UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS

Comparación de estrategias aglomerativas combinatorias tipo Ward usando conjuntos de datos quimioinformáticos y descriptores moleculares reales seleccionados por técnicas de aprendizaje automático

2011

Colectivo de autores

EDITORIAL FEIJÓO

Edición: Liset Ravelo Romero
Corrección: Fernando Gutiérrez Ortega
Diagramación: Roberto Suárez Yera

Editorial Feijóo, 2011

ISBN: 978-959-250-670-1

EDITORIAL
Feijóo

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

Resumen

El trabajo aborda la temática de los algoritmos de conglomerados combinatorios y su aplicación en problemas de la Quimioinformática como son el cribado virtual y las técnicas de selección de compuestos en conjunto de datos farmacológicos. Como tarea primaria e íntimamente ligada a los propósitos del trabajo, se propone una metodología para la selección de descriptores moleculares linealmente relevantes a las actividades biológicas bajo estudio mediante técnicas del Aprendizaje Automático, que garanticen el *principio de comportamiento de vecindad* y, por tanto, el buen rendimiento de los algoritmos a comparar. Posteriormente se proponen doce algoritmos aglomerativos jerárquicos y combinatorios, siete de los cuales son clásicos en la literatura especializada y otros cinco son novedosos en esta área del conocimiento. El objetivo fundamental del trabajo es comparar los nuevos algoritmos con el de elección o de Ward. Los experimentos se realizaron empleando ocho conjuntos de datos bien establecidos y validados en la literatura de la Química Medicinal que incluyen compuestos con usos en el tratamiento de problemas cardiovasculares, de enfermedades mentales, antiinflamatorios, entre otros. El tratamiento estadístico de los resultados permitó a los autores concluir que mediante las técnicas de selección de rasgos empleadas es posible establecer un sistema de meta-aprendizaje en conjuntos supervisados que guíen la posterior selección de descriptores moleculares en sistemas análogos pero no supervisados. Además, permite inferir que el grupo de actividades farmacológicas es modelizable a través de unas pocas familias de descriptores moleculares 3D. Finalmente, y como resultado más alentador, conlleva a argumentar que tres de los algoritmos propuestos se comportan de forma superior a la mayoría de los algoritmos clásicos y similarmente (o ligeramente superior en el caso más optimista) al algoritmo de Unión Completa y al de Ward.

Índice

Introducción

Las compañías farmacéuticas modernas se enfrentan con colecciones de compuestos cada vez más grandes. Las colecciones de compuestos siguen expandiéndose obedeciendo a fusiones, adquisiciones y a la explosión sintética provocada por la química combinatoria. Desde los años noventa del siglo pasado la industria farmacéutica se ha enfocado en la "diversidad" [1]. Las compañías han diversificado sus bases de datos corporativas, ya sea a través de la adquisición de compuestos provenientes de vendedores de estos o a través de síntesis propietarias de librerías combinatorias. En cualquier caso, se hace necesario analizar grandes números de compuestos para chequear su diversidad interna o la diversidad que aportan a los compuestos corporativos existentes. Tales aplicaciones se hacen comúnmente a través de las aplicaciones de conglomerados [2-11].

En los años noventa, el cribado masivo (HTS, de sus siglas en inglés, *High-Throughput Screening*) ha ayudado a mover el cuello de botella del descubrimiento de fármacos lejos de la "arena del cribado" hacia la "arena del análisis". El mero volumen de compuestos líderes y el énfasis en la automatización ha conllevado a la creación de algoritmos de descubrimiento de compuestos líderes que frecuentemente utilizan técnicas de análisis de conglomerados.

Las aplicaciones de estas técnicas a las estructuras químicas requieren representaciones químicas que en la mayoría de los casos toman la forma de cadenas binarias, e.g., MACCs keys [12], *Daylight fingerprints* [13] o BCI keys [14]. Tales representaciones son convenientes para comparaciones de medidas de similitud o métricas, e.g., el coeficiente de Tanimoto o la distancia Euclidiana. Precisamente, los algoritmos de conglomerados han florecido aparejados a la posibilidad de representar y comparar estructuras químicas usando tales medidas. En trabajos relativamente recientes de Brown y Martin [15], y Wild y Blankley [16], se ha mostrado que algunas aplicaciones jerárquicas tales como la de Ward, Unión Completa y Promedio de Grupos se comportan al menos razonablemente, en distintos grados, a la hora de agrupar moléculas químicamente activas en un mismo conglomerado, el método de Ward ha mostrado superioridad de forma general. Sin embargo, estos mismos resultados no son enteramente conclusivos; es todavía incierto cuál combinación de algoritmo de

conglomerados, medida de disimilitud, representación molecular y criterio de poda para determinar el número de conglomerados es la mejor o más robusta con respecto a la dependencia de los datos.

A partir del análisis anterior se puede esbozar como problema científico que en la actualidad no existe consenso en la literatura quimioinformática en cuanto a qué tipo y cantidad de descriptores moleculares se deben calcular, así como cuáles de ellos se deben seleccionar para la aplicación efectiva y eficiente de las técnicas de conglomerados. A pesar de la gran cantidad de algoritmos de conglomerados propuestos, con el de Ward como elección, poca o ninguna atención se le ha prestado a otros algoritmos que por su estrecha relación con este pudieran funcionar similarmente o mejor en problemas característicos. Sin embargo, es posible seleccionar los descriptores más apropiados al contexto quimioinformático estudiado, que al mismo tiempo faciliten el desempeño efectivo y eficiente de los algoritmos de conglomerados en experimentos de comparación. Por tanto, como objetivo global nuestro trabajo se propone desarrollar una metodología de selección de descriptores moleculares apropiados para cada contexto quimioinformático, que permitan la comparación adecuada de algoritmos de conglomerados novedosos con el algoritmo de Ward.

La presente monografía se estructura en tres capítulos. El primero de ellos está dedicado al Marco Teórico donde se brinda información acerca de los algoritmos de conglomerados, relacionada con la conceptualización, clasificación, evolución de las técnicas jerárquicas aglomerativas y aplicaciones en la Quimioinformática. El segundo se dedica a la presentación de los materiales y métodos utilizados a lo largo de la investigación, donde se muestra una descripción teórica de los algoritmos empleados, las bases de datos para la comparación y validación de los algoritmos, los descriptores calculados para los entes moleculares y las estrategias de cálculo molecular y selección de rasgos, finalmente se expone el diseño de experimentos más apropiado para la comparación de los algoritmos. En el tercero y último capítulo se describen los análisis de los resultados alcanzados en las etapas de selección de rasgos, y comparación y validación de los algoritmos.

Capítulo 1. Revisión Bibliográfica

1.1 Elementos del Análisis de Conglomerados

Agrupamiento

El análisis de grupos es descrito como una herramienta para el descubrimiento porque tiene la potencialidad de revelar relaciones basadas en datos complejos no detectadas previamente. Los algoritmos de agrupamiento son usados para encontrar una estructura de grupos que se ajuste al conjunto de datos, logrando homogeneidad dentro de los grupos y heterogeneidad entre ellos. Debe existir un alto grado de asociación entre los objetos de un mismo grupo y un bajo grado entre los miembros de grupos diferentes [17]. Cuando el agrupamiento se basa en la similitud de los objetos, se desea que los objetos que pertenecen al mismo grupo sean tan similares como se pueda y los objetos que pertenecen a grupos diferentes sean tan diferentes como sea posible, en otras palabras, seguir el principio de maximizar la similitud dentro del grupo y minimizar la similitud entre los grupos [17].

El concepto de “similitud” tiene que ser especificado acorde a los datos. En la mayoría de los casos los datos son vectores de valores reales, entonces se requieren algunas medidas (distancias, similitudes, o disimilitudes) para cuantificar el grado de asociación entre estos. La metodología de la aglomeración ha sido desarrollada y utilizada en una amplia variedad de áreas entre las que se pueden mencionar la Química, la Arqueología, la Astronomía, las Ciencias de la Computación, la Electrónica o la Medicina [18].

El proceso de aglomeración, por lo general, se compone de las etapas siguientes:

1. La generación de descriptores apropiados para cada combinación en el juego de datos.
2. La selección de una medida de similitud apropiada.
3. El uso de un método de aglomeración apropiado para el juego de datos.
4. El análisis de los resultados.

Los conglomerados pueden ser superpuestos o no superpuestos; si un elemento aparece en más de un conglomerado, los mismos son superpuestos. Por un extremo, cada elemento es miembro de todos los conglomerados en algún grado; un ejemplo de esto es la aglomeración difusa, en la cual el grado de membresía de un elemento individual está en el rango comprendido entre 0 y 1, y la suma total de dicha membresía a través de todos los conglomerados requiere ser 1. En el otro extremo se presenta la situación donde cada elemento es miembro de exactamente un conglomerado, en cuyo caso se dice que los conglomerados son no solapados. La mayoría de los métodos de aglomeración utilizados en juegos de datos químicos son no solapados, debidos a que el análisis de sus conglomerados resulta relativamente sencillo [18].

Si el juego de datos es analizado de forma iterativa, de tal forma que cada par de conglomerados es fusionado o un conglomerado resulta dividido, el resultado es jerárquico, con una relación padre-hijo, siendo establecida entre los conglomerados a cada nivel sucesivo de la iteración. Si un método de aglomeración jerárquico comienza con todos sus elementos simples, definiendo conglomerados por sí mismos, y más adelante se fusionan de forma iterativa hasta que todos los elementos se incluyan en un solo conglomerado, el método se dice que es aglomerativo [19].

1.2 Algoritmos Jerárquicos Aglomerativos

La familia de los algoritmos de conglomerados secuenciales, aglomerativos, jerárquicos, no superpuestos y combinatorios, o más simplemente en este trabajo, Algoritmos de Conglomerados Aglomerativos y Combinatorios (AC2), se implementan tradicionalmente usando el *algoritmo de la matriz almacenada*, llamado de esa forma debido a que el punto de partida es una matriz que contiene todas las proximidades entre los posibles pares en el juego de datos a ser aglomerado. Cada conglomerado inicialmente corresponde a un elemento individual. A medida que el procedimiento se ejecuta, cada conglomerado puede contener uno o más elementos. En cada iteración, un par de conglomerados es fusionado (aglomerado) y el número total de

conglomerados decrece en uno. El algoritmo de la matriz almacenada se describe como sigue:

1. Calcular la matriz de similitudes inicial, la cual contiene las similitudes entre todos los posibles pares de conglomerados en el juego de datos.
2. Recorrer la matriz para encontrar el par de conglomerados más similar, y fusionarlos en un nuevo conglomerado, reemplazando, por tanto, el par original.
3. Actualizar la matriz de similitudes mediante la desactivación de un conjunto de entradas del par original y actualizando el otro conjunto (el cual representa ahora el nuevo conglomerado) con las similitudes entre el nuevo conglomerado y los conglomerados restantes.
4. Repetir los pasos 2 y 3 hasta que solo quede un conglomerado.

Los distintos métodos AC2 difieren en la forma en la cual es definida la similitud entre los conglomerados, que más adelante será explicado.

Observando el algoritmo descrito, para un juego de datos de *N* componentes, se puede definir que el mismo tiene una complejidad temporal $O(N^2)$ y una complejidad espacial $O(N^2)$ para la creación y el almacenamiento de la matriz de similitudes, mientras que el proceso de aglomeración tiene una complejidad temporal $O(N^3)$. Este algoritmo, por lo tanto, demanda una gran cantidad de recursos para juegos de datos de tamaño superior.

1.3 Agrupamiento Jerárquico para Optimizar una Función Objetivo

En el artículo *"Hierarchical Grouping To Optimize an Objetive Function*" escrito en 1963 por Joe H. Ward [20], se describe un procedimiento para la formación de grupos jerárquicos de subconjuntos mutuamente excluyentes, donde los miembros de cada uno de los mismos son similares de forma maximal con respecto a características específicas. Dados n conjuntos, este procedimiento permite su reducción a *n-1* conjuntos mutuamente excluyentes mediante la consideración de la unión de todos los *n(n–1)/2* posibles pares y seleccionando una unión teniendo un valor máximo para la relación funcional o función objetivo, que refleja el criterio escogido por el investigador.

Formulación del Problema de Agrupamiento

En las situaciones de trabajo a menudo es deseable agrupar un gran número de objetos, símbolos, personas, etc., dentro un número menor de grupos mutuamente excluyentes, donde cada uno de sus miembros resulten ser lo más parecidos entre sí como sea posible. De esta manera el agrupamiento facilita la consideración y el entendimiento de las relaciones entre los individuos en colecciones de gran tamaño, incrementando así la eficiencia de su manejo. El agrupamiento, sin embargo, generalmente conlleva a pérdidas de información que deben ser cuantificadas en un criterio.

Función Objetivo

Sea un conjunto de n individuos con valores $\{x_1, x_2, \ldots, x_n\}$, una práctica común es usar el valor de la media para representar todos los valores, en vez de considerar los valores individuales. La "pérdida" de información de tratar todos los elementos como un solo grupo, con una media dada, puede ser indicada por la suma de cuadrados del error (ESS, de sus siglas en inglés, *Error Sum of Squares*).

En la suma de cuadrados el error está dado por la relación funcional,

$$\mathbf{ESS} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \qquad (\mathbf{1.1})$$

Donde, xi es el valor del puntaje para el *i - ésimo* individuo.

De manera similar si los individuos son clasificados en digamos una cantidad m de grupos, este agrupamiento puede ser evaluado como la suma de las m sumas de errores de la suma de cuadrados,

$$\mathbf{ESS}_{(m\ grupos)} = \mathbf{ESS}_{(Grupo\ 1)} + \mathbf{ESS}_{(Grupo\ 2)} + \ldots + \mathbf{ESS}_{(Grupo\ m)} \qquad (\mathbf{1.2})$$

Una relación funcional que proporciona un criterio de este tipo puede ser considerada como una función objetivo. En general, una función objetivo puede ser *cualquier relación funcional* que el investigador seleccione para reflejar la deseabilidad relativa del agrupamiento. En el caso planteado, la función objetivo es la "pérdida de

información" reflejada a partir del error de la suma de cuadrados. El valor más deseado de esta función objetivo es su mínimo, o sea, 0.0. La naturaleza del problema y el criterio, por supuesto, definen la selección e interpretación de la función objetivo.

Grupos Jerárquicos

El procedimiento de agrupamiento expuesto por Ward está basado en la premisa de que la mayor parte de la información, tal y como es indicada por la función objetivo, está disponible cuando un conjunto de n miembros está desagrupado. Por lo tanto, el procedimiento comienza con estos *n* miembros, denominados grupos, o subconjuntos, aun cuando contengan un solo miembro. El primer paso en este procedimiento es seleccionar dos de estos *n* subconjuntos de tal forma que, al unirlos, será reducido en uno el número de subconjuntos, mientras se obtiene el valor más óptimo posible de la función objetivo. Los *n-1* subconjuntos resultantes son entonces examinados para determinar si es posible la presencia de un tercer miembro que pueda integrarse al primer par seleccionado asegurando la optimalidad de la función objetivo, en este caso ya para *n-2* grupos. Este procedimiento puede continuarse si se desea hasta que todos los *n* miembros originales estén en un único grupo.

Procedimiento

El procedimiento comienza con un conjunto universal (U), $\{s_1, s_2, \ldots, s_n\}$, o sea, *n* conjuntos de solo un elemento. Estos son enumerados arbitrariamente para una conveniente identificación en el procesamiento computacional. Para reducir el número de subconjuntos a *n-1*, un nuevo subconjunto, que minimiza el valor de la función objetivo, es formado a partir de la unión de dos de los *n* subconjuntos originales, dígase:

$$[S(1,n)] \cup [S(2,n)] = \{s_1, s_2\} \qquad (\mathbf{1.3})$$

Esto requiere una evaluación de la función objetivo para cada una de las *n(n – 1)/2* uniones posibles de subconjuntos *S(i, n), i = 1, 2,.., n*, donde *i* se refiere al número de identificación del conjunto, y el parámetro n se refiere a la cantidad de conjuntos en

consideración. Ante cada unión el valor de la función objetivo es calculado y se espera que sea "igual o mejor" que la unión anterior. La identidad de la "mejor" unión, mantenida a lo largo de todo el proceso, facilita la identificación de la unión que tiene un valor de la función objetivo "igual o mejor" que alguna otra de las *n(n – 1)/2* uniones posibles. Esta unión es aceptada como un agrupamiento óptimo cuando el número de subconjuntos es reducido de *n* a *n-1*.

Para el propósito de mostrar la secuencia en la cual los subconjuntos son unidos en estudios de larga escala (*n = 100* a *1000*), resulta conveniente en el procesamiento computacional dar una designación para identificar cada nuevo subconjunto (resultante a partir de la aceptación de una unión) y sus miembros. Por tanto, la unión resultante en *n-1* subconjuntos es denotada:

$$\mathbb{S}(p_{n-1}, n-1) = [\mathbb{S}(p_{n-1}, n)] \cup [\mathbb{S}(q_{n-1}, n)] \qquad (\mathbf{1.4})$$

Donde, p_{n-1} es el menor de los dos números usados para identificar el subconjunto en el conjunto original de n elementos, siendo utilizado para identificar el nuevo subconjunto; mientras que q_{n-1} es el mayor de los dos números utilizados para identificar el subconjunto en el conjunto original de n elementos, pasando este a ser "inactivo" después de ser utilizado en esta fase de la muestra de la secuencia, en la cual dos subconjuntos resultan unidos.

El valor de la función objetivo es identificado de la misma forma para identificarlo con esta unión:

$$Z[p_{n-1}, q_{n-1}, n-1] \qquad (\mathbf{1.5})$$

El término en el extremo derecho *(n-1)*, muestra el número de subconjuntos restantes después de realizada la unión, los otros dos términos contienen los números de identificación originales de los subconjuntos unidos.

Luego, con *n-1* subconjuntos, debemos definir y considerar

$$\mathbb{S}(i, n-1) = \mathbb{S}(i, n) \qquad (\mathbf{1.6})$$

Donde $i = 1, 2, \ldots, n$; $i \neq p_{n-1}$; y $i \neq q_{n-1}$

Y $S(p_{n-1}, n-1) = [S(p_{n-1}, n)] \cup [S(q_{n-1}, n)]$ cuando $i = p_{n-1}$ (**1.7**)

La selección de una unión óptima para reducir *n-1* subconjuntos a *n-2* subconjuntos requiere de la evaluación y comparación de las *(n-1)(n-2)/2* uniones posibles de forma análoga a la usada para reducir n subconjuntos a *n-1*. Cuando esto se logre, la unión aceptada y su valor de la función objetivo asociado es designado

$$S(p_{n-2}, n-2) = [S(p_{n-2}, n-1)] \cup [S(q_{n-2}, n-1)] \quad (\mathbf{1.8})$$

y

$$Z[p_{n-2}, q_{n-2}, n-2] \qquad (p_{n-2} < q_{n-2}) \quad (\mathbf{1.9})$$

Este ciclo puede ser continuado, si se desea, hasta que todos los subconjuntos se encuentren unidos en el conjunto universal, **U**. En cualquier fase donde k subconjuntos mutuamente excluyentes están en consideración, el valor de la función objetivo, y la unión asociada puede ser expresada como:

$Z[i, j, k-1]$ Asociado con $[S(i,k)] \cup [S(j,k)]$ (**1.10**)

Donde, $i = 1, 2, \ldots, n-1$

$$i \neq q_{n-1}, q_{n-2}, \ldots, q_k$$

$$j \neq q_{n-1}, q_{n-2}, \ldots, q_k$$

Continuando la selección de una unión óptima, esta unión y su correspondiente valor de la función objetivo pueden ser designados:

$$S(p_{k-1}, k-1) = [S(p_{k-1}, k)] \cup [S(q_{k-1}, k)] \quad (\mathbf{1.11})$$

Y

$$Z[p_{k-1}, q_{k-1}, k-1] \qquad (p_{k-1} < q_{k-1}) \quad (\mathbf{1.12})$$

Por tanto, los elementos en cualquier subconjunto, S(i, k), pueden ser designados como:

$$S(i,k) = \{a_{m1}, a_{m2}, \ldots, a_{mo}, \ldots, a_{mt}\} \quad (\mathbf{1.13})$$

Donde, t es el número de elementos en el subconjunto y *ma* es el número de identificación del *a-ésimo* elemento en el mismo.

1.4 Teoría General de las Estrategias de Clasificación

En un artículo publicado por [21] llamado *"A general theory of classificatory sorting strategies"* en el año 1967 se describe que el comportamiento computacional de las estrategias de clasificación jerárquica depende de tres propiedades, las cuales quedan establecidas en el documento para 5 estrategias convencionales y cuatro medidas.

Las estrategias convencionales se demuestran como simples variantes de un sistema linear definido por cuatro parámetros. En este trabajo se define también una nueva estrategia, la cual permite variaciones continuas de la intensidad del agrupamiento mediante la variación de un parámetro.

Propiedades Generales

Por consideración general, supóngase que dos grupos (i) y (j) se unen para formar un grupo (k), entonces debemos distinguir entre tres tipos de medida: las (i)-medidas, que definen la propiedad para un grupo, las (i, j)-medidas, que definen una similitud o diferencia entre dos grupos, y las (i, j, k)-medidas, que definen alguna diferencia entre los grupos originales y el grupo formado por su unión. Las propiedades de las estrategias de selección no son invariantes bajo el cambio de medida, y las medidas a ser consideradas deben ser, por lo tanto, declaradas. Se considera que las propiedades generales de las estrategias jerárquicas se clasifican como:

Combinatoria o no combinatoria

Se asumen dos grupos (i) y (j) con n_i y n_j elementos, respectivamente, y como distancia entre los grupos una (i, j)-medida denotada como d_{ij}. Se asume que d_{ij} es la menor medida en el sistema a ser considerada para que, de esta forma, se formen un nuevo grupo (k) con $n\,(=n_i+n_j)$ elementos. Supóngase la matriz con todos los d_{ij} como entradas organizadas en columna, con los valores n_i como una fila adicional, y considérese un tercer grupo (h). Antes de la unión, los valores de d_{ht}, d_{hj}, d_{ij}, n_i y n_j

son todos conocidos y están incluidos en las columnas (i) y (j) de la matriz. Si d_{hk} puede ser calculada a partir de estos cinco valores, una columna (k) puede ser derivada a partir de las columnas (i) y (j) originales; la computadora necesita operar solo en pares de columnas y, como todas la medidas pueden ser calculadas a partir de medidas existentes previamente, los datos originales no necesitan ser almacenados desde el momento en que el primer conjunto de medidas es calculado. Por tanto una medida puede ser clasificada como combinatoria si cumple con la relación lineal:

$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta_{ij} + \delta \left| d_{hi} - d_{hj} \right| \quad (\mathbf{1.14})$$

Donde, los parámetros α_i, α_j, β y γ están determinados por la naturaleza de la estrategia. Nótese que cuando $\gamma = 0$ la cadena de medidas asociada con las uniones jerárquicas sucesivas tendrá una monotonía de la forma:

$$(\alpha_i + \alpha_j + \beta) \geq 0 \quad (\mathbf{1.15})$$

En contraste con lo anteriormente dicho, una estrategia no combinatoria es aquella donde la nueva medida no puede ser calculada a partir de la anterior, por lo cual los datos deben ser almacenados para ser utilizados en el cálculo de medidas requeridas más adelante en el análisis. Vale aclarar que las estrategias combinatorias han manifestado relevantes ventajas computacionales.

Compatible o incompatible

Una estrategia es aquella donde las medidas calculadas avanzado el análisis son exactamente del mismo tipo que las medidas iniciales entre los elementos; o sea, tienen la misma dimensión, están sujetas a las mismas restricciones, y pueden ser ilustradas a partir de un modelo exactamente comparable. Por otra parte, una estrategia incompatible es aquella donde no se cumple alguna de las propiedades mencionadas anteriormente, haciendo la interpretación de la estrategia realmente difícil.

Conservadora o Distorsionadora del Espacio

Las medidas primarias entre los elementos pueden definir un espacio con propiedades conocidas. Cuando los grupos comienzan a formarse no se obtiene que las medidas entre ellos formen un espacio con las propiedades originales. Si logran esto, y el modelo original se mantiene sin cambios, podemos decir que la estrategia es conservadora del espacio.

Por otra parte, existen estrategias en las cuales el modelo se comporta de manera que la vecindad inmediata del espacio se dilata o se contrae. Estas son clasificadas como estrategias distorsionadoras del espacio. En un sistema donde el espacio se contrae, un grupo aparecerá, en formación, para moverse cercano a algunos o todos los elementos restantes; de esta forma la posibilidad de que un elemento individual sea añadido a un grupo existente, en vez de actuar como núcleo de un nuevo grupo, se ve en aumento, siendo este sistema llamado como “cadena”.

Por otra parte, en un sistema en que el espacio se dilata ocurre exactamente lo contrario, los elementos individuales que aún no están en grupo son más propensos a formar núcleos de nuevos grupos. Una estrategia así agrupará cualquier dato para los cuales todos los d_{ij} no sean idénticos. Es inherente de esta forma la producción de grupos “no conformistas” de elementos periféricos.

1.5 Estrategias Estándares

Vecino más Cercano

Esta resulta la más antigua de las estrategias. La distancia entre los grupos es definida como la distancia (normalmente una (i, j)-medida) entre sus elementos más cercanos, uno de cada grupo. Resulta, en efecto, una estrategia combinatoria, en la cual es solo necesario seleccionar la menor medida en la unión; es inmediatamente derivable de la Ec. **1.14** mediante la condición ($\alpha_i = \alpha_j = {}^1/_2$; $\beta = 0$; $\gamma = -{}^1/_2$). Es compatible bajo todas las (i, j)-medidas, debido a que todas la medidas entre los grupos pueden ser encontradas en los elementos iniciales de la matriz. Cuando un grupo va en crecimiento simula moverse hacia algunos elementos mientras se aleja de otros, por lo

que es una estrategia donde el espacio se contrae, consecuencia de lo cual su tendencia al encadenamiento es notoria.

Vecino más Lejano

Resulta, en efecto, la antítesis de la estrategia anterior. En esta la distancia entre dos grupos es ahora definida como la distancia entre los elementos del par más lejano, uno de cada grupo. Resulta también una estrategia combinatoria, derivándose de la Ec. **1.14** por la condición ($\alpha_i = \alpha_j = {}^1/_2$; $\beta = 0$; $\gamma = {}^1/_2$). De manera similar a la estrategia anterior, esta es también compatible. Debido a que a medida que aumenta la cantidad de elementos en grupo este tiende a alejarse de algunos elementos, sin acercarse a elemento alguno, es una estrategia marcadamente dilatadora del espacio.

Centroide

En esta estrategia, el grupo, de manera algebraica, es considerado como un espacio euclidiano y es reemplazado, en formación, por las coordenadas de su centroide. Sus propiedades combinatorias no son invariantes bajo un cambio de la medida, y a continuación se procede a establecerlas para tres (i, j)-medidas.

Distancia cuadrada euclidiana

Debido a que esta es aditiva sobre los atributos, es necesario considerar un único atributo x. Sean las coordenadas del centroide de (i) denotadas como x_i. Entonces el centroide de (k) estará en $\frac{(n_i x_i + n_j x_j)}{n_k}$, y por definición,

$$d^2_{hk} = \left[x_h - \frac{n_i x_i + n_j x_j}{n_k}\right]^2 \qquad (\mathbf{1.16})$$

Realizando cambios mediante la multiplicación y reorganizando los elementos de la ecuación es muy fácil notar que el miembro derecho de la Ec. **1.14** es idénticamente igual a:

$$\frac{n_i}{n_k}(x_h - x_i)^2 + \frac{n_j}{n_k}(x_h - x_j)^2 - \frac{n_i}{n_k}\cdot\frac{n_j}{n_k}(x_i - x_j)^2$$

Expresión que es equivalente a:

$$\frac{n_t}{n_k} \cdot d_{ti}^2 + \frac{n_t}{n_k} d_{tj}^2 - \frac{n_t}{n_k} \cdot \frac{n_t}{n_k} \cdot d_{ij}^2 \qquad (\mathbf{1.17})$$

La estrategia para esta medida obtenida a partir de la Ec. **1.14** es de la forma:

$$\alpha_i = \frac{n_i}{n_k};\ \alpha_j = \frac{n_j}{n_k};\ \text{y}\ \beta = \alpha_i \cdot \alpha_j;\ \text{y}\ \gamma = 0 \qquad (\mathbf{1.18})$$

Coeficiente de correlación

La medida puede ser combinatoria mediante el almacenamiento de las covarianzas y varianzas en lugar de los coeficientes de correlación y el tamaño de los grupos. Escribimos cov_{ij} para la covarianza of (i) y (j), y v_i para la varianza de (i); el coeficiente de correlación es construido cuando se requiera a partir de la definición

$$r_{ij} = \frac{cov_{ij}}{(v_i v_j)^{1/2}}. \qquad (\mathbf{1.19})$$

De las definiciones de las cantidades concernientes, resulta fácil mostrar que

$$cov_{hk} = cov_{ht} + cov_{tj} \qquad (\mathbf{1.20})$$

$$v_k = v_i + v_j + 2cov_{ij}. \qquad (\mathbf{1.21})$$

Estas dos ecuaciones servirán para definir una solución combinatoria, aun cuando no puede ser basada en los coeficientes de correlación en sí, y no puede ser derivada de la Ec. **1.14**.

Distancia no métrica

Una solución combinatoria para esta medida debe requerir una solución para el problema siguiente: dadas las cantidades reales y positivas a, b y c, y dados los valores de $|a-b|$ y $|a-c|$, para derivar el valor de $|(a-b)+(a-c)|$. Esto es obviamente imposible; no existe, por lo tanto, una solución centroide combinatoria para esta medida, y los datos originales deben ser mantenidos almacenados.

Esta estrategia centroide es compatible para todos los coeficientes y es conservadora del espacio. La consiguiente simplicidad de este modelo no tiende a

presentar el sistema de forma agradable visualmente para los usuarios, pero no está excluida de las desventajas inherentes. En particular, la monotonicidad requerida por la Ec. **1.15** no se presenta, ni su inversión, particularmente para algunas medidas, puede ser extremadamente trabajosa.

Mediana

A la larga, una desventaja del centroide es que, si ni y nj están muy separados, el centroide de (k) tenderá a ir cerca de aquel con un grupo mayor, y permanecerá como parte de este grupo; las propiedades características del grupo menor están virtualmente perdidas.

La estrategia puede ser independiente del tamaño del grupo a través de la asignación arbitraria $n_i = n_j$; la aparente posición de (k) podrá yacer ahora entre (i) y (j), y los parámetros de la Ec. **1.14** reducidos a $\alpha_i = \alpha_j = 1/2$; $\beta = -1/4$ y $\gamma = 0$. En el modelo euclidiano, el nuevo grupo está situado en el punto medio del lado más corto del triángulo definido por (i), (j), (k); d_{hk} se encuentra a lo largo de la mediana de este triángulo, razón por la cual se sugirió el nombre de “mediana”.

Es combinatoria por definición, y completamente compatible para la distancia euclidiana cuadrada. Como la medida no métrica puede ser considerada como una distancia en un espacio no euclidiano, esto también puede ser tratado como si fuera compatible. Aunque el coeficiente de correlación puede ser manipulado (en la forma de $(1 - r_{ij})$) en el sistema, no se nos permite asignarle un significado geométrico útil y creemos que la estrategia debe ser considerada como incompatible para esta medida. El sistema es conservador del espacio, aunque la aparente posición de un grupo puede variar ampliamente. No es satisfecha la condición de la Ec. **1.15** y por tanto la estrategia puede pecar de fallos de monotonicidad.

Grupo Promedio (Group-Average)

Es combinatoria para todos los coeficientes; debido a que, si shi representa una medida sencilla inter-elemento entre (h) e (i), tenemos por definición:

$$d_{hk} = \frac{1}{n_h n_k} \cdot \sum_{h,k} s_{hk} = \frac{n_i}{n_k} \cdot \frac{1}{n_h n_i} \cdot \sum_{h,i} s_{hi} + \frac{n_j}{n_k} \cdot \frac{1}{n_h n_j} \cdot \sum_{h,j} s_{hj}$$

$$= \frac{n_i}{n_k} \cdot d_{hi} + \frac{n_j}{n_k} \cdot d_{hj} \qquad (\mathbf{1.22})$$

El sistema es por lo tanto obtenido a partir de la Ec. **1.14** cuando:

$$\alpha_i = \frac{n_i}{n_k};\ \alpha_j = \frac{n_j}{n_k};\ \text{y}\ \beta = \alpha_i \cdot \alpha_j;\ \text{y}\ \gamma = 0 \qquad (\mathbf{1.23})$$

Es completamente compatible, proporcionando el concepto de una medida promedio lo hace totalmente aceptable. El concepto de un coeficiente de correlación promedio no es del todo "feliz", y una solución más satisfactoria puede ser proporcionada para este caso mediante la asignación:

$$d_{ij} = \cos\left[\frac{1}{n_i n_j} \cdot \sum_{i,j} \cos^{-1} s_{ij}\right] \qquad (\mathbf{1.24})$$

El sistema es menos conservador del espacio que el centroide, debido a que no existen tendencias marcadas de contracción o dilatación, por lo cual puede ser considerado como una estrategia conservadora. Como $\alpha_i + \alpha_j + \beta = 1$, la Ec. **1.15** es satisfecha y el árbol resultante es necesariamente monótono. Hasta ese entonces esta estrategia no había recibido la atención que merecía.

Una Estrategia Flexible

A través de esta el sistema derivado a partir de la Ec. **1.14** mediante las restricciones ($\alpha_i + \alpha_j + \beta = 1$; $\alpha_i = \alpha_j$; $\beta < 1$; $\gamma = 0$); es combinatoria por definición. Es compatible para la distancia euclidiana aunque la propiedad estrictamente euclidiana está perdida; la restricción 1/0 puede fallar con la medida no métrica, pero con una cantidad tan arbitraria se cree que a esto no debe prestársele mucha importancia. Resulta sin sentido si se aplica al coeficiente de correlación, y para esta medida es completamente incompatible. Su flexibilidad yace en sus propiedades de distorsión del espacio. Como β se aproxima a la unidad, es muy probable que, luego de una unión, la distancia aparente desde el primer grupo al elemento más cercano

será siempre menor que alguna distancia entre elementos restantes. El sistema, en efecto, resulta altamente capaz de contraer el espacio, y apartando solo las ambigüedades iniciales, puede encadenar completamente tomando β lo suficientemente cerca a la unidad.

Si β se acerca a cero o se hace negativa, el sistema deja de contraer y adquiere una gran capacidad de dilatación del espacio, y los elementos correspondientemente más intensamente agrupados.

1.6 Aplicaciones en Química

Habiendo dado una breve reseña de los métodos de aglomeración jerárquica aglomerativos, nos disponemos a exponer, también de forma breve, las aplicaciones de los mismos en la química.

Las principales aplicaciones de estos métodos en esta ciencia están centradas en la química combinatoria, la adquisición de nuevos compuestos químicos y QSAR. Se hace mucho énfasis en el uso de estos métodos en las aplicaciones farmacéuticas debido a que en los análisis llevados a cabo en estas compañías tienden a procesar juegos de datos de gran tamaño. Encontramos, en especial, un amplio uso del método de Ward, habiendo sido comparado su comportamiento con otros métodos de selección de composiciones de manera aleatoria, devolviendo el método de Ward resultados mucho más consistentes, demostrándose también la relación existente entre el análisis de diversidad y el análisis de conglomerados, marcando así las pautas para realizar el uno apoyándose en el otro [22].

Entre las aplicaciones para la química de algoritmos tipo Ward, podemos mencionar un programa llamado VisualiSAR, el cual soporta la búsqueda de estructuras y el desarrollo de relaciones estructura-actividad en juegos de datos HTS. La compañía Johnson & Johnson ha incorporado también aglomeración de Ward en un sistema, llamado CerBeruS, cuyo uso específico es para el análisis de las bases de datos de dicha corporación, notándose gracias a esta aplicación dos ventajas de la investigación basada en conglomerados. Primero, si se encontraba una combinación exitosa, las combinaciones relacionadas podían ser probadas mediante la extracción de otras

posibles candidatas del conglomerado que las contiene. Segundo, el análisis de relaciones estrucura-actividad (SAR) puede ser formulado vinculando los resultados de las corridas con el examen de la jerarquía de conglomerados a diferentes niveles. Como consecuencia de esto, se desarrolló un procedimiento secuencial de dos etapas, soportado por la aglomeración para hacer HTS más eficiente.

Capítulo 2. Materiales y Métodos

2.1 Algoritmos de Conglomerados Secuenciales, Aglomerativos, Jerárquicos, No Superpuestos y Combinatorios

La familia de los algoritmos de conglomerados secuenciales, aglomerativos, jerárquicos, no superpuestos y combinatorios, o más simplemente en este trabajo, Algoritmos de Conglomerados Aglomerativos y Combinatorios (AC2) requiere que solo una matriz de similitud/disimilitud simétrica W permanezca en la memoria durante los cálculos [enfoque de la *matriz almacenada* de Anderberg [23]]. Los datos originales no necesitan ser guardados porque existe una solución *combinatoria* para recalcular las medidas inter-conglomerados usando la información contenida en W y en cierto arreglo que contiene los tamaños de los conglomerados. La primera aproximación para tal comportamiento fue aportada primeramente por Lance y Williams [21] que posteriormente fue extendida por Jambu y Lebeaux [24, 25], para derivar en la fórmula recurrente de Lance-Williams-Jambu:

$$w_{h,t\cup f} = \alpha_t w_{ht} + \alpha_f w_{hf} + \beta\, w_{tf} + \gamma\left|w_{ht} - w_{hf}\right| + \lambda_h w_h + \lambda_t w_t + \lambda_f w_f \qquad (\mathbf{2.1})$$

Si los conglomerados C_t y C_f se unen en un ciclo de aglomeración, entonces $w_{h,t\cup f}$ aporta el valor del criterio a ser usado en el próximo ciclo para el conglomerado $C_t \cup C_f$ con cualquier otro conglomerado C_h. Cada combinación de los parámetros $(\alpha_t, \alpha_f, \beta, \gamma, \lambda_h, \lambda_t, \lambda_f)$ define un algoritmo en específico. El valor actualizado, $w_{h,t\cup f}$, se determina acorde a seis valores de W:

$$\begin{pmatrix} w_{hh} & w_{ht} & w_{hf} \\ \cdot & w_{tt} & w_{tf} \\ \cdot & \cdot & w_{ff} \end{pmatrix} \qquad (\mathbf{2.2})$$

Donde: $w_{h,tf} = w\left[C_h \cup (C_t \cup C_f)\right]$; $w_{tf} = w(C_t \cup C_f)$; $w_{tt} = w_t = w(C_t)$

Más recientemente, Podani [26] propuso una nueva clasificación de los AC2s en algoritmos *basados en distancia* (d-AC2) o *basados en homogeneidad* (h-AC2), en

dependencia del tipo de medida de similitud/disimilitud que emplean como criterio de fusión:

1. Algoritmos d-AC2: cuando el criterio de fusión viene dado por una medida de distancia inter-conglomerados, o sea, cuando las entradas de W vienen dadas por $w_{ij} = d(C_i, C_j)$ o $w_{ij} = d^2(C_i, C_j)$.

2a. Algoritmos nh-AC2: cuando el criterio de fusión viene dado por una medida de homogeneidad intra conglomerados, o sea, cuando las entradas de W vienen dadas por $w_{ij} = h(C_i, C_j)$.

2b. Algoritmos ch-AC2: cuando el criterio de fusión viene dado por un cambio en la homogeneidad durante la fusión de dos conglomerados, o sea, cuando las entradas de W vienen dadas por $w_{ij} = h(C_i, C_j) - h(C_i) - h(C_j)$.

A medida que procede la aglomeración, los criterios del tipo 1 y 2b se minimizan, mientras que el criterio de tipo 2a se maximiza. De este modo, Podani en su trabajo colecta y clasifica 16 técnicas combinatorias (tres de su autoría) que se presentan en la Tabla 2.1.

Eficiencia de los algoritmos

Existen dos estrategias algorítmicas para llevar a cabo la actualización de la matriz W, estas son: *el par más cercano* [ver por ejemplo [19]] y *el vecino más cercano recíproco* [ver por ejemplo [27]]. En términos comparativos, a pesar de que el segundo algoritmo mejora al primero en cuanto a recursos computacionales empleados, destruye las propiedades ultramétricas de algunos métodos, trayendo consigo la aparición indeseada de *reversos* en los dendogramas respectivos [26]. En nuestro trabajo implementamos algunos de los algoritmos mostrados en la Tabla 2.1 y los comparamos con la versión correspondiente implementada en el software propietario SYN-TAX2000 [28], los mismos fueron (ver Tabla 2.2):

Tabla 2.1. Parámetros para métodos de aglomeración combinatorios d-AC2 (1-8) y h-AC2 (9-16)

Table 1. Parameters for combinatorial d-SAHN (1–8) and h-SAHN (9–16) clustering methods.

Clustering method	α_i	α_j	β	γ	λ_h	λ_i	λ_j
1. Single linkage (SL)	1/2	1/2	0	− 1/2	0	0	0
2. Complete linkage (CL)	1/2	1/2	0	1/2	0	0	0
3. Unweighted average (UPGMA)	$\frac{n_i}{n_h + n_i}$	$\frac{n_j}{n_h + n_j}$	0	0	0	0	0
4. Weighted average (WPGMA)	1/2	1/2	0	0	0	0	0
5. Centroid (UPGMC)	$\frac{n_i}{n_h + n_i}$	$\frac{n_j}{n_h + n_j}$	$-\frac{n_i n_j}{(n_i + n_j)^2}$	0	0	0	0
6. Median (WPGMC)	1/2	1/2	− 1/4	0	0	0	0
7. β-Flexible (β-FLEX)	$(1-\beta)/2$	$(1-\beta)/2$	<1	0	0	0	0
8. (β, γ)-Flexible ((β, γ)-FLEX)	$(1-\beta)/2$	$(1-\beta)/2$	unrestricted	unrestricted	0	0	0
9. Minimum increase of sum of squares (MISSQ)	$\frac{n_h + n_i}{n.}$	$\frac{n_h + n_j}{n.}$	$-\frac{n_h}{n.}$	0	0	0	0
10. Minimum sum of squares of new cluster (MNSSQ)	$\frac{n_h + n_i}{n.}$	$\frac{n_h + n_j}{n.}$	$\frac{n_i + n_j}{n.}$	0	$-\frac{n_h}{n.}$	$-\frac{n_i}{n.}$	$-\frac{n_j}{n.}$
11. Minimum increase of variance (MIVAR)	$\left(\frac{n_h + n_i}{n.}\right)^2$	$\left(\frac{n_h + n_j}{n.}\right)^2$	$-\frac{n_h(n_i + n_j)}{n^2}$	0	0	0	0
12. Minimum variance of new cluster (MNVAR)	$\left(\frac{n_h + n_i}{n.}\right)^2$	$\left(\frac{n_h + n_j}{n.}\right)^2$	$\left(\frac{n_i + n_j}{n.}\right)^2$	0	$-\left(\frac{n_h}{n.}\right)^2$	$-\left(\frac{n_i}{n.}\right)^2$	$-\left(\frac{n_j}{n.}\right)^2$
13. Minimum increase of weighted average distance (WMIDIS)	$\frac{b_{hi}}{b.}$	$\frac{b_{hj}}{b.}$	$\frac{b_{ij}}{b.} - \frac{1}{2}$	0	$-\frac{b_h + n_i n_j}{2b.}$	$-\frac{b. - 2b_i - 2n_h n_j}{4b.}$	$-\frac{b. - 2b_j - 2n_h n_i}{4b.}$
14. Minimum increase of unweighted average distance (UMIDIS)	$\frac{b_{hi}}{b.}$	$\frac{b_{hj}}{b.}$	$\frac{b_{ij}}{b.} - \frac{b_{ij}}{b_h + b_{ij}}$	0	*	*	*
15. Minimum average distance within new cluster (MNDIS)	$\frac{b_{hi}}{b.}$	$\frac{b_{hj}}{b.}$	$\frac{b_{ij}}{b.}$	0	$-\frac{b_h}{b.}$	$-\frac{b_i}{b.}$	$-\frac{b_j}{b.}$
16. λ-Flexible (λ-FLEX)	$(1-3\lambda)/3$	$(1-3\lambda)/3$	$(1-3\lambda)/3$	0	$\leqq 0$	$= \lambda_h$	$= \lambda_h$

$b_i = \binom{n_i}{2}$, n_i = number of objects in C_i, $n. = n_h + n_i + n_j$, * expression too long, see formula (14) in Appendix.

*Los métodos 10, 12, 15 pertenecen a la sub-clasificación nh-AC2; los métodos 9, 11, 13, 14 pertenecen a la clasificación ch-AC2; el método 16 se clasifica en ambos, en dependencia del valor de λ. Esta tabla fue reproducida con permiso del autor.

Tabla 2.2. Algoritmos de Conglomerados SAHN combinatorios empleados en el trabajo

Informe[a]	Nombre[b]	Tabla 1[c]
SL	Unión Simple (Vecino más Cercano)	1
CL	Unión Completa (Vecino más Lejano)	2
GA	Promedio de Grupos	3

SA	Promedio Simple	4
Cen	Centroide	5
Med	Mediana	6
MSSN	Suma de Cuadrado Mínima en Nuevo Conglomerado	10
MVN	Varianza Mínima en Nuevo Conglomerado	12
MADN	Distancia Promedio en Nuevo Conglomerado	15
MISS	Incremento Mínimo de la Suma de Cuadrados (Ward)	9
MIV	Incremento Mínimo de Varianza	11
MIAD	Incremento Mínimo de Distancia Promedio (No Ponderada)	14

[a]Notación y orden de los algoritmos de conglomerados empleados a lo largo del informe de tesis en correspondencia con la interfaz del SYN-TAX2000, [b]Nombre completo según su traducción al Español, [c]Posición que ocupan en la Tabla 2.1 donde aparecen los parámetros correspondientes de la Ec. **1**

2.2 Conjuntos de Datos

La medición del rendimiento de los índices de similitud, descriptores moleculares (DMs), e incluso enfoques de validación, es estrictamente dependiente de las moléculas/bases de datos de prueba, de la configuración del espacio químico y de la problemática tratada. Este problema se pudiera arreglar evaluando los métodos nuevos en bases de datos populares como el conjunto de datos (CD) de esteroides, el CD del NCI (*National Cancer Institute*), las bases de datos WDI (World Drug Index) y MDDR (MACCS Drug Data Report), o estableciendo metodologías concisas. Desafortunadamente, la comunidad científica internacional no ha adoptado ningún CD estándar para la comparación de medidas de similitud y DMs, probablemente por la imposibilidad de encontrar un grupo único de moléculas que reagrupe todas las necesidades de cribado de la Quimioinformática moderna [29]. Por este motivo se ha sugerido que, para validar un método nuevo, los investigadores deben presentar al menos 10 conjuntos con actividades diversas con más de un estándar de comparación [30]. Una revisión exhaustiva acerca de las bases de datos empleadas actualmente en la Quimioinformática, haciendo énfasis en las bases de datos farmacológicas se puede encontrar en [31]. La tendencia actual de dichos repositorios es pasar al dominio público [32]. Específicamente, en estudios comparativos reportados de algoritmos de conglomerados se han utilizado bases de datos para las cuales se dispone de propiedades físicas, químicas o biológicas [33].

Para nuestro estudio, se seleccionaron ocho conjuntos de datos diferentes de la Química Medicinal, usados originalmente por Sutherland *et al.* [34], y empleados más recientemente por otros investigadores [35-38]. La descripción de los mismos se muestra en la Tabla 2.3.

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

CD[a]	Diana farmacológica	Número de compuestos	Variable farmacocinética	Rango de Valores
ACE	Inhibidores de la enzima convertidora de angiotensina	114	pIC50	2,1-9,9
AchE	Inhibidores de la acetilcolinesterasa	111	pIC50	4,3-9,5
BZR	Ligandos para el receptor de la benzodiacepina	163	pIC50	5,5-8,9
COX-2	Inhibidores de la ciclooxigenasa	322	pIC50	4,0-9,0
DHFR	Inhibidores de la hidrofolato reductasa	397	pIC50	3,3-9,8
GPB	Inhibidores de la glicógeno fosforilasa b	66	pKi	1,3-6,8
THER	Inhibidores de la termolisina	76	pKi	0,5-10,2
THR	Inhibidores de la trombina	88	pKi	4,4-8,5

[a]Los conjuntos de datos se presentan en el orden de la fuente original [34]. Los mismos se pueden acceder libremente en ***http://www.cheminformatics.org/datasets/index.shtml***; [b]pIC50 = -logIC50, donde IC50 representa la mitad de la concentración inhibitoria máxima, se usa como una medida de la potencia del fármaco, pKi = - logKi, donde Ki representa la constante de inhibición del fármaco, también se usa como una medida de la potencia del fármaco.

2.3 Espacio Químico y Representación Molecular

Cercanamente aliado con la noción de similitud molecular está el concepto de *espacio químico*. Los espacios químicos proveen un medio para conceptualizar y visualizar la similitud molecular. El concepto de espacio químico se deriva de la noción de espacio usado en Matemáticas y consiste en un conjunto de moléculas y un conjunto de relaciones asociadas (e.g., similitudes, disimilitudes, distancias) entre las moléculas, lo cual le da al espacio una "estructura" [39].

El espacio químico se puede describir usando una codificación *basada en coordenadas* o una codificación *libre de coordenadas* de las estructuras químicas. En la codificación individual de moléculas (espacio basado en coordenadas) cada molécula

se describe mediante un vector de fragmentos o subestructuras, traducido posteriormente en un vector de DMs y, por tanto, tiene una posición absoluta en un espacio multidimensional, la dimensión de este espacio se especifica por el número de rasgos no correlacionados (e.g., descriptores de complejidad, descriptores de solubilidad, huellas dactilares, tripletes de farmacóforos, u otro vector de descriptores). Por otra parte, en la codificación por pares de moléculas (espacio libre de coordenadas) solo se calculan las distancias entre dos moléculas usando una medida de similitud explícita o implícita (quizás infinita). La posición absoluta de las moléculas en este espacio se puede calcular solamente si se miden todas distancias por pares y la dimensionalidad del espacio puede ser conocida (e.g., descriptores de pares de átomos, árboles de rasgos, enfoques de Subestructura Máxima Común) [40-42].

Un gran número de descriptores se han desarrollado para usarse en cálculos de similitud molecular. Los mismos son diseñados típicamente para brindar una descripción molecular que es transferible, en una representación conservadora de la información, a un espacio de descriptores abstracto [43]. Sin embargo, a medida que la dimensionalidad de los datos se incrementa, muchos tipos de análisis de datos y problemas de clasificación se vuelven significativamente difíciles. En ocasiones también los datos se vuelven crecientemente dispersos en el espacio que ocupan. Esto puede conducir a grandes problemas para ambos, para el aprendizaje supervisado y no supervisado. En la literatura este fenómeno se refiere como “la maldición de la dimensionalidad” [44]. Para propósitos de aglomeración, el aspecto más relevante de la maldición de la dimensionalidad concierne a la medida de distancia o similitud. Para ciertas distribuciones de datos, la diferencia relativa entre las distancias de los puntos más cercanos y lejanos a un punto independientemente seleccionado tiende a cero a medida que la dimensionalidad aumenta [45]. Una estrategia para solucionar esta dificultad es seleccionar un conjunto de descriptores en particular para los cuales se demostró que funcionan bien en un cierto problema. Otra estrategia es calcular primero un gran número de descriptores y luego eliminar aquellos descriptores del conjunto que muestran un coeficiente de correlación por encima de cierto valor. Un enfoque diferente es dejar que la computadora escoja la combinación óptima de descriptores para el

problema en cuestión [46]. En resumen, existe una amplia variedad de DMs y métricas usadas en los métodos de similitud molecular; parece ser, sin embargo, que el mejor rendimiento se logra adaptando dicha combinación al problema estudiado [47].

Cada uno de los ocho conjuntos de datos químicos, detallados en el epígrafe anterior, disponibles en ficheros de datos nombre1.sd, que contienen las estructuras 2D de las moléculas respectivas, junto a la variable respuesta o clase numérica (pIC50 o pKi) y otros datos menos relevantes, se re-optimizaron y convirtieron al formato nombre2.sdf con el Software Generador de Estructuras 3D CORINA esta vez utilizando el comando (ejecutable corina.exe a través de la consola cmd):

corina -d wh,rs,neu,r2d,errorfile=errors.sdf nombre1.sd nombre2.sdf.sdf

Los parámetros entrados fueron: wh, escribir hidrógenos adicionados; rs, eliminar fragmentos pequeños; neu: neutralizar cargas formales; r2d, eliminar entradas 2D del fichero de salida; errorfile=errors.sdf, escribir moléculas fallidas/erróneas a errors.sdf [48].

Los ocho ficheros de salidas resultantes, de estructuras 3D optimizadas, nombre2.sdf se cargaron en el Software de Cálculos de Descriptores Moleculares DRAGON [49] y se calcularon todas las familias de DMs disponibles (3 224 descriptores en total). No se hizo pre filtrado y todos los descriptores, junto a la variable respuesta correspondientes a cada una de los ocho conjuntos de datos, se salvaron en un único fichero nombre3.txt.

Cada uno de estos ficheros se cargó en el software Excel 2007 [50] y se eliminaron los rasgos binarios.

Cada uno de los ocho ficheros anteriores se prepararon adecuadamente y transformaron al formato nombre4.csv para cargarlos al Software de Minería de Datos Weka v. 3-6-2 [51], donde se sometieron a un tratamiento que comprendió el pre-filtrado, re-escalado y selección de rasgos. Todas las etapas se sistematizaron y aplicaron en un flujo de conocimiento como se aprecia en la figura 1.

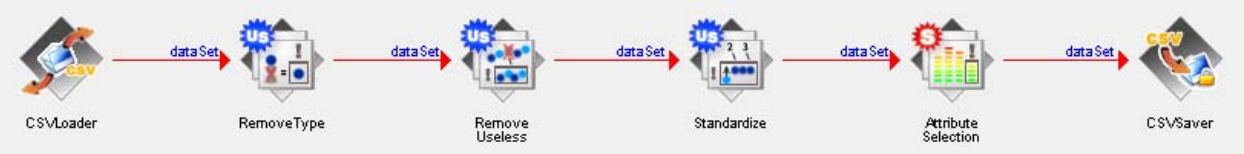


Figura 1. Diagrama del Flujo de Conocimiento empleado en el trabajo para el pre-filtrado, re-escalado y selección de rasgos, según la interfaz directa del Weka

Los elementos comprendidos en el flujo de conocimiento fueron: CSVLoader, lee una fuente que se encuentre en formato separado por coma o tab; RemoveType, elimina atributos de un tipo dado, en este caso se configuró para eliminar atributos nominales; RemoveUseless, elimina atributos que varían mucho o no varían en absoluto, se dejaron los parámetros por defecto; Standardize, estandariza todos los atributos numéricos para que tengan media 0 y varianza 1, en nuestro caso también se estandarizó la clase numérica; AtributeSelection, es un filtro supervisado que se puede utilizar para seleccionar atributos, se utilizó el evaluador CfsSubsetEval con el resto de los parámetros que aparecían por defecto, este evaluador selecciona subconjuntos de rasgos con una alta correlación con la clase y una baja correlación entre ellos [52], como los rasgos se escalaron de modo que tuvieran igual media (media 0) y varianza (varianza 1) la medida de correlación es intraclásica y entonces el evaluador selecciona los rasgos con una alta consistencia con la clase y una baja consistencia entre ellos [53], a pesar de su simplicidad, este evaluador ha sido utilizado por otros investigadores sobre los mismos conjuntos de datos con resultados relativamente buenos en la exactitud de los clasificadores empleados [54, 55]; CSVSaver, escribe a un destino en formato separado por coma.

Los ocho ficheros resultantes se acondicionaron al formato nombre5.dat para la matriz de datos, y, var.lab y case.lab para las etiquetas de las variables y casos, respectivamente, según los requerimientos para cargarlos en el Software para Análisis de Datos en Ecología y Sistemática SYN-TAX2000 [28], donde se corrieron los algoritmos de conglomerados d-AC2, h-AC2 y ch-AC2 explicados anteriormente, con la métrica euclidiana como distancia entre los centroides. Para la resolución de ataduras se seleccionó la estrategia de *fusión sub-óptima* con lo cual queda fijada la estrategia

de fusión en el *par más cercano*. Para cada una de las 12 corridas (12 algoritmos de conglomerados) correspondientes a cada uno de los 8 conjuntos de datos (96 corridas en total) se salvaron los ficheros de salida: análisis.txt, dendograma.den y árbol.emf en la carpeta del CD correspondiente. Con la prestación del software *partición a partir de un dendograma* se cargó el fichero dendograma.den y se obtuvo la partición para un número de conglomerados de 2 (caso binario).

Para cada uno de los conjuntos de datos se cargaron cada una de las 8 particiones obtenidas (particiones binarias) junto a la dicotomización binaria propuesta por Bruce *et al.* [56] en un fichero de Excel donde se calcularon 5 medidas de la calidad externa de los conglomerados F-measure (o coeficiente de Dice en el caso binario), exactitud global (Qt), el coeficiente de correlación de Mattews y el coeficiente de Tanimoto. La primera medida se ha utilizado tradicionalmente como medida subjetiva y externa de la calidad de la aglomeración [57]; las siguientes dos como medidas de la exactitud de los algoritmos de predicción [58]; la tercera medida, sin embargo, se ha utilizado, con los mejores resultados, como medida de similitud en el cribado virtual de bases de datos químicas empleando huellas dactilares (descriptores binarios) [59]. Una revisión exhaustiva de medidas de calidad tanto internas como externas se puede encontrar en [17].

2.4 Análisis Estadístico

El interés principal de la investigación estuvo focalizado en la comparación de los AC2s con el algoritmo propuesto originalmente por Ward. A pesar de que el teorema *ningún almuerzo es gratis* (NFL, de sus siglas en inglés, *No Free Lunch*) nos permite predecir que nuestro objetivo es inútil cuando se promedian los resultados de los modelos para la *población* de los conjuntos (bases) de datos [60], el principio de la *longitud mínima de la descripción* (MDL, de sus siglas en inglés, *Minimun Description Length*) nos permite predecir que dada una *muestra* de tales conjuntos (bases) de datos es posible encontrar el modelo con el mejor balance entre ajuste y simplicidad [61]. La diferencia está en que mientras NFL opera sobre el dominio completo de todas

las hipótesis, MDL trata de generalizar su hipótesis construida sobre los datos observados [62].

Partiendo de estas premisas, para cada CD se midió la misma cualidad (cada una de las 4 medidas subjetivas o de calidad externa de la aglomeración, ver final del epígrafe anterior) obtenida con cada uno de los 12 algoritmos de conglomerados. Dado un arreglo de los 8 conjuntos de datos, la matriz de datos resultante **MD** (8X12) se puede considerar como consistente en 12 muestras de algoritmos de conglomerados que están ajustados (esto es, estratificados, cada estrato contribuyendo con una observación para cada muestra) de acuerdo a cada CD sobre el cual actúan y que afecta así mismo la cualidad medida para cada uno de los mismos [63]. Luego de una transformación tipo-Friedman o RT-2 [64], la nueva matriz de rangos $\mathbf{MD_r}$ es susceptible a un análisis de varianza de dos criterios por rangos de Friedman [65], o mediante la versión mejorada de Iman y Davenport [66-68]. Equivalentemente, se puede llevar a cabo una evaluación de comunidad de juicios entre 8 ranqueos hechos sobre 12 objetos de Kendall [69], o mediante la versión mejorada de Fagot [70].

Capítulo 3. Análisis de los Resultados

3.1 Comportamiento de los Descriptores Moleculares

En la Tabla 3.1 se muestran los descriptores linealmente relevantes a cada una de las actividades estudiadas, seleccionados por la técnica de selección de rasgos CfsSubsetEval del Weka.

Tabla 3.1. Lista de DMs seleccionados por la técnica de selección de rasgos CfsSubsetEval del Weka

BD[a]	DM[b]	Bloque[c]	Dim[d]	BD[a]	DM[b]	Bloque[c]	Dim[d]
	MAXDP	topological descriptors	2D		Lop	topological descriptors	2D
	PW4	topological descriptors	2D		D/Dr09	topological descriptors	2D
	Lop	topological descriptors	2D		X1A	connectivity indices	2D
	BIC5	information indices	2D		G(N..O)	geometrical descriptors	3D
	ATS4m	2D autocorrelations	2D		RDF060m	RDF descriptors	3D
	MATS8m	2D autocorrelations	2D		RDF060v	RDF descriptors	3D
	MATS3p	2D autocorrelations	2D		Mor12u	3D-MoRSE descriptors	3D
	EEig03d	edge adjacency indices	2D		Mor30m	3D-MoRSE descriptors	3D
	EEig11d	edge adjacency indices	2D		Mor08v	3D-MoRSE descriptors	3D
	EEig12d	edge adjacency indices	2D		Mor30v	3D-MoRSE descriptors	3D
	DISPp	geometrical descriptors	3D	cox-2	Mor12e	3D-MoRSE descriptors	3D
	RDF035u	RDF descriptors	3D		E3u	WHIM descriptors	3D
ace	RDF035m	RDF descriptors	3D		P1v	WHIM descriptors	3D
	RDF035e	RDF descriptors	3D		E1e	WHIM descriptors	3D
	RDF035p	RDF descriptors	3D		R6u+	GETAWAY descriptors	3D
	Mor23m	3D-MoRSE descriptors	3D		R3m+	GETAWAY descriptors	3D
	Mor26v	3D-MoRSE descriptors	3D		H-049	atom-centred fragments	1D
	Mor26p	3D-MoRSE descriptors	3D		O-058	atom-centred fragments	1D
	E3u	WHIM descriptors	3D		F03[N-O]	2D frequency fingerprints	2D
	E1p	WHIM descriptors	3D		F05[N-N]	2D frequency fingerprints	2D
	C-006	atom-centred fragments	1D		F07[N-F]	2D frequency fingerprints	2D
	C-026	atom-centred fragments	1D	dhfr	nR05	constitutional descriptors	0D
	ALOGP2	molecular properties	Otros		D/Dr10	topological descriptors	2D
	F03[O-O]	2D frequency fingerprints	2D		GATS7m	2D autocorrelations	2D
	F06[O-O]	2D frequency fingerprints	2D		GATS6p	2D autocorrelations	2D
ache	D/Dr07	topological descriptors	2D		BELm2	Burden eigenvalues	2D
	IC4	information indices	2D		BELe1	Burden eigenvalues	2D
	SIC5	information indices	2D		RCI	geometrical descriptors	3D
	BIC5	information indices	2D		Mor10u	3D-MoRSE descriptors	3D
	MATS4m	2D autocorrelations	2D		Mor03m	3D-MoRSE descriptors	3D

	MATS4p	2D autocorrelations	2D
	GATS4m	2D autocorrelations	2D
	GATS6e	2D autocorrelations	2D
	GATS5p	2D autocorrelations	2D
	JGI10	topological charge indices	2D
	RDF045u	RDF descriptors	3D
	RDF090u	RDF descriptors	3D
	RDF155u	RDF descriptors	3D
	RDF090m	RDF descriptors	3D
	RDF090e	RDF descriptors	3D
	RDF155e	RDF descriptors	3D
	Mor22m	3D-MoRSE descriptors	3D
	Mor11e	3D-MoRSE descriptors	3D
	Mor32e	3D-MoRSE descriptors	3D
	E3u	WHIM descriptors	3D
	G2m	WHIM descriptors	3D
	G3v	WHIM descriptors	3D
	G1e	WHIM descriptors	3D
	E3p	WHIM descriptors	3D
	R6e+	GETAWAY descriptors	3D
	nR=Cs	functional group counts	1D
	nArCONR2	functional group counts	1D
	H-053	atom-centred fragments	1D
	O-058	atom-centred fragments	1D
bzr	TI2	topological descriptors	2D
	Vindex	information indices	2D
	ATS7e	2D autocorrelations	2D
	J3D	geometrical descriptors	3D
	HOMA	geometrical descriptors	3D
	RDF020u	RDF descriptors	3D
	RDF030m	RDF descriptors	3D
	RDF055m	RDF descriptors	3D
	RDF020p	RDF descriptors	3D
	RDF030p	RDF descriptors	3D
	Mor09u	3D-MoRSE descriptors	3D
	Mor04v	3D-MoRSE descriptors	3D
	G3p	WHIM descriptors	3D
	H7m	GETAWAY descriptors	3D
	R6u	GETAWAY descriptors	3D
	R3m	GETAWAY descriptors	3D
	Mor04m	3D-MoRSE descriptors	3D
	Mor09e	3D-MoRSE descriptors	3D
	R5u	GETAWAY descriptors	3D
	C-033	atom-centred fragments	1D
	O-057	atom-centred fragments	1D
	F04[C-N]	2D frequency fingerprints	2D
	F04[N-O]	2D frequency fingerprints	2D
	X5A	connectivity indices	2D
	BIC1	information indices	2D
	MATS8v	2D autocorrelations	2D
	MATS7e	2D autocorrelations	2D
gbp	Mor13m	3D-MoRSE descriptors	3D
	R5m+	GETAWAY descriptors	3D
	C-006	atom-centred fragments	1D
	H-046	atom-centred fragments	1D
	F02[N-O]	2D frequency fingerprints	2D
	F07[O-O]	2D frequency fingerprints	2D
	X5v	connectivity indices	2D
	IC1	information indices	2D
	GATS5m	2D autocorrelations	2D
	GATS7p	2D autocorrelations	2D
	RDF065m	RDF descriptors	3D
	Mor17m	3D-MoRSE descriptors	3D
	Mor31m	3D-MoRSE descriptors	3D
therm	Mor16e	3D-MoRSE descriptors	3D
	Du	WHIM descriptors	3D
	R5p	GETAWAY descriptors	3D
	nCt	functional group counts	1D
	nROH	functional group counts	1D
	F01[O-S]	2D frequency fingerprints	2D
	F03[C-N]	2D frequency fingerprints	2D
	F09[C-N]	2D frequency fingerprints	2D
thr	TI2	topological descriptors	2D
	MATS5v	2D autocorrelations	2D
	EEig02x	edge adjacency indices	2D
	JGI7	topological charge indices	2D
	P2u	WHIM descriptors	3D
	E2p	WHIM descriptors	3D
	E3s	WHIM descriptors	3D
	H8m	GETAWAY descriptors	3D

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

[a]CD: Conjuntos de Datos utilizados (ver Tabla 2.3); [b]DM: notación de los Descriptores Moleculares seleccionados; [c]Bloque: Clasificación acorde al Bloque o familia de descriptores a la que pertenecen; [c]Dim: Clasificación acorde a la Dimensionalidad o complejidad de la representación molecular que emplean

Se decidió estudiar el comportamiento de las distribuciones empíricas de frecuencias de los DM seleccionados (ver Tabla 3.1), para cada CD según el bloque a que pertenecen, y compararlas con la distribución obtenida por la fusión de los DMs para todos los CDs y con la distribución empírica *a priori* del DRAGON. Para ello se empleó una prueba de bondad de ajuste χ^2, la Tabla 3.2 muestra los resultados.

Tabla 3.2. Significación de la prueba χ^2 de bondad de ajuste entre distribuciones empíricas de frecuencias de DMs

Objetivo[a]	Fusión	ace	ache	bzr	cox-2	dhfr	gbp	therm	thr
Fusión	1	0,4300	0,5770	0,7245	0,8601	0,0081**	0,7834	0,8080	0,2837
DRAGON	~ 0***	0,0011**	~ 0***	0,0022**	0,0027**	0,0031**	0,0152*	0,1840	0,0045**

[a]Columna que representa las distribuciones objetivo, las restantes columnas con las distribuciones a comparar con las primeras; *Pruebas medianamente significativas ($p < 0,05$), **pruebas significativas ($p < 0,01$), ***pruebas muy significativas ($p < 0,001$)

De los resultados anteriores se puede inferir que existe una similitud significativa entre las distribuciones de DMs de cada CD por separado en comparación con la distribución de DMs de los CDs fusionados (**Fusión**) o, de otra forma, la fusión de los CDs es representativa de cada uno de los CD por separado, quizás con excepción del CD **dhfr**. Sin embargo, cabe inferir que existe una disimilitud significativa entre las distribuciones de DMs de cada CD por separado en comparación con la distribución empírica *a priori* del DRAGON, lo cual indica que la selección de DMs no fue casual sino que estuvo guiada por una relación DMs-Actividad o DMs-Fenómeno farmacológico en específico. Los Gráficos 3.1 y 3.2 muestran más explícitamente el comportamiento de los DMs para los datos fusionados teniendo en cuenta la dimensionalidad y los bloques o familias a que pertenecen, correspondientemente.

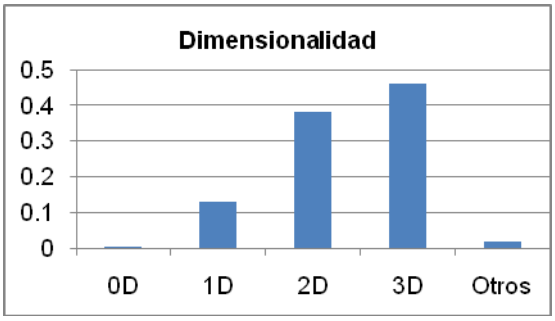


Gráfico 1a. Histograma de frecuencias relativas del comportamiento de los DMs en los CDs fusionados según su dimensionalidad

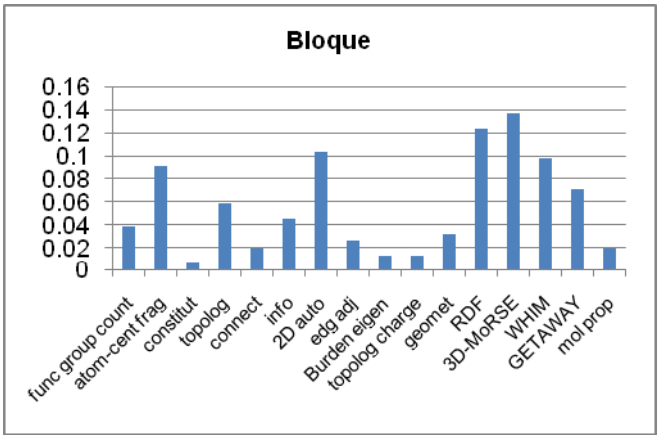


Gráfico 1b. Histograma de frecuencias relativas del comportamiento de los DMs en los CDs fusionados según el bloque a que pertenecen

Resulta interesante observar que la información química que se relaciona más directamente con las actividades farmacológicas bajo estudio está codificada mayoritariamente por DMs 2D y 3D ya que entre ambos aportan aproximadamente el 84 % de los rasgos de interés. Un análisis más refinado de comparación de proporciones, basado en una prueba binomial, arrojó que solo estos dos grandes "grupos" se difieren significativamente de la uniformidad ($p \sim 0$). Un análisis similar, pero ahora realizado con los bloques de DMs, mostró (ver Tabla 3.3) que solo las familias 2D autocorrelations (2D), RDF descriptors (3D), 3D-Morse descriptors (3D) y WHIM descriptors (3D) difieren significativamente de la uniformidad y, por tanto, tuvieron la incidencia más marcada en el resultado anterior. En términos químico-teóricos estos dos resultados significan que, adicionalmente a la topología molecular, la configuración espacial de los átomos juega un rol dominante en la interacción ligando-receptor, lo cual es consistente con lo observado por otros investigadores [30]. A pesar de que nuestro estudio no abarca un gran número de dianas/actividades farmacológicas, este resultado soporta en parte el uso directo de solo estas cuatro familias de DMs en contextos farmacológicos donde solo se disponga de información estructural de las entidades químicas.

Tabla 3.3. Resultados de la prueba binomial para los bloques de DMs

Bloque	Sig[a]	Bloque	Sig[a]	Bloque	Sig[a]

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

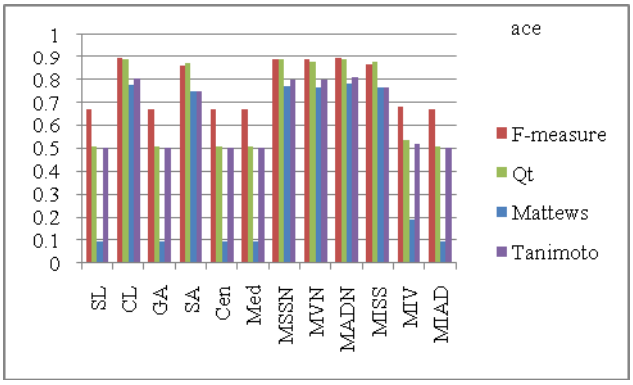
[a]Signifación de la prueba binomial para comparación de proporciones, el valor contrastado de probabilidad fue de p = 0,0625 en todos los casos (distribución uniforme). Se utilizó la aproximación mediante la distribución normal con la corrección X + 0,5 para la variable aleatoria. *Pruebas significativas (p < 0,05), ***pruebas extremadamente significativas (p < 0,001)

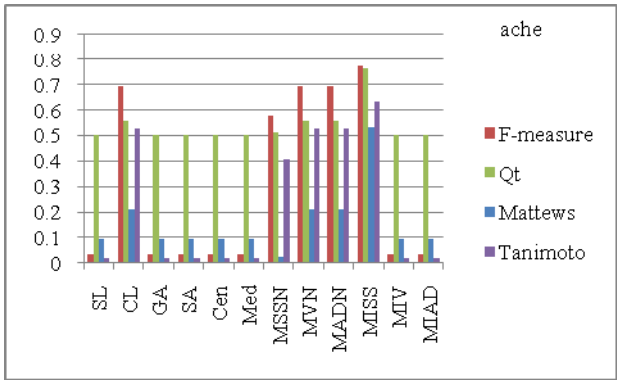
3.2 Comparación de los Algoritmos de Conglomerados

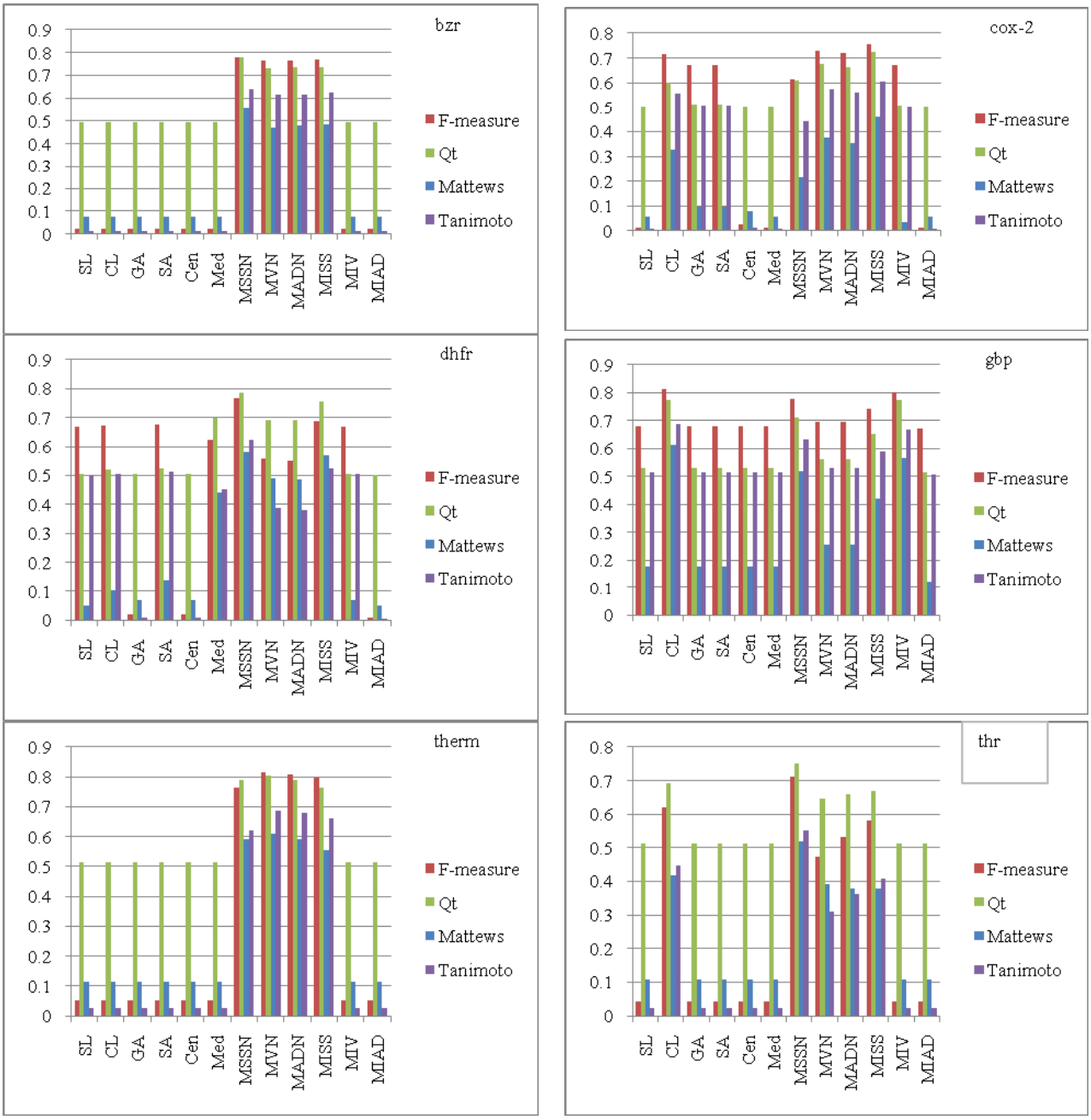
Un análisis preliminar, basado en la inspección visual de los dendogramas generados por los 12 AC2 sobre los ocho CDs permitió detectar algunas regularidades en el comportamiento de dichos algoritmos (ver Anexos A1-A25 como ejemplos). Los algoritmos SL y MIAD presentan fuerte tendencia a la concatenación en todos los casos; GA y SA no recuperan apropiadamente la estructura de los datos en la mayoría de los casos, excepto para el conjunto de datos ACE, donde se aprecia una buena resolución de los casos para un tamaño dos de conglomerados; seguidamente, Cen y Med presentan serios reversos (falla la monotonicidad) en sus árboles de aglomeración; por su parte, MIV presenta una tendencia a la concatenación y reversos en la mayoría de los casos, excepto para el conjunto de datos GBP, donde también se observa una buena resolución cuando se consideran dos conglomerados. Sin embargo, los algoritmos CL, MSSN, MVN, MIDN, ISS son capaces de resolver claramente cada uno de los CDs. Como tendencia general, a distancias relativamente pequeñas ya se han fusionado todos los casos repartidos en dos grandes, que posteriormente se unen a una distancia relativamente grande en comparación con los niveles iniciales de fusión. La presencia de reveros en los algoritmos Cen y Med se puede explicar teóricamente porque los mismos fallan en cumplir las condiciones de monotonicidad para algoritmos del tipo d-AC2 [71, 72]. Similarmente, los algoritmos MIV y MIAD fallan en cumplir las condiciones de monotonicidad para algoritmos del tipo h-AC2 [26, 73]. La tendencia a la

concatenación del SL y la incapacidad de los algoritmos GA y SA de lidiar con conglomerados de forma diferente a la hiperesférica también se ha resaltado en la literatura especializada [74], lo cual pudiera estar relacionado con nuestros hallazgos. Como explicación adicional *a posteriori* la resolución de cada CDs en dos grandes grupos parece justificar la dicotomización de las variables farmacocinéticas pIC50 y pKi hechas empíricamente en estudios QSAR anteriores [56].

Para corroborar cuantitativamente el análisis anterior, se calcularon y graficaron las medidas subjetivas F-measure (F-measure), exactitud global (Qt), el coeficiente de correlación de Mattews (Mattews) y el coeficiente de Tanimoto (Tanimoto) para evaluar la calidad externa de las aglomeraciones obtenidas. Los Gráficos 3.3 muestran el comportamiento de dichas medidas. Se puede observar claramente que Qt y Mattews explican estable y coherentemente las tendencias observadas anteriormente en los dendogramas, un valor de Qt por encima de 0,5 indica una asignación no aleatoria de los casos a las particiones, la significación de Mattews se puede evaluar directamente mediante su relación con el estadístico chi-cuadrado $\chi^2 = (Mattews)^2 * N$, donde N es el tamaño del CDs. Por otra parte, las medidas F-measure y Tanimoto se comportan inestablemente a través de los CDs, sobrevalorando el comportamiento de los peores algoritmos en la mitad de los casos (ver ace, cox-2, dhfr, gbp). Una vez más se evidenció que SL, GA, SA, Cen, Med, MIV y MIAD son los peores algoritmos frente a CL, MSSN, MVN, MADN y MIV como los mejores para la problemática tratada.







Gráficos 3.3. Comportamiento de las medidas de calidad externa de las aglomeraciones obtenidas con los doce AC2 sobre los conjuntos de datos descritos anteriormente

En consecuencia con los resultados anteriores, se realizó una prueba de comparación múltiple entre estos cinco últimos algoritmos, empleando Qt como variable aleatoria, a través de un análisis de varianza de dos criterios de Friedman. También se compararon los valores Qt promediados para cada CDs con los valores promedios reportados por otros investigadores en estudios donde se emplearon los mismos CDs pero métodos más

sofisticados de QSAR y del Aprendizaje Automático, los datos y resultados se muestran en la **Tabla 3.4.**

Tabla 3.4. Datos para la comparación múltiple entre los mejores AC2 y con otros algoritmos empleados en QSAR y Aprendizaje Automático

	Datos I[a]					**Datos II[b]**			
CDs	**CL**	**MSSN**	**MVN**	**MADN**	**MISS**	**AC2[c]**	**Bruce[d]**	**Sönströd[e]**	**Johansson[f]**
ace	88,60	88,60	87,72	88,60	87,72	88,25	87,84	84,53	90,12
ache	55,86	51,35	55,86	55,86	76,58	59,10	73,33	64,80	70,22
bzr	49,69	77,91	73,01	73,62	73,62	69,57	76,44	72,17	75,56
cox-2	59,94	60,87	67,39	66,46	72,36	65,40	75,30	70,30	74,92
dhfr	52,14	78,59	69,27	69,02	75,57	68,92	82,17	76,50	80,67
gbp	77,27	71,21	56,06	56,06	65,15	65,15	74,47	65,40	76,07
therm	51,32	78,95	80,26	78,95	76,32	73,16	70,37	64,40	69,57
thr	69,32	75,00	64,77	65,91	67,05	68,41	68,77	61,67	72,90
Fried[g]			0,481					**0,002****	
RP[h]	2,50	3,69	2,63	2,81	3,38	1,75	3,38	1,63	3,25
Wilcon[i]	AC2-Bruce			**0,039***		Bruce-Sönströd			**0,004****
	AC2-Sönströd			0,473		Bruce-Johansson			0,473
	AC2-Johansson			**0,012***		Sönströd-Johansson			**0,004****

[a]Valores de Qt para la comparación múltiple entre los mejores AC2 obtenidos; [b]Valores de Qt para la comparación múltiple de los ACs con otros métodos reportados por otros autores; [c]Se tomó el valor promedio de los AC2 en cada CDs; [d]Valores promedios de Qt reportados por este autor de los métodos estudiados en cada CDs, los métodos estudiados fueron: tree, bagged tree, boosted tree, random forest, SVM, tunned forest, tunned SVM [56]; [e]Valores promedios de Qt reportados por este autor de los métodos estudiados en cada CDs, los métodos estudiados fueron: RIPPER o JRip (decision lists), C4.5 o J48 (decision tres), Chipper (decision lists) [55]; [f]Valores promedios de Qt reportados por este autor de los métodos estudiados en cada CDs, los métodos estudiados fueron: MLP, RBF, SVM, Bag-M_W, Bag-RBF, Bag-M_B, Avg-M_A, GAS, NB neural networks [54]; [h]Significación exacta para la prueba χ^2 de comparación múltiple de Friedman; [g]Rango medio asignado a cada algoritmo por la prueba de Friedman, mientras más grande mejor es el algoritmo; [i] Significación exacta para la prueba de Wilcoxon de una cola para la comparación *post hoc* de pares de métodos; *Pruebas significativa ($p < 0,05$), **pruebas muy significativas ($p < 0,01$)

A partir de los resultados anteriores se pueden hacer observaciones interesantes. Del valor de la significación de la prueba de Friedman para los **Datos I** ($p > 0,05$) se puede inferir que no existen diferencias significativas en cuanto a las medianas de los puntajes Qt de los algoritmos AC2 sobre los CDs estudiados. Este resultado es alentador, pues muestra ***tres*** algoritmos novedosos con un rendimiento similar al algoritmo de Ward (MISS), el cual es

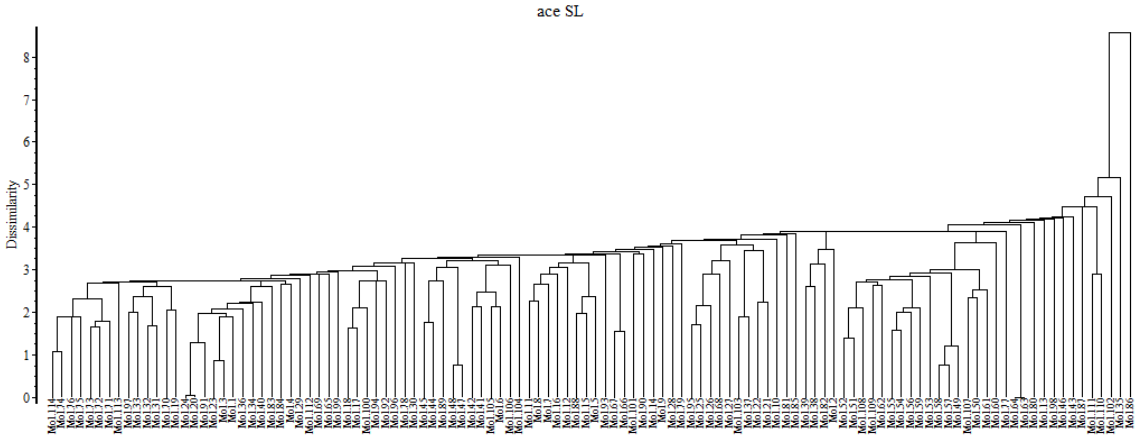
considerado por la comunidad científica como el algoritmo de elección en problemas quimioinformáticos [33]. Por otra parte, resultados análogos obtenidos para los **Datos II** nos permiten inferir que existen diferencias significativas ($p < 0,01$) entre las medianas de los métodos contrastados. La prueba *post hoc* de Wilcoxon nos permitió encontrar que las diferencias anteriores pueden ser atribuibles a la formación de dos grupos homogéneos ($p > 0,05$), AC2-Sönströd y Bruce-Johansson, que presentan diferencias significativas entre sí ($p < 0,01$). Esto significa que, en términos promedios, los algoritmos AC2 rinden igual de bien que algunos algoritmos del aprendizaje automático como algunas técnicas simples de árboles y listas de decisión, mientras que son superados por técnicas de multi clasificación que emplean árboles de decisión, máquinas de soporte vectorial y redes neuronales, estas dos últimas se consideran como las más potentes y prometedoras dentro de las técnicas supervisadas en problemas quimioinformáticos [75, 76]. Si tenemos en cuenta que los resultados anteriores se obtuvieron empleando descriptores (descriptores de Sutherland *et al.*) distintos a los nuestros, entonces habría que analizarlos en el contexto representación-método; de esta forma, dichos resultados constituyen una referencia invaluable de la calidad de los descriptores seleccionados y de los mejores algoritmos de conglomerados analizados en este trabajo.

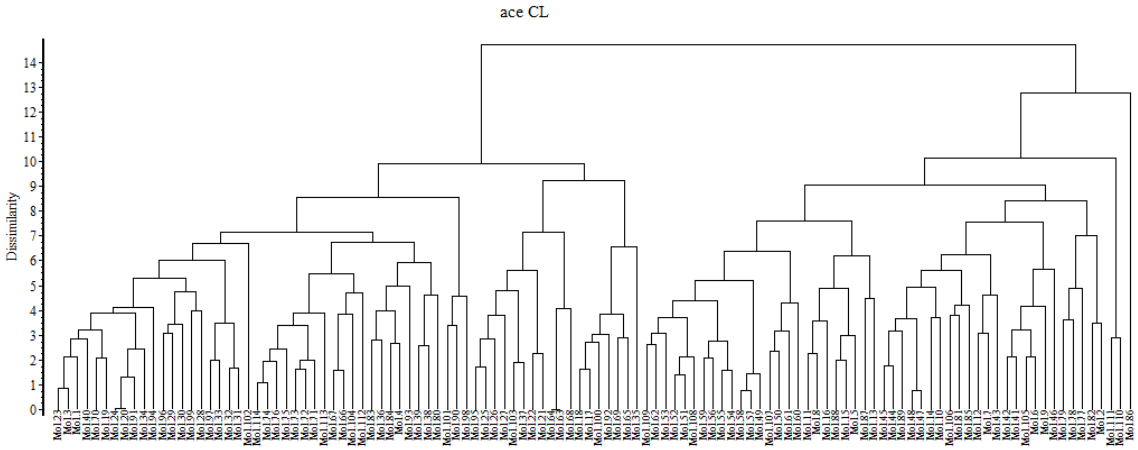
Conclusiones

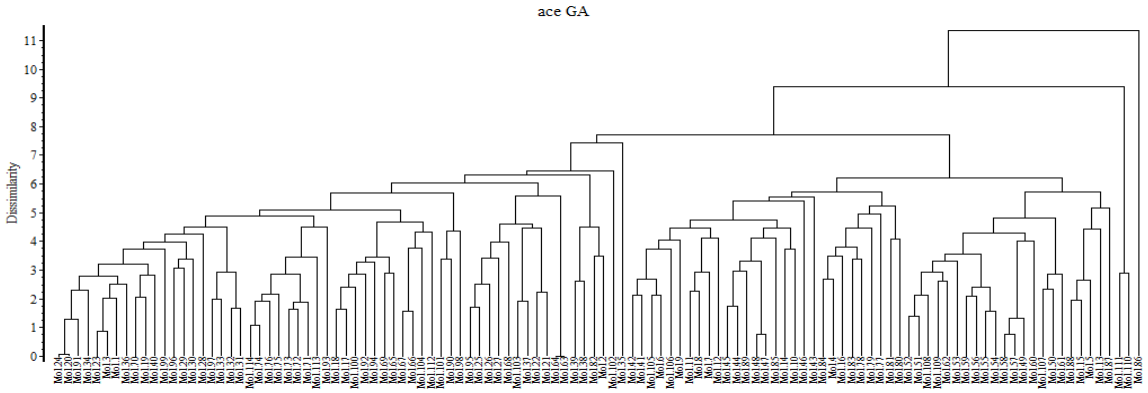
En el trabajo se abordó una metodología para la comparación de algoritmos combinatorios SAHN, usando para ello bases de datos de la Química Medicinal, cuyos elementos se representaron por descriptores moleculares de naturaleza numérica real, seleccionados mediante técnicas del Aprendizaje Automático *(Machine Learning)*. A través de la técnica de selección de rasgos CfsSubsetEval implementada en el Weka, aplicada a ocho conjuntos de datos de la Química Medicinal, se seleccionaron los rasgos linealmente más relevantes a cada una de las actividades farmacológicas. Como tendencia promedio, las familias de descriptores 2D y 3D atraparon mayoritariamente la información química directamente relacionada con dicho contexto farmacológico. Específicamente, las familias 2D autocorrelations (2D), RDF descriptors (3D), 3D-Morse descriptors (3D) y WHIM descriptors (3D) atraparon la información química implicada en estos fenómenos de interacción ligando-receptor. De los doce algoritmos comparados, cuatro presentan un rendimiento similar al de Ward, tres de los cuales son completamente novedosos en la Quimioinformática. Para trabajos futuros se planea ampliar el estudio comparativo para abarcar además los nuevos métodos ideados en nuestro grupo, así como ampliar el dominio de aplicación de los mismos para tener en cuenta otras bases de datos QSPR, QSAR, toxicidad, metabolismo, de permeabilidad, docking, mecanísticas, así como conjuntos generados por técnicas de simulación estadística para lograr inferencias estadísticas más robustas acerca del comportamiento y calidad de los algoritmos comparados.

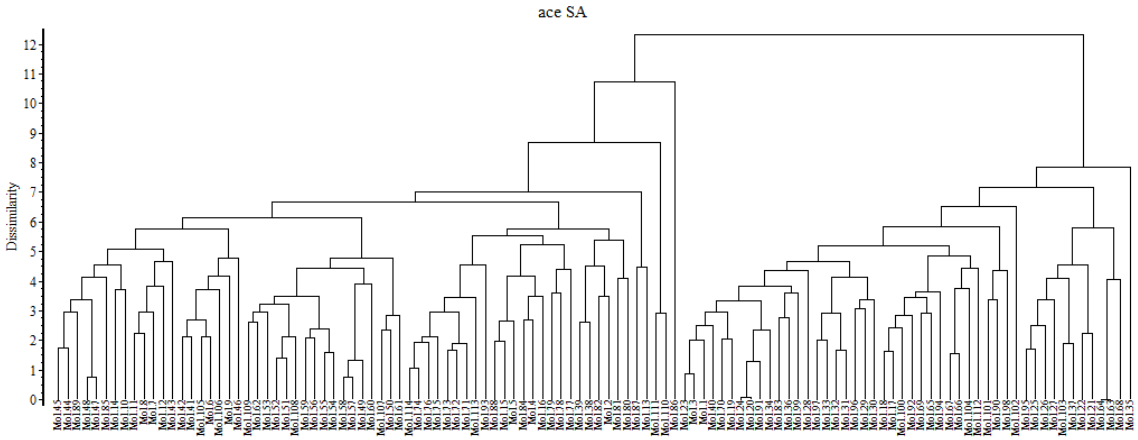
Anexos

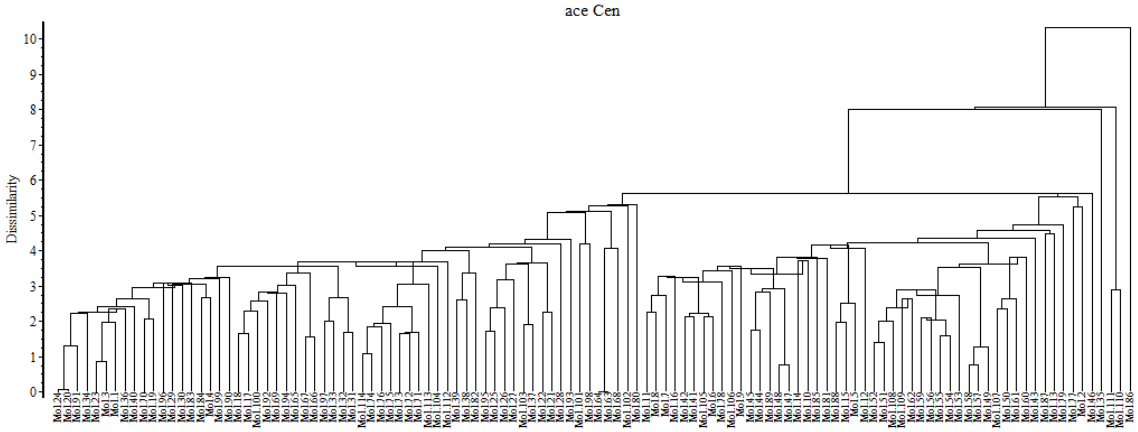
Figuras A1-A25. Dendogramas asociados a los algoritmos SL, CL, GA, SA, Cen, Med MSSN, MVN, MIDN, ISS, MIV y MIAD para los conjuntos de datos ACE (1-12) y THR (13-
25)

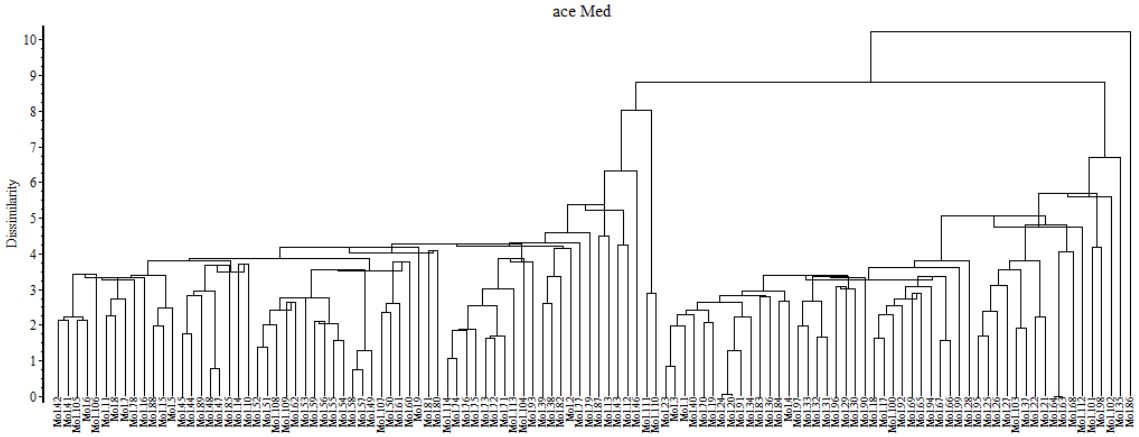


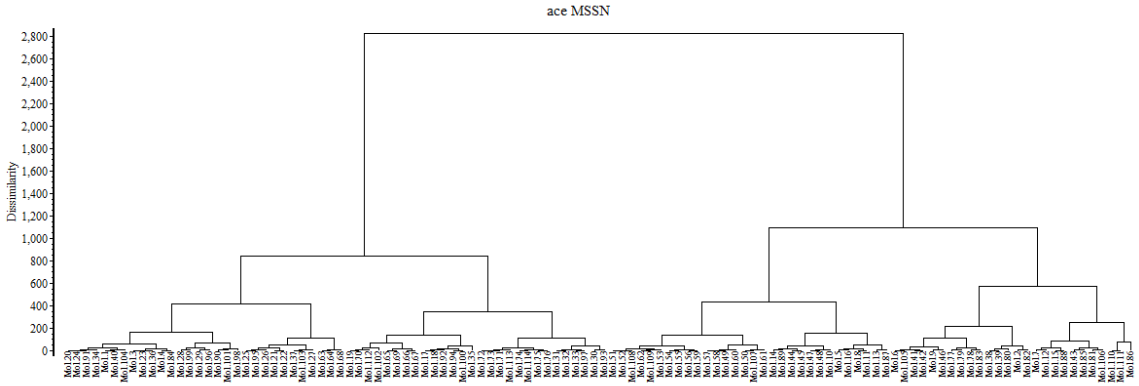


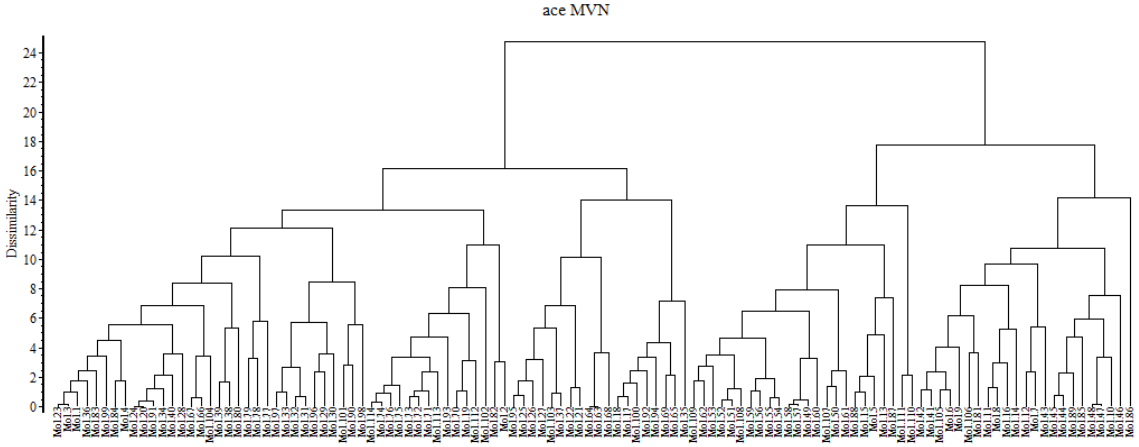

ace GA
Dissimilarity

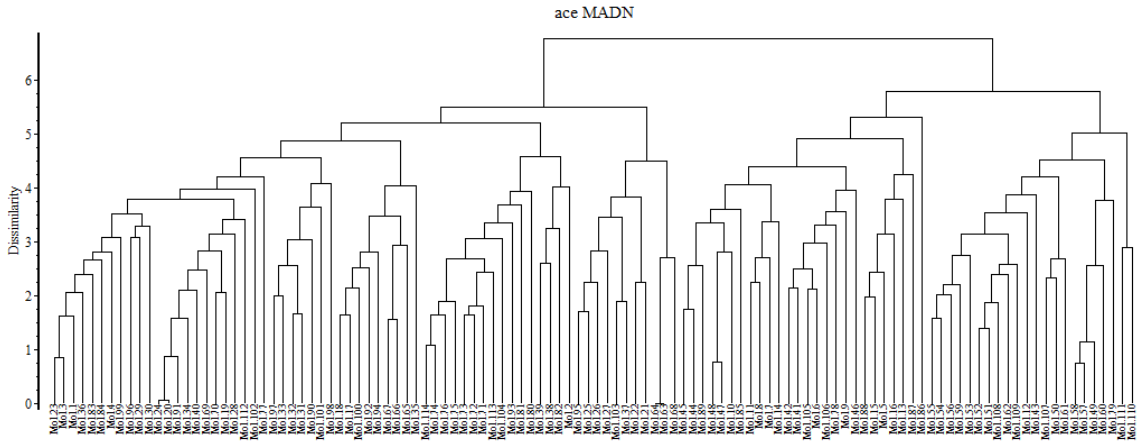

ace SA
Dissimilarity

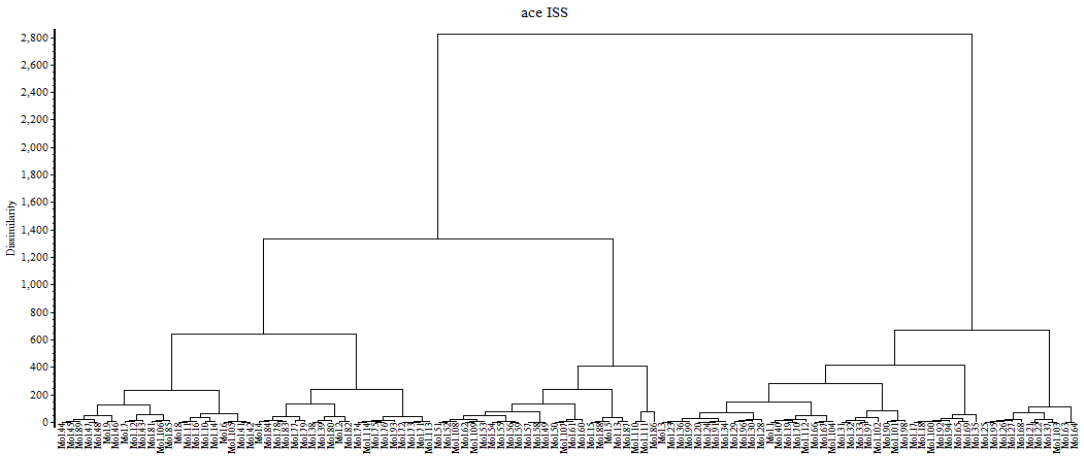

ace Cen
Dissimilarity

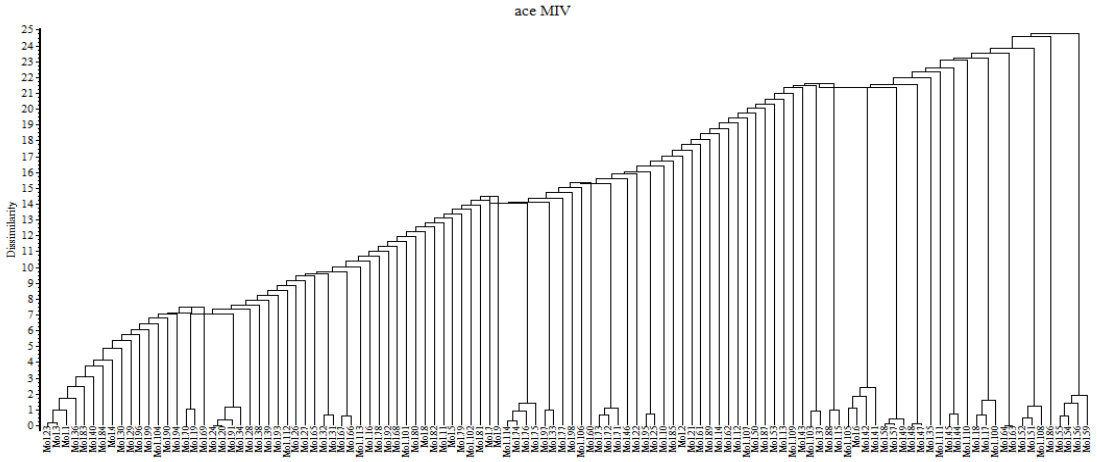

ace Med
Dissimilarity
0
1
2
3
4
5
6
7
8
9
10

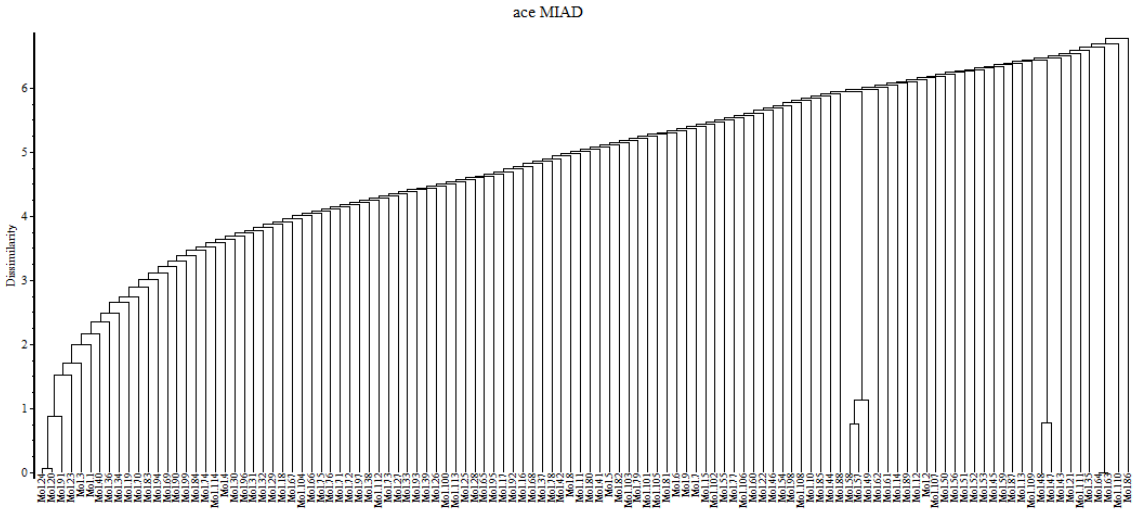

ace MSSN
Dissimilarity
0
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,200
2,400
2,600
2,800

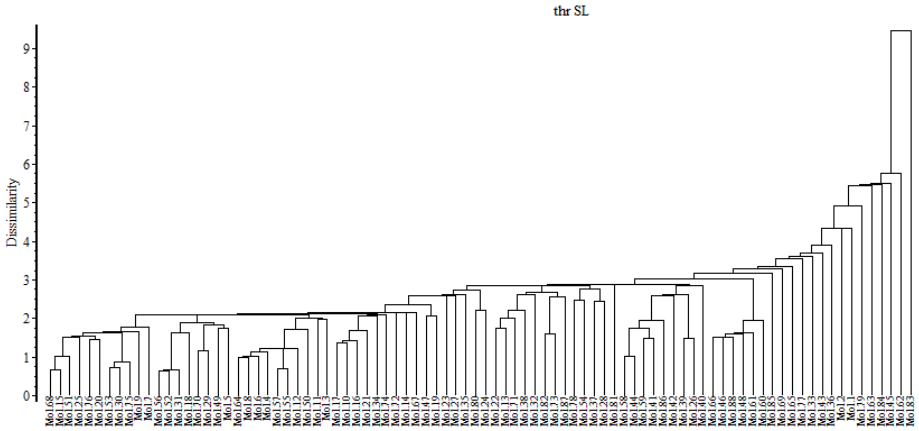

ace MVN
Dissimilarity
0
2
4
6
8
10
12
14
16
18
20
22
24

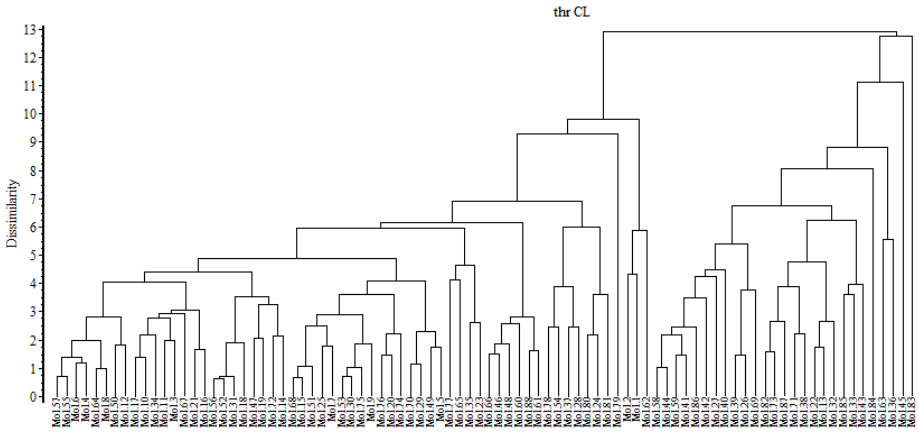

ace MADN
Dissimilarity
0
1
2
3
4
5
6

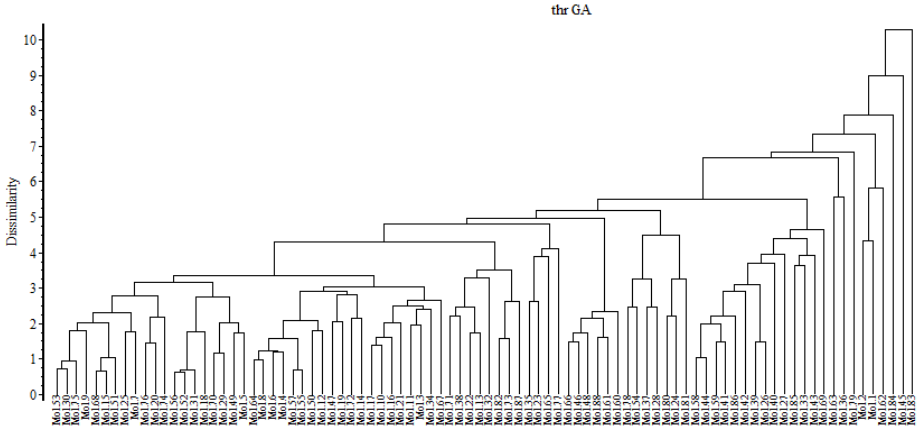

ace ISS
Dissimilarity
0
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,200
2,400
2,600
2,800

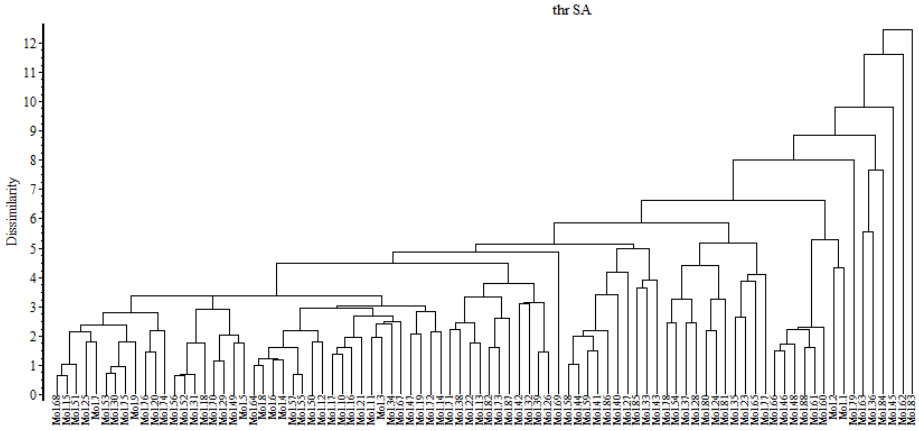

ace MIV
Dissimilarity

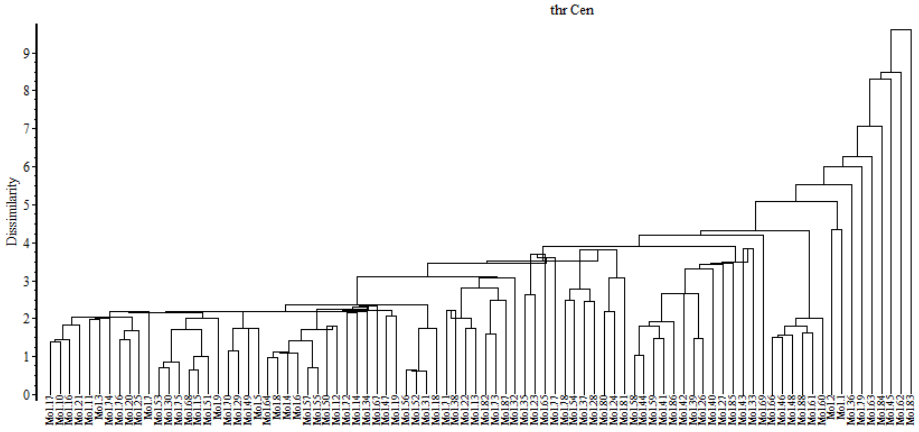

ace MIAD
Dissimilarity

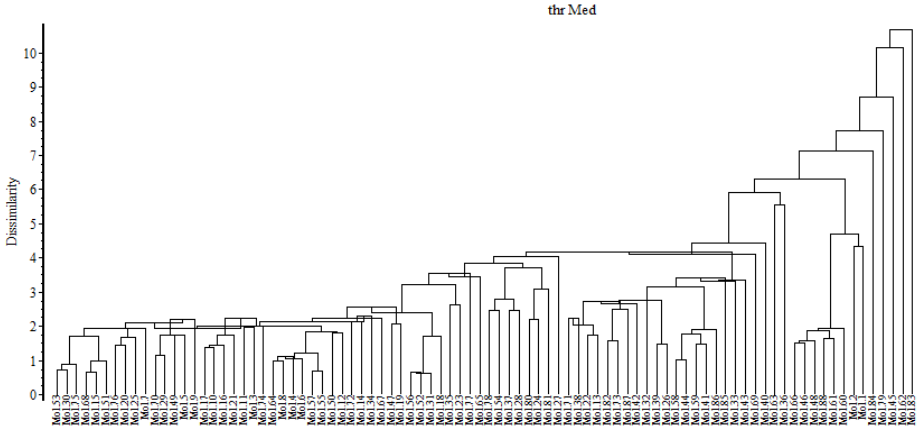

thr SL
Dissimilarity

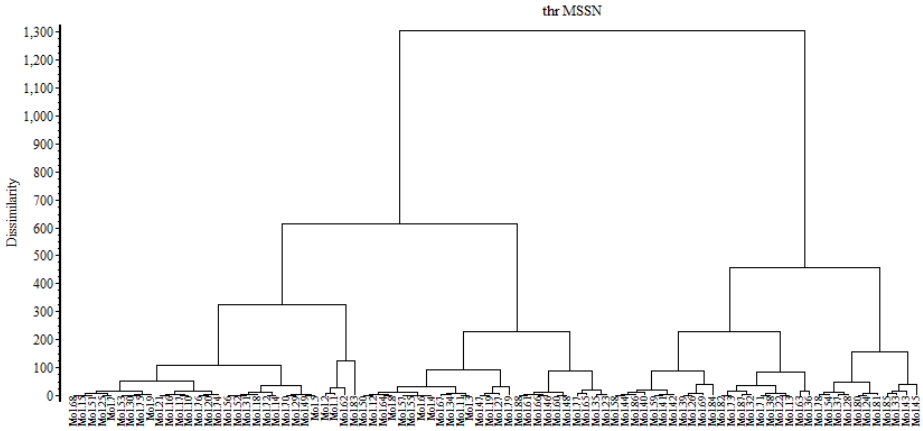

thr CL
Dissimilarity

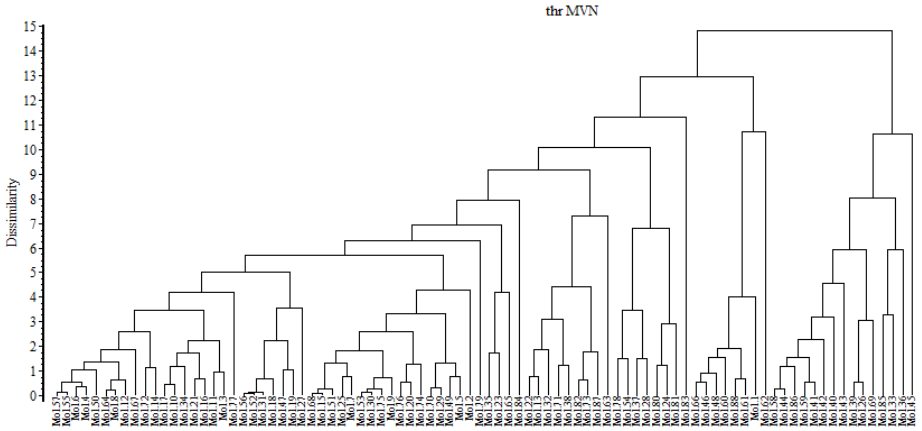

thr GA
Dissimilarity

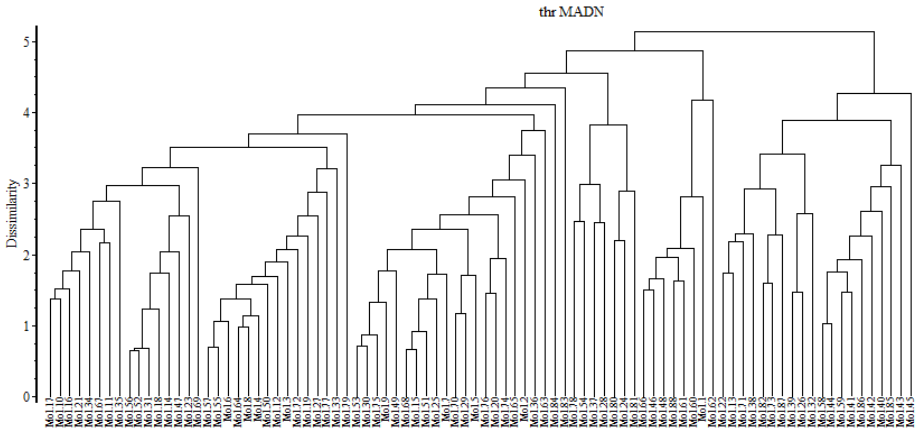

thr SA
Dissimilarity

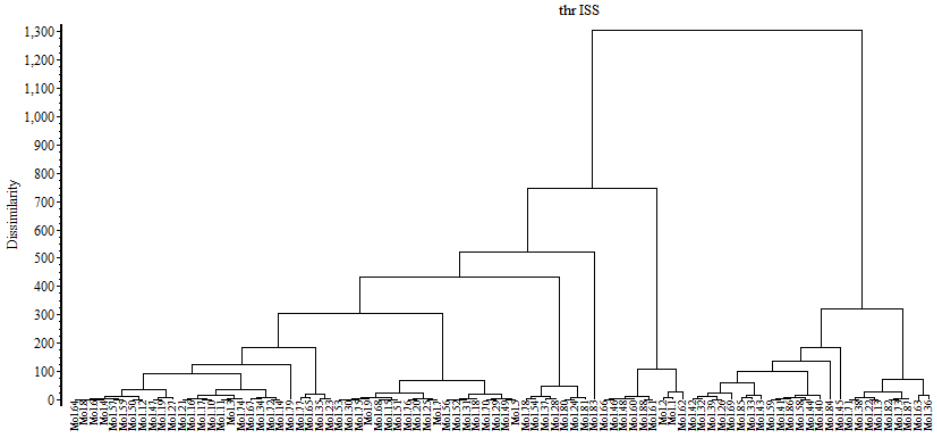

thr Cen
Dissimilarity

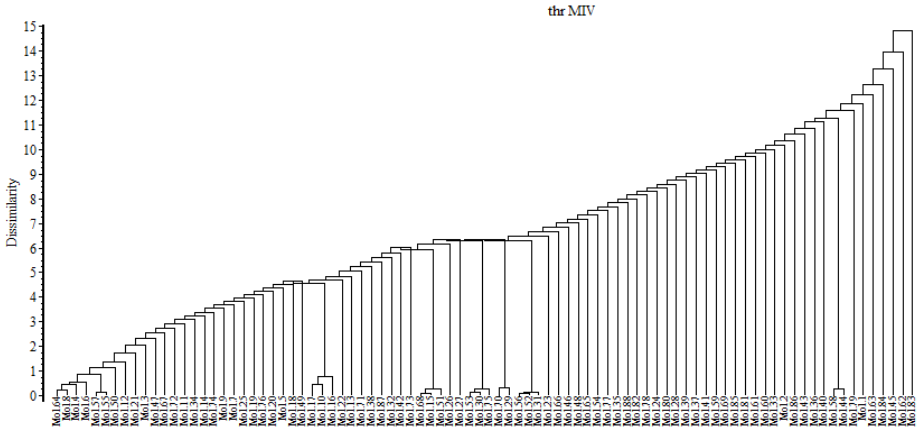

thr Med
Dissimilarity
0
1
2
3
4
5
6
7
8
9
10

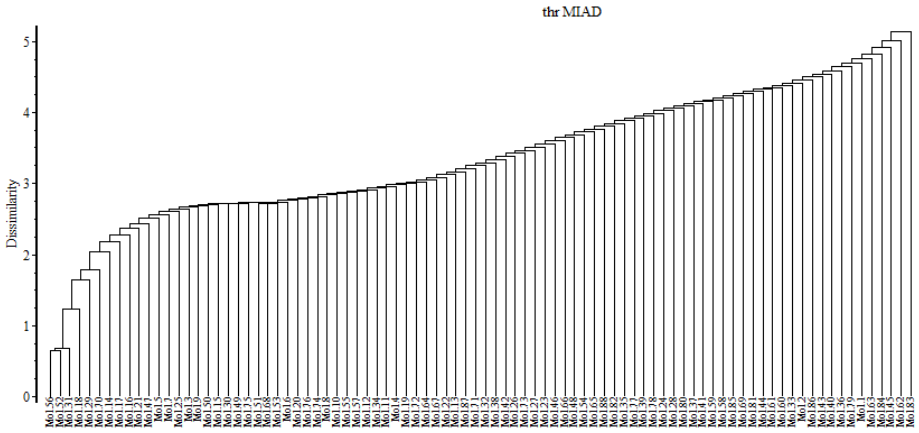

thr MSSN
Dissimilarity
0
100
200
300
400
500
600
700
800
900
1,000
1,100
1,200
1,300


thr MVN
Dissimilarity
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15


thr MADN
Dissimilarity
0
1
2
3
4
5


thr ISS
Dissimilarity
0
100
200
300
400
500
600
700
800
900
1,000
1,100
1,200
1,300


thr MIV
Dissimilarity
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15


thr.MIAD
Dissimilarity
0
1
2
3
4
5

Referencias Bibliográficas

1. Shemetulskis, N.E., Dunbar, J.B., Dunbar, B.W., et al., "Enhancing the Diversity of a Corporate Database Using Chemical Database Clustering and Analysis", *J. Comput-Aided Mol. Des.*, vol. 9, pp. 407-416, 1995.
2. Barnard, J.M. and Downs, G.M., "Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures", *J. Chem. Inf. Comput. Sci.*, vol. 32, pp. 644-649, 1992.
3. Martin, E.J., Blaney, J.M., Spellmeyer, D.C., et al., "Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery", *J. Med. Chem.*, vol. 38, pp. 1431-1436, 1995.
4. Sadowski, J., Wagener, M., and Gasteiger, J., "Assessing Similarity and Diversity of Combinatorial Libraries By Spatial Autocorrelation Functions and Neural Networks", *Agnew. Chem., Int. Ed. Engl.* , vol. 34 (23), pp. 2674-2677, 1995.
5. Martin, E.J. and Critchlow, R.E., "Beyond Mere Diversity: Tailoring Combinatorial Libraries for Drug Discovery", *J. Comb. Chem.*, vol. 1, pp. 32-45, 1999.
6. Warr, W.A., "Combinatorial Chemistry and Molecular Diversity. An Overview", *J. Chem. Inf. Comput. Sci.*, vol. 37, pp. 134-140, 1997.
7. Menard, P.R., Mason, J.S., Morize, I., et al., "Chemistry Space Metrics in Diversity Analysis, Library Design, and Compound Selection", *J. Chem. Inf. Comput. Sci.* , vol. 38, pp. 1204-1213, 1998.
8. Turner, D.B., Tyrell, S.M., and Willett, P., "Rapid Quantification of Molecular Diversity for Selective Database Acquisition", *J. Chem. Inf. Comput. Sci.*, vol. 37, pp. 18-22, 1997.
9. Ashton, M.J., Jaye, M.C., and Mason, J.S., "New Perspectives in Lead Generation ii: Evaluating Molecular Diversity", *Drug Discovery Today*, vol. 1, pp. 71-78, 1996.
10. Engels, M.F.M., Thielemans, T., Verbinnen, D., et al., "Cerberus: A System Supporting the Sequential Screening Process", *J. Chem. Inf. Comput. Sci.*, vol. 40, pp. 241-245, 2000.

11. Rhodes, N., Willett, P., Dunbar, J.B., et al., "Bitstring Methods for Selective Compound Acquisition", *J. Chem. Inf. Comput. Sci.*, vol. 40, pp. 210-214, 2000.
12. MDL Information Systems, http://www.mdli.com/.
13. Daylight Chemical Information Systems, http://www.daylight.com.
14. Barnard Chemical Information, http://www.bci1.demon.co.uk/.
15. Brown, R.D. and Martin, Y.C., "Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection", *J. Chem. Inf. Comput. Sci.*, vol. 36, pp. 572-584, 1996.
16. Wild, D.J. and Blankley, C.J., "Comparison of 2D Fingerprint Types and Hierarchy Level Selection Methods for Structural Grouping Using Wards Clustering", *J. Chem. Inf. Comput. Sci.* , vol. 40, pp. 155-162, 2000.
17. Arco-García, L., "Agrupamiento basado en el concepto de intermediación diferencial y la aplicación de la teoría de los conjuntos aproximados para valorar resultados de agrupamientos", *Computer Science Department*, UCLV, 2008.
18. Downs, G.M. and Barnard, J.M., "Clustering Methods and Their Uses in Computational Chemistry", *Reviews in Computational Chemistry*, vol. 18, pp. 41, 2002.
19. Wishart, D., "An Algorithm for Hierarchical Classifications", *Biometrics*, vol. 25, (1), pp. 165-170, 1969.
20. Ward, J.H., "Hierarchical Grouping to Optimize an Objetive Function", *Journal of the American Statistical Association*, vol. 58, (301), pp. 9, 1963.
21. Lance, G.N. and Williams, W.T., "A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems", *The Computer Journal*, vol. 9, (4), pp. 373-380, 1967.
22. Downs, G.M. and Barnard, J.M., "Clustering Methods and Their Uses in Computational Chemistry", en *Reviews in Computational Chemistry* ed., Vol. 18, Eds. K.B. Lipkowitz and D.B. Boyd:Wiley-VCH, John Wiley and Sons, Inc., 2002.
23. Anderberg, M.R., *Cluster analysis for applications*. New York:Wiley, 1973.
24. Jambu, M. and Lebeaux, M.O., *Classification automatique pour l'analyse des données (1. Méthodes et algorithmes. 2. Logiciels)*. Paris, France: Dunod, 1978.

25. Jambu, M. and Lebeaux, M.O., *Cluster analysis and data analysis*. Amsterdam, Oxford: North-Holland 1983.
26. Podani, J., "New combinatorial clustering methods", *Vegetatio*, vol. 81, (1-2), pp. 61-77, 1989.
27. Murtagh, F., "A Survey of Recent Advances in Hierarchical Clustering Algorithms", *The Computer Journal* vol. 26, (4), pp. 354-359, 1983.
28. Podani, J., "SYN-TAX2000. Computer Programs for Data Analysis in Ecology and Systematics. User's Manual ", *Scientia Publishing, Budapest*, 2001.
29. Maldonado, A.G., Doucet, J.P., Petitjean, M., et al., "Molecular similarity and diversity in chemoinformatics: From theory to applications", *Molecular Diversity*, vol. 10, pp. 39-79, 2006.
30. Sheridan, R.P. and Kearsley, S.K., "Why do we need so many chemical similarity search methods?" *Drug Discov. Today*, vol. 7, (17), pp. 903-911, 2002.
31. Jónsdóttir, S.Ó., Jørgensen, F.S., and Brunak, S., "Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates", *Bioinformatics*, vol. 21 (10), pp. 2145-2160, 2005.
32. Bender, A., "Compound bioactivities go public", *Nature Chemical Biology*, vol. 6 pp. 309, 2010.
33. Willett, P., "Searching Techniques for Databases of Two- and Three-Dimensional Chemical Structures", *Journal of Medicinal Chemistry*, vol. 48, (13), pp. 4183-4199, 2005.
34. Sutherland, J.J., O'Brien, L.A., and Weaver, D.F., "A Comparison of Methods for Modeling Quantitative Structure−Activity Relationships", *Journal of Medicinal Chemistry*, vol. 47, (22), pp. 5541-5554, 2004.
35. Bruce, C.L., Melville, J.L., Pickett, S.D., et al., "Contemporary QSAR Classifiers Compared", *Journal of Chemical Information and Modeling*, vol. 47, (1), pp. 219-227, 2007.
36. Johansson, U., Löfström, T., and Norinder, U., "Evaluating Ensembles on QSAR Classification", in *3rd Skövde Workshop on Information Fusion Topics*, Eds. R. Johansson, J. van Laere, and J. Mellin: Univeristy of Skövde, 2009, pp. 49-54.

37. Sönströd, C., Johansson, U., and Norinder, U., "Generating Comprehensible QSAR Models", in *3rd Skövde Workshop on Information Fusion Topics 2009*, Eds. R. Johansson, J. van Laere, and J. Mellin, Skövde, Sweden: University of Skövde, 2009.
38. Culp, M., Johnson, K., and Michailidis, G., "The Ensemble Bridge Algorithm: A New Modeling Tool for Drug Discovery Problems", *Journal of Chemical Information and Modeling*, vol. 50, (2), pp. 309-316, 2010.
39. Johnson, M.A., "A review and examination of mathematical spaces underlying molecular similarity analysis", *J. Math. Chem.* , vol. 3, pp. 117-145, 1989
40. Maggiora, G.M. and Shanmugasundaram, V., "Molecular Similarity Measures", en *Chemoinformatics*ed., Vol. 275, Eds. J. Bajorath: Humana Press, 2004.
41. Agrafiotis, D.K., Bandyopadhyay, D., Wegner, J.K., et al., "Recent Advances in Chemoinformatics", *J. Chem. Inf. Model.*, vol. 47, pp. 1279-1293, 2007.
42. Wegner, J.K., Fröhlich, H., Mielenz, H.M., et al., "Data and Graph Mining in Chemical Space for ADME and Activity Data Sets", *QSAR Comb. Sci.*, vol. 25, (3), pp. 205-220, 2006.
43. Bender, A. and Glen, R.C., "Molecular similarity: a key technique in molecular informatics", *Org. Biomol. Chem.*, vol. 2, pp. 3204-3218, 2004.
44. Janecek, A.G.K. and Gansterer, W.N., "On the Relationship Between Feature Selection and Classification Accuracy", in *JMLR: Workshop and Conference Proceedings* 2008, pp. 90-105.
45. Steinbach, M., Ertöz, L., and Kumar, V., "The Challenges of Clustering High Dimensional Data", 2000.
46. Böcker, A., Schneider, G., and Teckentrup, A., "Status of HTS Data Mining Approaches", *QSAR Comb. Sci.*, vol. 23, pp. 207-213, 2004.
47. Glen, R.C. and Adams, S.E., "Similarity Metrics and Descriptor Spaces – Which Combinations to Choose?" *QSAR Comb. Sci.* , vol. 25, (12), pp. 1133-1142, 2006.
48. Sadowski, J. and Schwab, C.H., "3D Structure Generator CORINA", *3D Structure Generator CORINA*, Generation of High-Quality Three-Dimensional Molecular Models, 2008.

49. Talete srl, "DRAGON for Windows ", *DRAGON for Windows.* Software for Molecular Descriptors Calculations, 2007.
50. Microsoft Corporation, "Microsoft ® Office Excel ®, 2006.
51. Hall, M., Frank, E., Holmes, G., et al., "The WEKA Data Mining Software: An Update", *SIGKDD Explorations*, vol. 11, (1), 2009.
52. Hall, M.A., "Correlation-based Feature Subset Selection for Machine Learning",*Department of Computer Science*, The University of Waikato, 1998.
53. McGraw, K.O. and Wong, S.P., "Forming Inferences About Some Intraclass Correlation Coefficients", *Psychological Methods*, vol. 1, (1), pp. 30-46, 1996.
54. Johansson, U., Löfström, T., and Norinder, U., "Evaluating Ensembles on QSAR Classification", in *3rd Skövde Workshop on Information Fusion Topics 2009*, Eds. R. Johansson, J. van Laere, and J. Mellin, Skövde, Sweden: Univeristy of Skövde, 2009 pp. 49-54.
55. Sönströd, C., Johansson, U., and Norinder, U., "Generating Comprehensible QSAR Models", in *3rd Skövde Workshop on Information Fusion Topics 2009*, Eds. R. Johansson, J. van Laere, and J. Mellin, Skövde, Sweden: University of Skövde, 2009
56. Bruce, C.L., Melville, J.L., Pickett, S.D., et al., "Contemporary QSAR Classifiers Compared", *J. Chem. Inf. Model.*, vol. 47, pp. 219-227, 2007.
57. Stein, B., Meyer zu Eissen, S., and Wißbrock, F., "On Cluster Validity and the Information Need of Users", in *3rd IASTED Int. Conference on Artificial Intelligence and Applications (AIA 03)*, Eds. M.H. Hanza, Benalmádena, Spain: ACTA Press, ©IASTED 2003, 2003, pp. 216-221.
58. Baldi, P., Brunak, S., Yves, C., et al., "Assesing the accuracy of prediction algorithms for classification: an overview", *BIOINFORMATICS*, vol. 16, (5), pp. 412-424, 2000.
59. Willett, P., "Similarity-based virtual screening using 2D fingerprints", *Drug Discovery Today*, vol. 11, (23/24), pp. 1046-1053, 2006.

60. Wolpert, D.H., "The supervised learning no-free-lunch Theorems", in *Proc. 6th Online World Conference on Soft Computing in Industrial Applications*, 2001, pp. 25-42.
61. Grünwald, P.D., "Minimum Description Length Tutorial", en *Advances in Minimum Description Length. Theory and Applications*ed., Eds. P.D. Grünwald, I.J. Myung, and M.A. Pitt: Cambridge, Mass. : MIT Press, ©2005., 2005.
62. Piël, N., "No-Free-Lunch and the Minimum Description Length", *No-Free-Lunch and the Minimum Description Length*, vol., pp. 2, 2007.
63. Kruskal, W.H. and Wallis, W.A., "Use of Ranks in One-Criterion Variance Analysis", *Journal of the American Statistical Association*, vol. 47, (260), pp. 583-621, 1952
64. Conover, W.J. and Iman, R.L., "Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics", *The American Statistician*, vol. 35, (3), pp. 124-129, 1981.
65. Friedman, M., "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance", *Journal of the American Statistical Association*, vol. 32, (200), pp. 675-701, 1937.
66. Demšar, J., "Statistical Comparisons of Classifiers over Multiple Data Sets", *Journal of Machine Learning Research* vol. 7, pp. 1-30, 2006.
67. García, S. and Herrera, F., "An Extension on “Statistical Comparisons of Classifiers over Multiple Data Sets” for all Pairwise Comparisons", *Journal of Machine Learning Research* vol. 9 pp. 2677-2694, 2008
68. Iman, R.L. and Davenport, J.M., "Approximations of the critical region of the Friedman statistics ", *Communications in Statistics - Theory and Methods*, vol. 9, (6), pp. 571-595, 1980.
69. Kendall, M.G. and Smith, B.B., "The Problem of m Rankings", *The Annals of Mathematical Statistics*, vol. 10, (3), pp. 275-287, 1939.
70. Fagot, R.F., "An ordinal coefficient of relational agreement for multiple judges", *Psychometrika*, vol. 59, (2), pp. 241-251, 1994.
71. Milligan, G.W., "Ultrametric hierarchical clustering algorithms", *Psychometrika* vol. 44, pp. 343-346, 1979.

72. Batagelj, V., "Note on ultrametric clustering algorithms", *Psychometrika* vol. 46, pp. 351-352, 1981.
73. Diday, E., "Inversions en classification hiérarchique: application á la construction adaptive d'indices d'agrégation. " *Rev. Stat. Appl.*, vol. 31, pp. 45-62, 1983.
74. Downs, G.M. and Barnard, J.M., "Clustering Methods and Their Uses in Computational Chemistry", en *Reviews in Computational Chemistry* ed., Vol. 18, Eds. K.B. Lipkowitz and D.B. Boyd:Wiley-VCH, John Wiley and Sons, Inc. , 2002.
75. Ivanciuc, O., "Applications of Support Vector Machines in Chemistry", en *Reviews in Computational Chemistry*ed., Vol. 23, Eds. K.B. Lipkowitz and T.R. Cundari:Wiley-VCH, Weinheim, 2007.
76. Eckert, H. and Bajorath, J., "Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches", *Drug Discovery Today* vol. 12, (5/6), pp. 225-233, 2007.