Patrick - AID488906 - Soergel	0,7200	0,6900	0,8000	0,7700
Basado en Grafo - AID941 - Dice-Sorensen	0,7650	0,7951	0,7951	0,7831
Basado en Grafo - AID488886 - Dice-Sorensen	0,7313	0,7313	0,7313	0,7313
Basado en Grafo - AID488906 - Dice-Sorensen	0,7300	0,7400	0,7400	0,7400
Jarvis Patrick - AID941 - Dice-Sorensen	0,7590	0,7891	0,7891	0,7831
Jarvis Patrick - AID488886 - Dice-Sorensen	0,7313	0,7313	0,7313	0,7313
Jarvis Patrick - AID488906 - Dice-Sorensen	0,7300	0,7400	0,7400	0,7400
Basado en Grafo - AID941- Jaccard	0,7891	0,8072	0,8373	0,8012
Basado en Grafo - AID488886 - Jaccard	0,7761	0,7761	0,7761	0,7761

Basado en Grafo - AID488906 - Jaccard	0,7400	0,7400	0,7500	0,7300
Jarvis Patrick - AID941 - Jaccard	0,7891	0,8012	0,8313	0,8072
Jarvis Patrick - AID488886 - Jaccard	0,7761	0,7761	0,7761	0,7761
Jarvis Patrick - AID488906 - Jaccard	0,7500	0,7400	0,7500	0,7300
Basado en Grafo - AID941 - Tanimoto	0,7530	0,7349	0,7891	0,7469
Basado en Grafo - AID488886 - Tanimoto	0,7611	0,7611	0,7910	0,7611
Basado en Grafo - AID488906 - Tanimoto	0,6300	0,6000	0,7300	0,6500
Jarvis Patrick - AID941 - Tanimoto	0,7289	0,7349	0,7891	0,7228
Jarvis Patrick - AID488886 - Tanimoto	0,7611	0,7611	0,7910	0,7611
Jarvis Patrick - AID488906 - Tanimoto	0,6300	0,6000	0,7300	0,6300
Basado en Grafo - AID941 - Ruzicka	0,7108	0,7469	0,7891	0,7710
Basado en Grafo - AID488886 - Ruzicka	0,7910	0,7910	0,8202	0,7910
Basado en Grafo - AID488906 - Ruzicka	0,7500	0,6900	0,8000	0,7700
Jarvis Patrick - AID941 - Ruzicka	0,7108	0,7530	0,7891	0,7108
Jarvis Patrick - AID488886 - Ruzicka	0,7910	0,7910	0,8208	0,7910
Jarvis Patrick - AID488906 - Ruzicka	0,7300	0,6900	0,8000	0,7700
Basado en Grafo - AID941 - Sorensen	0,6385	0,6506	0,8253	0,7530
Basado en Grafo - AID488886 - Sorensen	0,7910	0,7462	0,8805	0,8059
Basado en Grafo - AID488906 - Sorensen	0,7400	0,7400	0,7500	0,7300
Jarvis Patrick - AID941 - Sorensen	0,5843	0,6506	0,8012	0,6686
Jarvis Patrick - AID488886 - Sorensen	0,7910	0,7462	0,8805	0,8059
Jarvis Patrick - AID488906 - Sorensen	0,6500	0,7100	0,7900	0,7200
$\overline{X}$	**0,7358**	**0,7368**	**0,7898**	**0,7525**
S	**0,0526**	**0, 0518**	**0,0378**	**0,0423**
Rango medio negativo	**14,18**		**15,50**	
Rango medio positivo	**12,07**		**0,00**	
	Wilcoxon Test **Z** = -0,175 **Asymp. Sig. (2-tailed)** = 0,861		**Wilcoxon Test** **Z** = -4,785 **Asymp. Sig. (2-tailed)** = 0,000	

Analizando estos últimos resultados se llega a la conclusión de que los algoritmos Rasgo 2 y Rasgo 3 se comportan de manera similar ya que el parámetro ***Asymp. Sig.(2-tailed)*** en esta comparación es mayor que el nivel de significancia (0.05). También se llega a la conclusión de que los algoritmos PMC y Rasgo1 se comportan de manera diferente debido que el parámetro ***Asymp. Sig.(2-tailed)*** en esta comparación es menor que el nivel de significancia (0.05), por lo que se puede afirmar que el algoritmo PMC para el cálculo de similitud es de todos el que obtiene la mayor cantidad de moléculas activas asociadas a los fragmentos representativos encontrados. Demostrando el segundo supuesto planteado en el epígrafe 3.1.

3.4.- Validación de los clústeres representativos obtenidos.

Los clústeres representativos obtenidos del ensayo AID941 por los dos métodos de *clustering* y con el algoritmo PMC para el cálculo de similitud seleccionado como el mejor en el epígrafe anterior se muestran en la Tabla 24.

Tabla 24. Clústeres obtenidos utilizando el algoritmo PMC en el ensayo AID941.

Métodos	Función	Características			
		Clúster	Clústeres repres.	Moléculas activas / fragm. repres.	% de moléculas activas
Basado en Grafo	**Dice-Sorensen**	465	149	132	79,51
	Jaccard	544	161	139	83,73
	Tanimoto	528	152	131	78,91
	Sorensen	695	161	137	82,57
	Ruzicka	588	158	131	78,91
	Soergel	588	158	131	78,91
Jarvis Patrick	**Dice-Sorensen**	438	143	131	78,91
	Jaccard	507	157	138	83,13
	Tanimoto	495	147	131	78,91
	Sorensen	548	146	133	82,12
	Ruzicka	545	153	131	78,91
	Soergel	545	153	131	78,91

Para validar que los clústeres obtenidos agrupan correctamente los fragmentos de las moléculas del ensayo AID941, primeramente, se seleccionó Soergel por ser una de las funciones de similitud que muestran el menor por ciento de moléculas activas asociadas a los fragmentos representativos, asumiendo el criterio que, si el agrupamiento es correcto con la función de similitud que ofrece menor valor de por ciento de moléculas activas, con el resto de las funciones también será correcto. Los resultados de los clústeres encontrados por orden del fragmento se muestran en la Tabla 25.

Tabla 25. Clústeres encontrados por orden de fragmento.

Orden del fragmento	Cantidad de fragmentos	Cantidad de clúster	Cantidad de clúster representativos	Cantidad de fragmentos represe.
2	989	205	37	111
3	622	166	46	118
4	348	101	34	118
5	172	65	22	54
6	55	27	10	21
7	22	9	4	9
8	10	5	1	2
9	15	7	2	4

10	2	1	1	2
11	2	1	1	2
12	2	1	0	0
Total	**2239**	**588**	**158**	**441**

Luego se analizaron los 37 clústeres representativos de orden 2 y se seleccionaron los 12 clústeres que muestran fragmentos estructuralmente iguales, lo que supone a simple vista que el método de *clustering* utilizado no realiza un agrupamiento adecuado. Para demostrar el supuesto se obtiene la correlación que existe entre los centroides de los 12 clústeres seleccionados, utilizando la prueba estadística tabla de contingencia. Los resultados de las correlaciones se muestran en la Tabla 26.

Tabla 26. Correlación entre los clústeres obtenidos de orden 2.

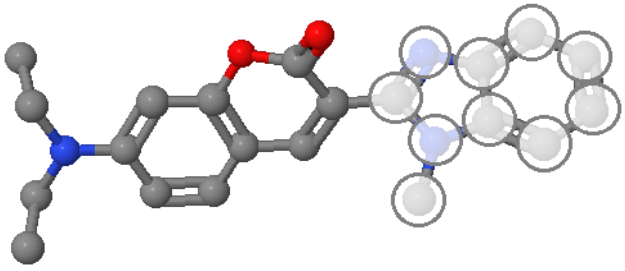
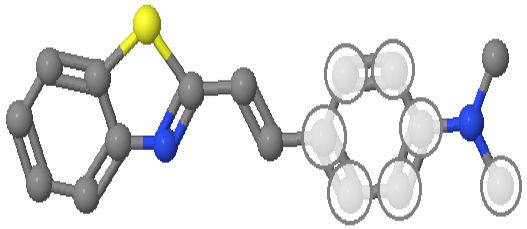
C16 = ANILLO6-HETEROATOMO
C17 = ANILLO9-METILO
C30 = ANILLO6-HETEROATOMO
C32 = ANILLO13-HETEROATOMO
C33 = ANILLO6-METILO
C34 = ANILLO5-HETEROATOMO
C38 = ANILLO6-ANILLO13
C44 = ANILLO6-ANILLO9
C51 = ANILLO6-ANILLO13
C61 = ANILLO6-ANILLO13
C80 = ANILLO9-METILO
C82 = ANILLO9-HETEROATOMO

	C17	**C30**	**C32**	**C33**	**C34**	**C38**	**C44**	**C51**	**C61**	**C80**	**C82**
C16	0,845	0,942	0,952	0,869	0,864	0,313	0,505	0,048	0,189	0,884	0,986
C17		0,915	0,903	**0,998**	0,955	0,679	0,710	0,389	0,507	0,925	0,822
C30			**0,995**	0,932	0,928	0,354	0,449	0,028	0,182	0,941	0,951
C32				0,919	0,911	0,323	0,431	0,002	0,148	0,934	0,965
C33					0,970	0,650	0,691	0,356	0,479	0,945	0,850
C34						0,599	0,649	0,312	0,446	**0,990**	0,858
C38							0,941	0,934	0,964	0,509	0,215
C44								0,884	0,936	0,576	0,389
C51									0,981	0,219	-0,071
C61										0,355	0,066
C80											0,890

Como se observa en los resultados de las correlaciones existe una fuerte correlación entre los clústeres: 17 y 33 (anillo9-metilo y anillo6-metilo), 30 y 32 (anillo6-heteroatomo y anillo13-heteroatomo) y por último entre el 34 y el 80 (anillo5-heteroatomo y anillo9-metilo). Por lo que se sigue suponiendo que el método de *clustering* utilizado no realiza un agrupamiento adecuado, pero si analizamos los valores medios de las propiedades químico-físicas de los clústeres 17 y 33, así como la representación estructural de los fragmentos agrupados en cada clúster que se muestra en la Tabla 27. Se puede notar que existen diferencias entre las propiedades Si y $\Re_{3D}$ del centro descriptor 1 y la propiedad $\Re_{3D}$ del centro descriptor 2, además si analizamos la Figura 3 podemos apreciar que los fragmentos agrupados en los clústeres son estructuralmente diferentes entre los clústeres y semejantes dentro del clúster.

Si analizamos las propiedades mostradas de los clústeres 30 y 32 en el Tabla 28 y los fragmentos agrupados en cada clúster mostrado en la Figura 4. Ocurre algo parecido con los clústeres anteriores, pero en este caso la diferencia radica en las propiedades Si y $\Re_{3D}$ del centro descriptor 1 y los fragmentos agrupados son diferentes estructuralmente entre estos nuevos clústeres y los anteriores pero semejantes al interior de cada clúster.

Tabla 27. Valores de las propiedades de los clústeres 17 y 33.

	CENTRO DESCRIPTOR 1			CENTRO DESCRIPTOR 2			
CLUSTER	Si	$\Re_{3D}$	Λ_{3D}	Si	$\Re_{3D}$	Λ_{3D}	**DISTANCIA**
17	**13,2693**	**28,8139**	2,0318	10,9049	**10,0009**	0,0630	2,5937
33	**10,7942**	**23,3113**	2,1894	8,3566	**7,6738**	-0,0981	3,2562
							

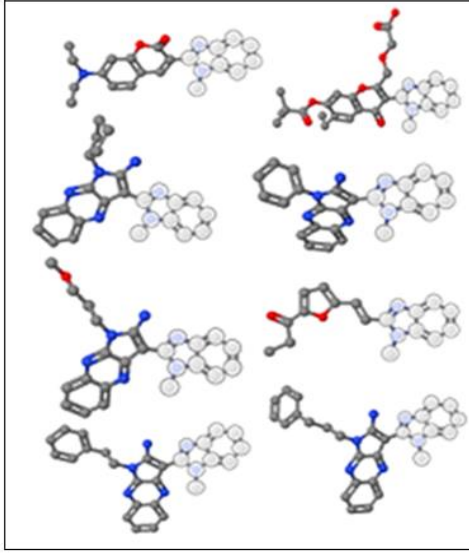
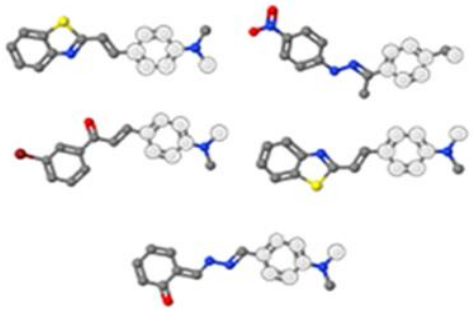
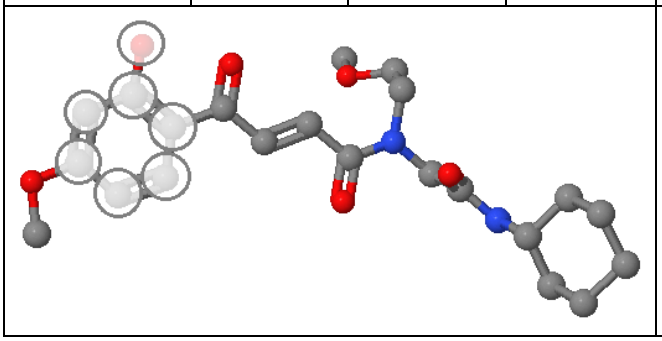
FRAGMENTOS AGRUPADOS EN EL CLÚSTER 17	FRAGMENTOS AGRUPADOS EN EL CLÚSTER 33
	

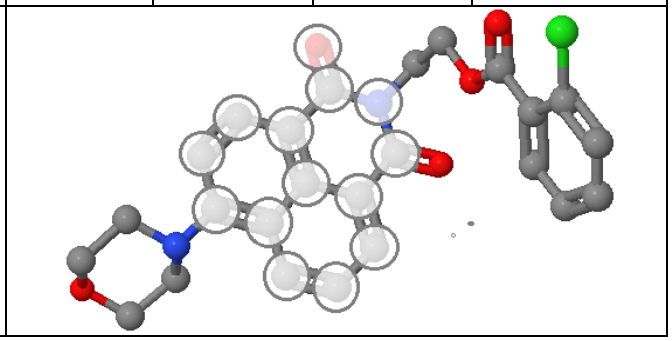
Figura 3. Fragmentos agrupados en los clústeres 17 y 33. Fuente: elaboración propia.

Tabla 28. Valores de las propiedades de los clústeres 30 Y 32.

	CENTRO DESCRIPTOR 1			CENTRO DESCRIPTOR 2			
CLUSTER	Si	$\Re_{3D}$	Λ_{3D}	Si	$\Re_{3D}$	Λ_{3D}	**DISTANCIA**
30	**9,7148**	**30,7791**	3,2483	4,5569	-0,2678	-0,5519	2,1950

32	**19,2088**	**58,5628**	1,5044	5,4996	-3,9915	-0,1258	3,0532





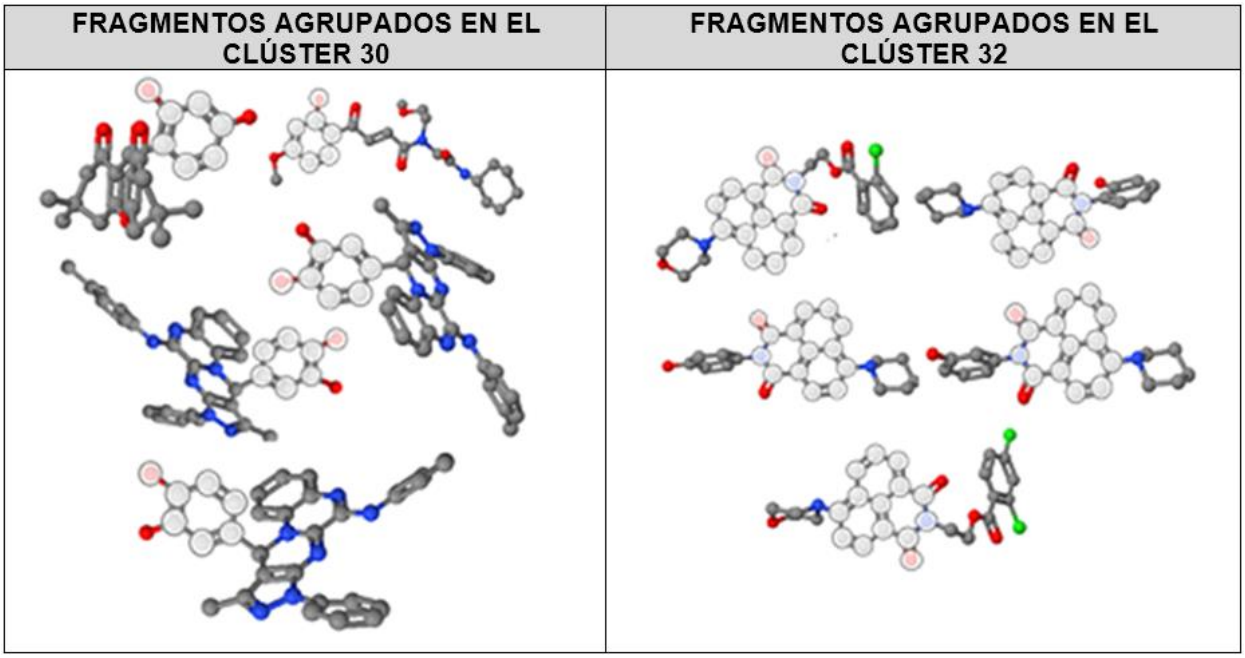
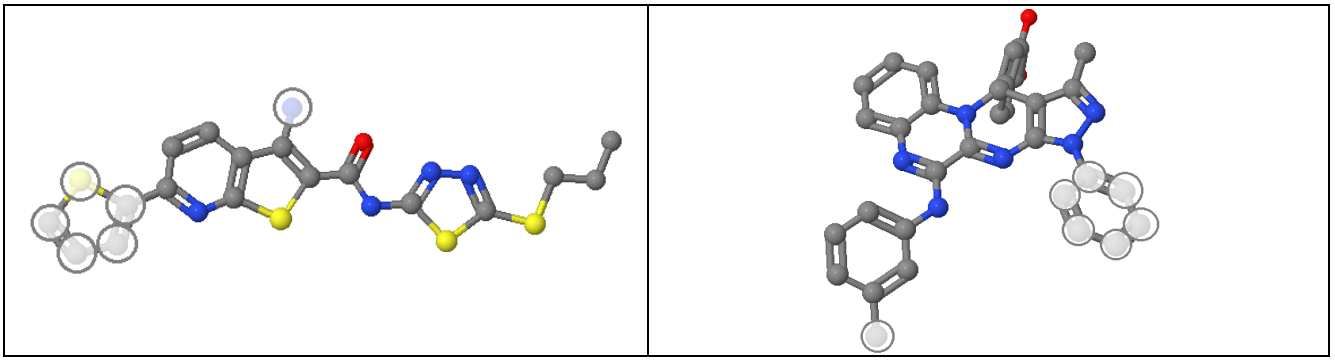


Figura 4. Fragmentos agrupados en los clústeres 30 y 32. Fuente: elaboración propia.

En el caso de los clústeres 34 y 80, las propiedades mostradas en la Tabla 29 demuestran que existen diferencias en la propiedad Si del centro descriptor 1 y la distancia entre ambos centros descriptores. Mientras que la Figura 5 sigue mostrando que los fragmentos agrupados son diferentes estructuralmente entre estos nuevos clústeres y los anteriores pero semejantes al interior de cada clúster.

Tabla 29. Valores de las propiedades de los clústeres 34 Y 80.

	CENTRO DESCRIPTOR 1			**CENTRO DESCRIPTOR 2**			
CLUSTER	Si	$\mathfrak{R}_{3D}$	Λ_{3D}	Si	$\mathfrak{R}_{3D}$	Λ_{3D}	**DISTANCIA**
34	9,2301	**27,6278**	2,2163	8,0946	7,9808	-2,1020	**7,5235**
80	10,5490	**31,5739**	2,7599	8,8241	7,6896	0,4961	**10,8457**



FRAGMENTOS AGRUPADOS EN EL CLÚSTER 34	FRAGMENTOS AGRUPADOS EN EL CLÚSTER 80

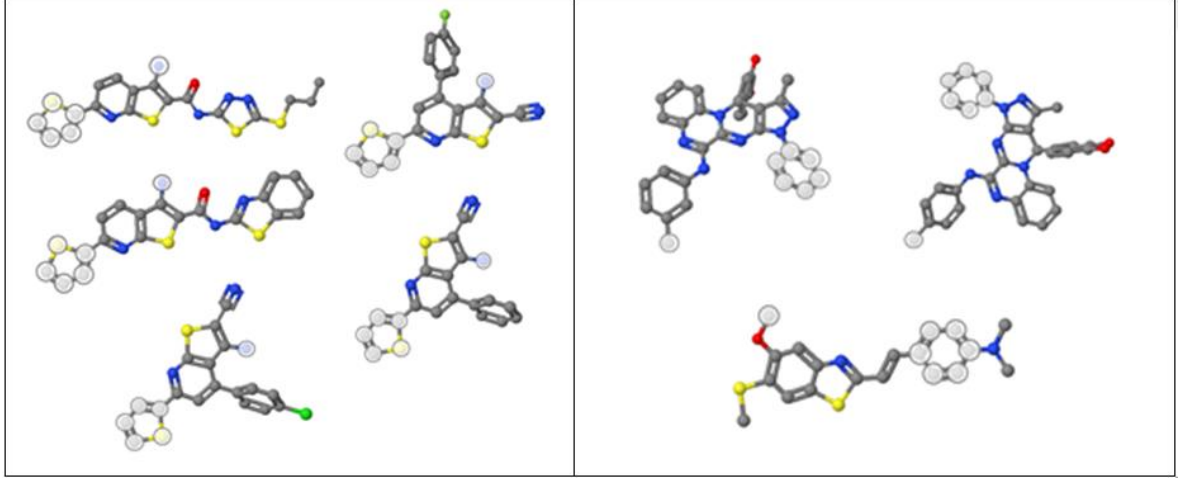


Figura 5. Fragmentos agrupados en los clústeres 34 y 80. Fuente: elaboración propia.

Otro método para validar los resultados del agrupamiento es a través del cálculo de la distancia que existe entre los centroides (39). Las mismas se calcularon utilizando la función de similitud Soergel y los resultados obtenidos se muestran en la Tabla 30.

De los resultados podemos comprobar que todas las distancias entre los centroides de los clústeres son mayores que el umbral (0,09) utilizado en esta función para la aplicación de los métodos de *clustering*, por lo que se demuestra que existe una mayor distancia entre los clústeres obtenidos que la distancia existente entre los fragmentos de cada clúster.

Tabla 30. Distancia entre los centroides de los clústeres utilizando la función de similitud Soergel.

C16 = ANILLO6-HETEROATOMO	C33 = ANILLO6-METILO	C51 = ANILLO6-ANILLO13
C17 = ANILLO9-METILO	C34 = ANILLO5-HETEROATOMO	C61 = ANILLO6-ANILLO13
C30 = ANILLO6-HETEROATOMO	C38 = ANILLO6-ANILLO13	C80 = ANILLO9-METILO

	C32 = ANILLO13-HETEROATOMO				C44 = ANILLO6-ANILLO9			C82 = ANILLO9-HETEROATOMO			
	C17	**C30**	**C32**	**C33**	**C34**	**C38**	**C61**	**C44**	**C51**	**C80**	**C82**
C16	0.441	0.416	0.444	0.531	0.525	0.483	0.494	0.621	0.463	0.614	0.552
C17	0.000	0.541	0.409	0.487	0.487	0.216	0.226	0.427	0.161	0.629	0.702
C30		0.000	0.504	0.357	0.349	0.581	0.555	0.517	0.517	0.456	0.417
C32			0.000	0.596	0.605	0.518	0.452	0.665	0.408	0.678	0.671
C33				0.000	0.197	0.516	0.472	0.341	0.447	0.323	0.435
C34					0.000	0.500	0.482	0.368	0.430	0.345	0.528
C38						0.000	0.199	0.362	0.289	0.633	0.718
C61							0.000	0.333	0.126	0.611	0.692
C44								0.000	0.377	0.461	0.559
C51									0.000	0.604	0.682
C80										0.0000	0.4024

Por todo lo argumentado anteriormente se puede afirmar que el método de *clustering* empleado sí logra agrupar dentro de un mismo clúster los fragmentos similares de acuerdo a sus propiedades químico-físicas y su estructura, por lo que se descarta el supuesto de que a simple vista el método de *clustering* utilizado no realiza un agrupamiento adecuado.

3.5.- Fragmentos representativos obtenidos en el ensayo AID491.

El objetivo general planteado en esta investigación fue identificar fragmentos representativos en una colección de grafos moleculares a partir de métodos de *clustering* utilizando descriptores híbridos. Con los dos métodos de *clustering* y el algoritmo PMC para el cálculo de similitud propuestos se demuestra en el epígrafe anterior, que alcanzan resultados aceptables a la hora de agrupar los fragmentos de las moléculas presentes en un ensayo químico. En otras palabras, permiten identificar los fragmentos representativos en una colección de grafos, de esta forma dándole solución al problema planteado al inicio de la investigación. Se logra obtener 465 fragmentos representativos de los 2239 presentes en las 330 moléculas del ensayo, representando estos fragmentos a 131 moléculas activas de las 166 que presenta el ensayo. En la Figura 6, se observa una muestra de los fragmentos representativos del ensayo AID941.

Otro de los aspectos que queda demostrado en la presente investigación es que se puede identificar fragmentos representativos en una colección de grafos, utilizando descriptores híbridos ponderados por propiedades químico-físicas independiente de su estructura química, siendo esto nuevo campo de investigación en el estudio de similitud molecular.

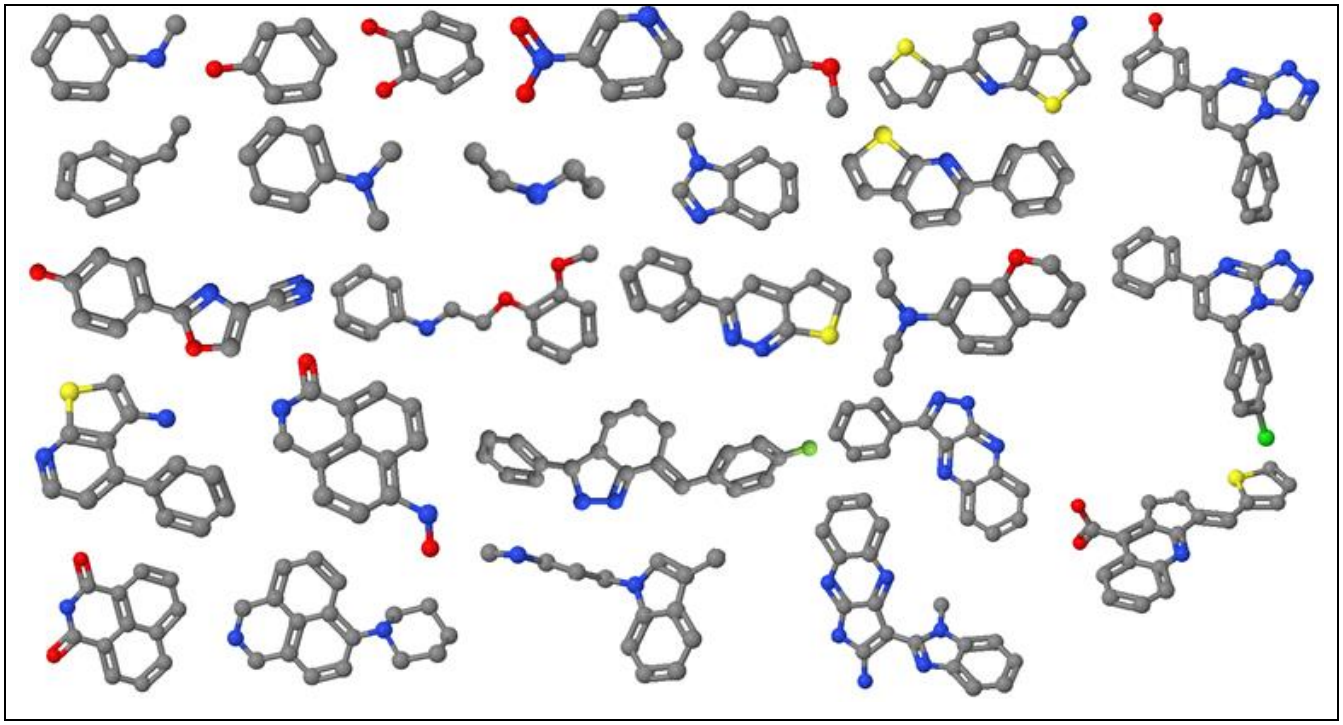


Figura 6. Muestra de fragmentos representativos del ensayo AID941. Fuente: elaboración propia.

3.6.- Conclusiones del capítulo

En este capítulo se muestran los resultados obtenidos al aplicar los dos métodos de *clustering* (Basado en Grafo y Jarvis Patrick) utilizando los diferentes algoritmos para el cálculo de similitud implementados y las diferentes funciones de similitud seleccionadas; afirmándose que se comportan de manera similar. Se selecciona por ofrecer mejores resultados el algoritmo PMC para el cálculo de similitud. Además, se exponen los elementos fundamentales que validan la eficacia entre los clústeres conformados de acuerdo a las propiedades químico–físicas que caracterizan a un fragmento. Finalmente se obtienen los fragmentos representativos del conjunto de fragmentos obtenidos en el ensayo AID941.

Conclusiones

- Se implementaron cuatro algoritmos para el cálculo de similitud entre fragmentos en una colección de grafos utilizando descriptores híbridos ponderados por propiedades químico-físicas, siendo el algoritmo que utiliza la *Propiedad Máxima Común* el que ofrece mejores resultados.

- Se implementaron dos métodos de *clustering* (Basado en Grafo y Jarvis Patrick) para agrupar fragmentos similares de acuerdo a sus propiedades químico-físicas, los cuales obtienen resultados similares.

- Se identificaron 441 fragmentos representativos de grados diferentes de los 2239 presentes en las 330 moléculas del ensayo AID941, utilizando el método de *clustering* Basado en Grafo y el algoritmo PMC con la función de similitud Soergel, los cuales pueden ser los responsables de la actividad farmacológica del 79% de las moléculas activas de dicho ensayo.

Recomendaciones

- Extender el uso de los dos métodos de *clustering* (Basado en Grafo y Jarvis Patrick) y el algoritmo de similitud utilizando *Propiedad Máxima Común* a otras colecciones de grafos moleculares, para continuar con su evaluación.

- Trabajar en la optimización de la cantidad de fragmentos representativos obtenidos con la aplicación de los dos métodos de *clustering* (Basado en Grafo y Jarvis Patrick) y el algoritmo de similitud utilizando *Propiedad Máxima Común*.

- Continuar utilizando el algoritmo de similitud molecular con el concepto de *Propiedad Máxima Común* en otros algoritmos de *clustering* existentes.

Referencias Bibliográficas

1. DEL RÍO, J. *Segunda revolución terapéutica: realidades y proyecciones. La nueva frontera del medicamento.* Madrid : Instituto de Ciencias del Hombre, 1985.

2. BAKKE, Olav M. *Drug safety discontinuations in the United Kingdom, the United States, and Spain from 1974 through 1993: a regulatory perspective.* Clinical Pharmacology & Therapeutics, 1995, vol. 58, no 1, p. 108-117.

3. Orozco López, Fabián. *Diseño Racional de Compuestos Espirotiazolidónicos, Pirazolotiazolidónicos y Pirazolo ß-Lactámicos con Potencial Actividad Antimicrobiana. Tesis Mag, Univ. Nacional de Colombia, Fac. Cien.* Bogotá : s.n., 2014.

4. López Marín, Nelia y Rodríguez Fernández, Rolando. *Uso de bibliotecas químicas virtuales para el diseño de medicamentos.* La Habana : s.n.

5. PubChem. [En línea] National Center for Biotechnology Information. [Citado el: 13 de 6 de 2016.] http://pubchem.ncbi.nlm.nih.gov/.

6. National Cancer Institute. [En línea] [Citado el: 13 de 6 de 2016.] https://resresources.nci.nih.gov/resources/.

7. The Cheminformatics and QSAR Society. [En línea] [Citado el: 13 de 6 de 2016.] http://www.qsar.org/.

8. International Academy of Mathematical Chemistry. [En línea] [Citado el: 13 de 6 de 2016.] http://www.iamc-online.org/.

9. Daylight Chemical Information Systems. [En línea] [Citado el: 13 de 6 de 2016.] http://www.daylight.com.

10. Sunset Molecular Discovery. WOMBAT. . [En línea] [Citado el: 13 de 6 de 2016.] http://sunsetmolecular.com.

11. Barragán, Carlos González y Juaristi, Eusebio. Bases de Datos de Química. Red Latinoamericana de Química. [En línea] Academia Mexicana de Ciencias. [Citado el: 13 de 6 de 2016.] http://www.relaq.mx/RLQ/bases_de_datos.html.

12. Ehrlich, P. *Über den jetzigen Stand der Chemotherapie.* 2006.

13. Escalona, Julio. C., Carrasco, Ramón y Padrón, Juan A. *Introducción al diseño racional de fármacos.* La Habana : Editorial Universitaria, 2008. ISBN 978-959-16-0647-1.

14. Johnson y Maggiora. *Concepts and Applications of Molecular Similarity.* New York : s.n., 1990.

15. Martin, Y. C., Kofron, J. L. y & Traphagen, L. M. *Do structurally similar molecules have similar biological activity?* 2002.

16. 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.

17. Maldonado, A. G., Doucet, J. P. y Petitjean, M. Molecular similarity and diversity in chemoinformatics: From theory to applications. 2006.

18. Han, J., Pei, J., Yin, Y. Mining frequent patterns without candidate generation. In Proceedings of the 2000 ACM-SIGMOD International Conference on Management of Data (SIGMOD'2000), Dallas, TX (2000) 1–12.

19. Agrawal, R., Srikant, R. Fast algorithms for mining association rules. In Proceedings of the 1994 International Conference on Very Large Data Bases (VLDB'94) . Santiago, Chile : s.n., (1994) 487–499.

20. Rivera Borroto, O. M., y otros. Novel similarity measures for the effective and efficient retrieval of pharmacological data sets. 2011.

21. Sheridan, R. P. y & Kearsley, S. K. Why do we need so many chemical similarity search methods? Drug Discov Today. 2002.

22. Paneque Pérez, Juan Luis y Hernández Govea , María Cecilia . *Búsqueda de fragmentos similares pertenecientes a grafos moleculares utilizando índices topográficos e híbridos.* La Habana : Universidad de las Ciencias Informáticas, 2015.

23. *Similitud Molecular Empleando Índices Híbridos.* Antelo Collado, Aurelio, y otros. La Habana : s.n., 2016.

24. *Hybrid reduced graph for SAR studies, SAR and QSAR in Environmental Research.* Carrasco Velar, R., y otros. 2013.

25. Rabal Gracia, Obdulia. Herramientas de cribado virtual aplicadas a inhibidores de tirosina quinasas. Contribución al desarrollo del programa PRALINS para el diseño de quimiotecas combinatorias. (tesis de doctoral). Universitat Ramon Llull. Barcelona : s.n., 2015.

26. Reddy, A. S., y otros. Virtual screening in drug discovery-A computational perspective. 2007.

27. *Acoplamiento inverso y mapeo de farmacóforo como herramientas para encontrar nuevos blancos farmacológicos de compuestos naturales.* Maldonado Rojas, Wilson, Noguera Oviedo, Katia y

Oliveroverbel, Jesús. 140, s.l. : Academia Colombiana de Ciencias Exactas, Físicas y Naturales, 2012, Vol. 36.

28. *Strategy in Drug Design. Cluster Analysis as an Aid in the Selection of Substituents.* Hansch, Corwin. Los Angeles : s.n., 1973.

29. *Substructural Analysis. A Novel Approach to the Problem of Drug Design.* Cramer, Redl, George y Berkoff, Charles E. Pennsylvania : s.n., 1973.

30. Rivera Borroto, Óscar Miguel, y otros. Perspectiva general sobre el proceso de desarrollo de fármacos y las técnicas de cribado virtual basadas en la similitud molecular. [En línea] Laboratorio de Bioinformática, Centro de Estudios de Informática, Facultad de Matemática, Física y Computación, Universidad Central “Marta Abreu” de Las Villas, Santa Clara, 2013. [Citado el: 13 de 6 de 2016.] http://www.anales.ranf.com/2013/vol4/1450.html.

31. Menéndez Velázquez, Amador. *Una breve introduccion a la teoría de grafos.* Oviedo : Universidad de Oviedo, 1998.

32. *Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions.* Sung Hyuk, Cha. s.l. : International Journal of Mathematical Models and Methods in Applied Sciences, 2007, Vol. 1.

33. *Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software.* Garre, Miguel, Cuadrado, Juán José y Sicilia, Miguel Ángel. Madrid : Universidad de Alcalá.

34. Romero-Campero, Francisco J. Búsqueda de patrones: técnicas de clustering. [En línea] 2013. [Citado el: 7 de 6 de 2016.] https://www.cs.us.es/~fran/curso_unia/clustering.html.

35. Bouckaert, Remco R., y otros. *WEKA Manual for Version 3-7-10.* 2013.

36. Langfelder, Peter. WGCNA: an R package for weighted correlation network analysis. [En línea] [Citado el: 1 de 6 de 2016.] https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/.

37. *Using Fuzzy clustering with bioinformatics data.* Gasparovica, Madara, Aleksejeva, Ludmila y Nazaruks, Vladislavs. 2013.

38. *ANALYSIS OF SUBSPACE CLUSTERING OF MOLECULES USING CHAMELEOCLUST, AN EVOLUTIONARY ALGORITHM .* Peignier, Sergio y Castañeta , Heriberto. La Paz : s.n., 2015, Vol. 32.

39. Cartijo Bon, José Francisco. *Técnicas no supervizadas. Metodos de Agrupamiento.* 2001.

40. Improve Outcomes Software. [En línea] [Citado el: 3 de 6 de 2016.] http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/JarvisPatrick_Clustering_Overview.html.

41. *Predicción de actividad anticancerígena de compuestos orgánicos partiendo de descriptores, utilizando programación genética.* Costales Leiva, , Lien , y otros. La Habana : s.n., 2007.

42. The Eclipse Foundation. [En línea] 2004. [Citado el: 3 de 5 de 2016.] https://eclipse.org/.

43. JMOL. [En línea] 2015. [Citado el: 12 de 3 de 2016.] http://jmol.sourceforge.net/.

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 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.
- **Propiedad Máxima Común:** Dados los grafos $G1$ y $G2$, se entiende por fragmentos de Propiedad Máxima Común P ($G1$, $G2$), a los subgrafos máximos comunes 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.
- **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.