Asociación Argentina
de Mecánica Computacional

Mecánica Computacional Vol XXV, pp. 993-1011
Alberto Cardona, Norberto Nigro, Victorio Sonzogni, Mario Storti. (Eds.)
Santa Fe, Argentina, Noviembre 2006

ANÁLISIS NUMÉRICO DE DIFERENTES CRITERIOS DE SIMILITUD EN ALGORITMOS DE CLUSTERING

Axel J. Soto[a,b], Ignacio Ponzoni[a,b] y Gustavo E. Vazquez[b]

[a]Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC)
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur
Av. Alem 1253 – 8000 - Bahía Blanca
ARGENTINA
[b]Planta Piloto de Ingeniería Química (PLAPIQUI)
Universidad Nacional del Sur - CONICET
Complejo CRIBABB – Camino La Carrindanga km. 7 – CC 717 - Bahía Blanca
ARGENTINA
asoto@plapiqui.edu.ar; ip@cs.uns.edu.ar; gvazquez@criba.edu.ar;

Palabras Clave: Análisis de agrupamiento, quimioinformática, redes neuronales, logP

Resumen. En el presente trabajo se analizan diferentes metodologías y criterios para realizar análisis de agrupamiento sobre datos multivariados. El análisis de agrupamiento tiene por objetivo formar grupos de elementos, de manera tal que los pertenecientes a un mismo grupo sean parecidos entre sí y distintos a los miembros de los restantes grupos. Se describen consideraciones para los dos grandes tipos de métodos: jerárquicos y de partición. Los primeros proveen una estructura de grupos a diferentes niveles de granularidad según su nivel de similitud, mientras que los segundos dividen el conjunto muestral en grupos internamente homogéneos. En el caso de los métodos jerárquicos, se analiza en detalle las diferentes medidas de asociación y distancia utilizadas por el método, así como también el ligamiento usado para recalcular las distancias. La elección del índice de distancia es de suma importancia, dado que esta medida define el criterio por el cual dos elementos son considerados semejantes. Para los métodos de partición, se analizan las medidas de homogeneidad que definen la selección de los elementos dentro de cada grupo.

Nuestra propuesta tiene como objetivo, a mediano plazo, definir características comunes en los elementos, que nos permitan trabajar con modelos de predicción de propiedades fisicoquímicas, de manera que cada uno de los modelos difiera acorde al grupo sobre el cual fue clasificado. En particular para este trabajo nuestros experimentos se aplicaron sobre información multivariada de compuestos químicos para predicción de la propiedad logP (grado de hidrofobicidad de una sustancia). La técnica empleada en la predicción fueron redes neuronales y su validación fue realizada con otro conjunto de datos sin entrenar. Finalmente, se analiza la importancia de la justificación e interpretación de la clasificación seleccionada, así como también del grado de similaridad que cada grupo presenta.

1 INTRODUCCIÓN

La búsqueda de características similares dentro de un conjunto de datos para identificar grupos, también denominados *clusters*, constituye una técnica exploratoria de interés en variadas aplicaciones (Eisen *et al*, 1994; Slonim, 2002; Heeringa 2004; Guha et al. 2000). En particular, este tipo de técnicas exploratorias resultan de provecho en la comprensión de relaciones complejas de carácter multivariado. Básicamente, el agrupamiento nos provee de medios para identificar, a través de un espacio de múltiples dimensiones, elementos anómalos (*outliers*) e hipótesis concernientes a relaciones entre los datos.

Sin embargo, en la tarea de buscar grupos dentro de datos multivariados, existe una gran variedad de métodos (Johnson and Wichern, 1992; Anderberg, 1973) y dada las distintas restricciones y/o aplicaciones de cada uno de ellos, se requiere un conocimiento general del tema y un criterio juicioso antes de realizar pruebas “a ciegas”.

En el presente trabajo se realiza una revisión de los principales algoritmos de agrupamiento de datos multivariados junto con la aplicación de una selección de estas técnicas a una base de datos de compuestos químicos. Dada la complejidad inherente en la predicción de propiedades fisicoquímicas de compuestos, el objetivo de nuestro trabajo es aplicar uno de estos métodos para detectar grupos en función de su estructura molecular, y a partir de allí intentar aprender la relación del compuesto con propiedades fisicoquímicas, focalizando el análisis en este artículo sobre la propiedad logP (nivel de hidrofobia de un compuesto).

El trabajo se encuentra organizado de la siguiente manera. En la sección 2 se presenta una extensa revisión de distintas técnicas de análisis de agrupamiento junto con sus características. En la sección 3 se detalla la aplicación de algunas de las técnicas antes vistas a un conjunto de datos multivariados. Finalmente, en la sección 4 se encuentran las conclusiones del trabajo, junto con las consideraciones y planeamientos que se tendrán a futuro sobre esta línea de investigación.

2 TÉCNICAS PARA ANÁLISIS DE AGRUPAMIENTO

El análisis de Agrupamientos, de Conglomerados o de Clusters busca satisfacer dos objetivos básicos:

a) Similaridad intraclase, es decir formar grupos de elementos de manera tal que los pertenecientes a un mismo grupo sean parecidos entre sí.

b) Distanciamiento interclase, es decir formar grupos de elementos tales que los elementos pertenecientes a diferentes grupos sean lo más disímiles posible.

Es importante aclarar que el análisis de agrupamiento es distinto a los llamados métodos de clasificación. Estos últimos parten del conocimiento de un número fijo de grupos y en donde su objetivo es la asignación de nuevas observaciones a uno de estos grupos. El análisis de agrupamiento es una técnica que no requiere suposiciones sobre el número o la estructura de los grupos, sino que el agrupamiento es realizado en base a medidas de similaridad (o distancia) sobre los datos, y lo que se busca es descubrir el número y la composición de los grupos. Se lo conoce, también, como una técnica de clasificación *no supervisada*. Normalmente se usa para agrupar elementos en función de determinados atributos o variables (Modo Q), aunque también se puede usar para agrupar las variables en función de los valores que presenta en los individuos (Modo R). En este último caso el objetivo del análisis es detectar variables redundantes o correlacionadas.

Dentro del análisis de agrupamiento existen, básicamente, dos tipos de métodos: los jerárquicos y los de partición. En el caso de los primeros, se intenta ordenar los elementos a distintos niveles de similitud; mientras que los segundos, meramente, asignan cada elemento a un grupo de manera tal de obtener conjuntos homogéneos.

Para graficar la dificultad en la tarea de definición de grupos, considérese el ejemplo de agrupar, en diferentes sectores de un museo paleontológico, los fósiles de los animales, siguiendo un criterio lógico. Ahora, ¿qué significa un criterio lógico? ¿Por especie? ¿Por era geológica? ¿Una combinación de otros criterios (*e.g.* tamaño como combinación de: altura, ancho, largo)? ¿Aplicar un criterio y luego otro? Todo dependerá de la información que se desee diferenciar. Por otra parte, ¿cuál sería un índice apropiado para medir el grado de similitud o distancia entre dos individuos? ¿Y entre dos grupos de individuos? Éstos y otros interrogantes son las que se intentan abordar en la presente sección.

2.1 Medidas de asociación y distancia

Cualquier necesidad de producir una estructura basada en grupos partiendo de un conjunto complejo de datos requiere de una medida de "cercanía" o "similaridad". Como se dijo anteriormente, con frecuencia existe un gran problema de subjetividad en la elección de la medida de similaridad. También resulta relevante el tipo de las variables involucradas en la comparación (discreta, continua, binaria), su escala de medida (nominal, ordinal, intervalo, frecuencia) u otro conocimiento propio de la variable.

Las medidas de asociación asignan valores numéricos a cada par de elementos (o par de variables si es modo R). Estos valores crecen a medida que los elementos (variables) son más parecidos, de acuerdo con el criterio de asociación elegido. Este criterio suele estar definido por un *Índice de Asociación,* el cual se calcula en función de la matriz de entrada X_{nxp} (n cantidad de datos y p cantidad de variables). La mayoría de los índices son acotados, tomando valor "1" como máximo y "0" ó "-1" como mínimo. A su vez, los valores resultantes de aplicar un índice de asociación pueden ser volcados a una matriz de asociación A_{nxn}, en donde el elemento $A_{1,5}$ contiene la asociación entre los elementos 1 y 5. Generalmente, la matriz A es simétrica y con unos en la diagonal principal.

De igual modo que las medidas de asociación, se pueden definir medidas de distancia entre elementos (no utilizado para variables). Estos valores aumentan cuanto más disímiles sean los elementos comparados y sus datos pueden ser volcados a una matriz de distancias D_{nxn}, generalmente simétrica y con ceros en la diagonal principal. La utilización de matrices de distancia es más usual que las de asociación para el caso de la comparación entre elementos.

2.2 Propiedades de la distancia

Sea $d(i, j)$ la distancia entre los elementos i y j de la matriz de datos X_{nxp}:

1) $d(i, j) \geq 0$
2) $d(i,i) = 0$
3) $d(i, j) = d(j,i)$ (simetría)
4) $d(i, j) \leq d(i,m) + d(m, j)$ (desigualdad triangular)
5) $d(i, j) = 0$ sí y solo sí los elementos i y j son iguales
6) El índice es euclideano (i.e. las distancias entre n objetos puede ser representado en un espacio $\Re_k$, de modo tal que la distancia Euclídea entre cada par de puntos (i, j) sea igual a la correspondiente $d(i, j)$)

Es aconsejable que una distancia verifique las tres primeras propiedades; si también verifica las dos siguientes, es una métrica, lo que garantiza la representación de tres elementos como puntos en un plano y separados por las distancias correspondientes.

2.3 Medidas de distancias generales

Un gran número de distancias es determinado por la distancia de Minkowski, según sea el valor del parámetro $m \geq 1$ (ecuación 1). Esta familia de distancias posee una monotonía decreciente de los resultados con el aumento de m. La monotonía es importante dado que para ciertos procedimientos de clustering, la elección de diferentes medidas de distancia (en relación monotónica) no afecta el orden relativo de las agrupaciones.

$$d(i,j) = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{m}\right)^{1/m} \tag{1}$$

Para $m = 1$, $d(i,j)$ mide la distancia Manhattan o "city block". Para $m = 2$, $d(i,j)$ mide la distancia Euclídea. Sólo ésta última, de todas las basadas en Minkowski, cumple la propiedad 6, mientras que las otras restantes son sólo métricas. Para cualquier m, una condición necesaria es que sea aplicada a datos expresados en las mismas unidades en todas sus columnas (o, en su defecto, se deberá estandarizar previamente la matriz de datos).

Otra distancia que parte de la formulación de Minkowski, es la llamada de Bray-Curtis (ecuación 2), que es similar a la métrica de Manhattan en el sentido que la acota entre "0" y "1" al dividirla por los totales de ambas filas. Al igual que la anterior, requiere que sea aplicada a datos expresados en las mismas unidades en todas sus columnas, aunque en este caso no cumple la propiedad de ser métrica.

$$d(i,j) = \frac{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|}{\sum_{k=1}^{p} x_{ik} + \sum_{k=1}^{p} x_{jk}} \tag{2}$$

Similar a la anterior, resulta la distancia de Gower (ecuación 3), dado que también acota la de Manhattan entre "0" y "1", estandarizando la contribución de cada variable entre "0" y "1", y luego dividiendo por la cantidad de variables a comparar (p). En este caso se trata de una métrica y puede ser aplicada a datos no negativos con variables medidas en distintas unidades.

$$d(i,j) = \frac{1}{p} \sum_{k=1}^{p} \frac{\left| x_{ik} - x_{jk} \right|}{R_k}, \text{ donde } R_k = \max_h x_{hk} - \min_h x_{hk} \tag{3}$$

La última de las distancias acotadas basadas en Manhattan que presentaremos en este trabajo es la distancia de Canberra (ecuación 4), la cual divide cada diferencia entre las variables a comparar por su suma. El caso 0/0 se considera como 0 y está pensada para actuar como una medida de distancia binaria para la *doble ausencia* y la *presencia-ausencia*, y como medida cuantitativa para la *doble presencia*. Al igual que en el caso anterior es una métrica, puede usarse para datos con variables en diferentes unidades y requiere valores no negativos.

$$d(i,j) = \frac{1}{p} \sum_{k=1}^{p} \frac{\left| x_{ik} - x_{jk} \right|}{x_{ik} + x_{jk}} \tag{4}$$

Otros tipos de distancia, toman en cuenta la relación proporcional de los valores en las distintas variables, característica útil en la comparación de elementos con variables medidas en diferentes unidades. Por ejemplo, dos flores de tamaños bien distintos, pero con la

geometría de sus pétalos similares y con valores proporcionales de las variables entre sí en ambas muestras, son identificadas como similares en este tipo de distancias, mientras que se marcarían como disímiles para las basadas en Minkowski. Un caso particular de este tipo de distancia es la Coseno (ecuación 5), la que se calcula como uno menos el coseno del ángulo de los vectores de los puntos a comparar.

$$d(i,j) = 1 - \cos(\vec{i},\vec{j}) = 1 - \frac{\vec{i}}{\|\vec{i}\|} \bullet \frac{\vec{j}}{\|\vec{j}\|} \tag{5}$$

Asimismo, la distancia Cuerda de Orloci fue pensada para datos de abundancia, pero que se les quiere eliminar el efecto del tamaño de la muestra. Para esto, se normaliza cada vector a comparar y se calcula la distancia euclídea entre ellos. La normalización se realiza con la longitud propia de cada vector (ecuación 6). Otros métodos realizan estandarizaciones de cada variable por su desvío como se hace en la distancia S-euclidean (de Euclídea estandarizada) o, más general aún, por su matriz de covarianza tal como se realiza en la de Mahalanobis (ecuación 7).

$$d(i,j) = \sqrt{\sum_{k=1}^{p} \left(\frac{x_{ik}}{\|\vec{x}\|} - \frac{x_{jk}}{\|\vec{x}\|} \right)^2} \tag{6}$$

$$d(i,j) = \sqrt{(\vec{i} - \vec{j})\,\mathrm{cov}(\vec{i},\vec{j})^{-1}(\vec{i} - \vec{j})} \tag{7}$$

Existen muchos otros tipos de distancias que se aplican en diferentes contextos. Estos contextos difieren en si los datos son medidos mediante valores binarios o porcentajes, o de si se aplican para comparar distancias entre elementos o entre variables.

Por otra parte, a partir de las distancias acotadas entre “0” y “1” se pueden definir índices de asociación $s(i,j)$ como $1 - d(i,j)$. Si bien algunos índices, como el de Gower, fueron construidos de esta manera, otros surgieron antes que las distancias asociadas, como en el caso del índice de Czekanowski (complemento del índice de Bray-Curtis).

2.4 Agrupamiento jerárquico

El objetivo básico de los algoritmos de agrupamiento, consiste en obtener grupos de elementos a partir de la matriz de distancia (o asociación) y sin tener que examinar todas las posibles combinaciones de agrupamiento.

El agrupamiento por jerarquías (*hierarchical clustering*) puede realizarse mediante una serie sucesiva, ya sea de uniones (aglomerativo) o de divisiones (divisivo). Los métodos jerárquicos aglomerativos empiezan con los elementos individuales, considerándose cada uno de ellos como un grupo. Los objetos más similares son los primeros en unirse, y los grupos así formados continúan uniéndose en forma progresiva de acuerdo a sus similaridades. Finalmente, todos los grupos quedan fusionados en uno sólo. El Algoritmo 1 esquematiza la lógica de este método.

```
1- Inicializar cada punto como un cluster.
2- Repetir Mientras haya más de un cluster.
2.1-   Calcular la similaridad entre todos los pares de clusters.
2.2-   Encontrar los pares de clusters más similares.
2.3-   Unir esos dos clusters formando un nuevo cluster
   FIN Repetir Mientras
```

Algoritmo 1: Método Jerárquico Aglomerativo

En contrapartida, los métodos jerárquicos divisivos trabajan en el sentido opuesto. Un único grupo inicial de elementos se divide en dos subgrupos, de manera que los elementos de un subgrupo sean "bien disímiles" con los elementos del otro. Estos subgrupos son divididos sucesivamente en subgrupos disímiles, hasta que queden tantos subgrupos como cantidad de elementos a agrupar.

Los resultados de aplicar el método, ya sea aglomerativo o divisivo, pueden ser ilustrados mediante un diagrama de dos dimensiones conocido como dendrograma, el cual se muestra en la Figura 1 para el caso de un aglomerativo. Como puede verse, el dendrograma muestra el orden de las uniones (divisiones) que se producen y además la distancia a la que se unen dos grupos.

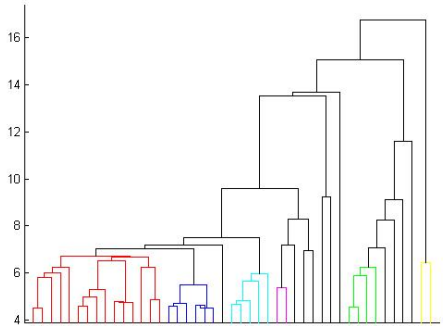


Figura 1: Dendrograma aglomerativo (distancia máxima intergrupos = 8). Cada color indica un grupo distinto.

Cabe aclarar que el método carecería de sentido, si no se estableciera un método que "corte" el dendrograma a un cierto nivel. Para solucionar este problema se pueden tomar distintas alternativas modificando la condición de corte. En vez de continuar hasta que quede un solo cluster, se puede fijar una cantidad de clusters mayor que uno y menor que la cantidad de elementos del conjunto inicial. Otro criterio puede ser fijar una altura, *i.e.* cortar el algoritmo cuando los clusters que se fusionan superan una cierta distancia (Figura 1). Quizás un criterio más natural, que fijar la cantidad de grupos o distancia máxima, sea establecer un criterio cuantitativo de inconsistencia del ligamiento. El mismo intenta identificar divisiones naturales en los datos, comparando la distancia de un ligamiento con la distancia a la que se unen los ligamientos por debajo de esa unión. Esto se realiza así, dado que un ligamiento que se produce a casi la misma distancia a la que se produce en sus hijos, identifica que no existen diferencias grandes entre los dos grupos unidos. En esta sección nos concentraremos en los métodos aglomerativos y, en particular, en los basados en métodos de ligamiento. Revisiones sobre métodos divisivos y otras técnicas aglomerativas, puede encontrarse en (Anderberg, 1973; Everitt, 1974).

Los métodos por ligamientos son igualmente útiles para agrupar tanto elementos como variables, hecho que no es posible para todos los métodos aglomerativos. En el paso 2.2 del Algoritmo 1, se calcula la distancia entre clusters, etapa conocida con el nombre de ligamiento. El mismo puede hacerse de distintos modos, y esto da lugar a distintas variantes de agrupamiento. En particular analizaremos las siguientes formas de ligamiento: simple, completa, promedio (ponderado y no ponderado), centroide (ponderado y no ponderado), mediana y el ligamiento de Ward (1963).

En el caso del ligamiento simple (mínima distancia o vecino más próximo), la distancia entre dos grupos está dada por la distancia de los elementos más cercanos de cada uno de los grupos. En el ligamiento completo (máxima distancia o vecino más lejano) los grupos se

fusionan de acuerdo a la distancia de los elementos más lejanos de cada grupo. Para el ligamiento promedio (*UPGMA, unweighted pair groups method with arithmetic averages)*, la distancia computada entre dos grupos es la correspondiente al promedio de las distancias que conforman ambos grupos. El ligamiento por centroide (UPGMC) utiliza la distancia Euclídea entre los centroides de los dos clusters a comparar. En la Figura 2 se esquematizan estos ligamientos.

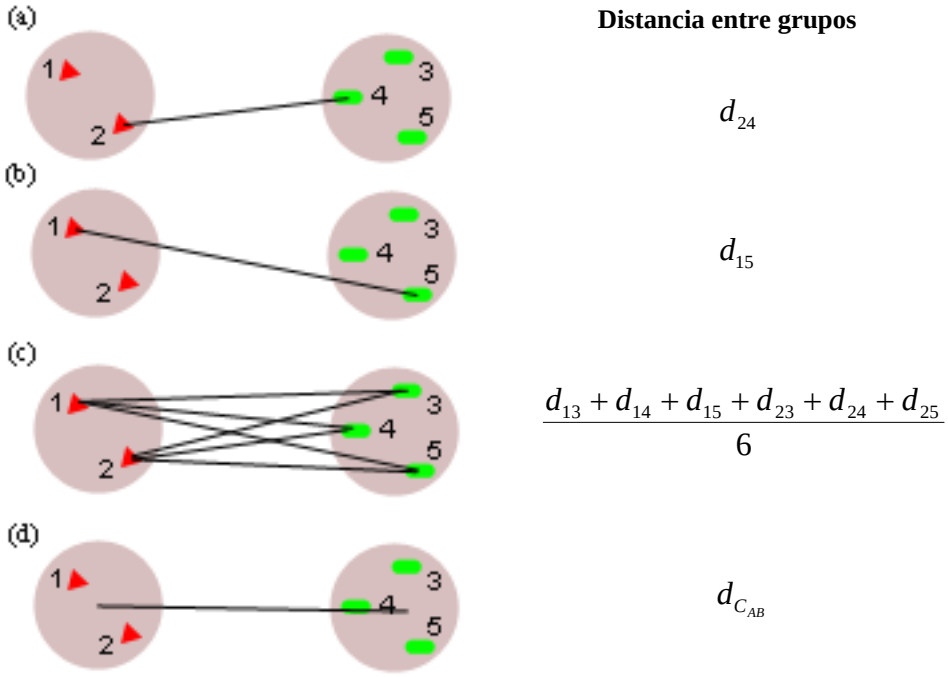


Figura 2: Distancia entre clusters para (a) ligamiento simple (b) ligamiento completo (c) ligamiento promedio (d) ligamiento por centroide

Existen otros ligamientos más complejos de graficar, tal como el ligamiento promedio ponderado (*WPGMA, weighted pair groups method with arithmetic averages*) y el ligamiento por centroide ponderado (*WPGMC, weighted pair groups method with arithmetic centroids*), en los cuales se pondera el cálculo de las distancias según sea el orden en que fueron fusionados los grupos. En la ecuación 8 se muestra el criterio seguido en estos dos últimos casos, al calcular las distancias de un nuevo cluster, formado por R y S, al unirse con T. Finalmente, el ligamiento de Ward, utiliza como criterio el aumento del valor de la suma de las distancias al cuadrado de cada elemento al centroide, como resultado de la unión de dos grupos. El algoritmo busca juntar los dos elementos que produce el menor incremento de W (ecuación 9)

$$d(\{R,S\},T) = \frac{d(R,T)+d(S,T)}{2} \tag{8}$$

$$W = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{ig}-\bar{x}_g)^T(x_{ig}-\bar{x}_g), \text{ donde G = cantidad de grupos y } n_g = \text{cardinalidad del grupo} \tag{9}$$

Como conclusión de los ligamientos mostrados, se puede decir que el ligamiento simple posee dificultades para detectar clusters que no están claramente separados. Una tendencia

común es que los clusters se vayan agrupando en forma “longitudinal”, dando lugar al problema del “encadenamiento”, en donde los elementos de los extremos de la cadena pueden quedar a gran distancia entre sí (Figura 3). Por otra parte, este ligamiento resulta útil para descubrir grupos que no tienen una forma elíptica.

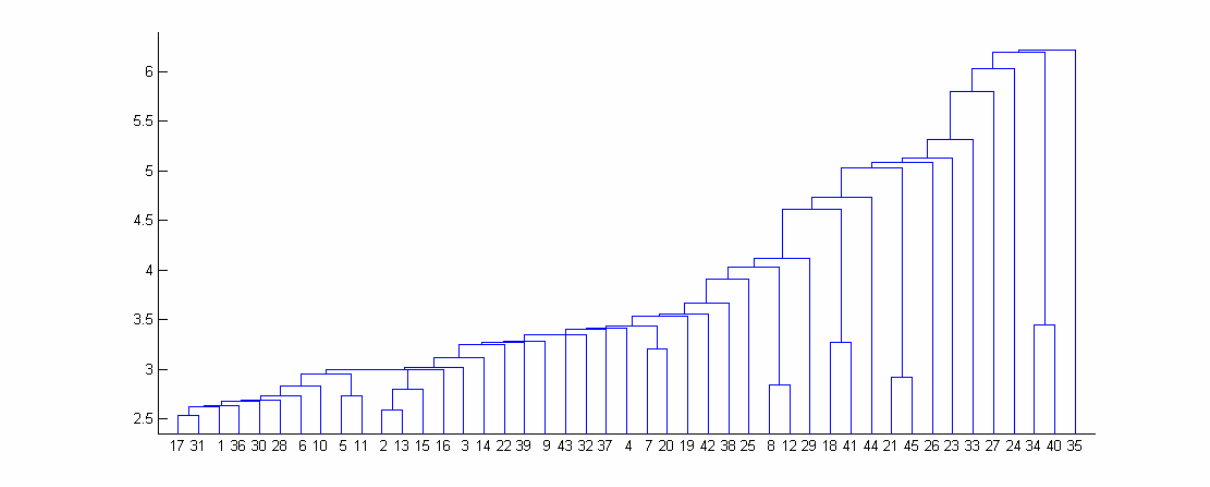


Figura 3: Ligamiento simple – problema del encadenamiento

Por otra parte, el ligamiento completo tiende a generar grupos más compactos y estables, ya que justamente en cada paso busca unirse con el grupo menos distinto. El ligamiento por centroide, representa una simplificación del por promedio no ponderado, sin embargo puede dar lugar a dendrogramas no monótonos (o invertidos), tal como el mostrado en la Figura 4.

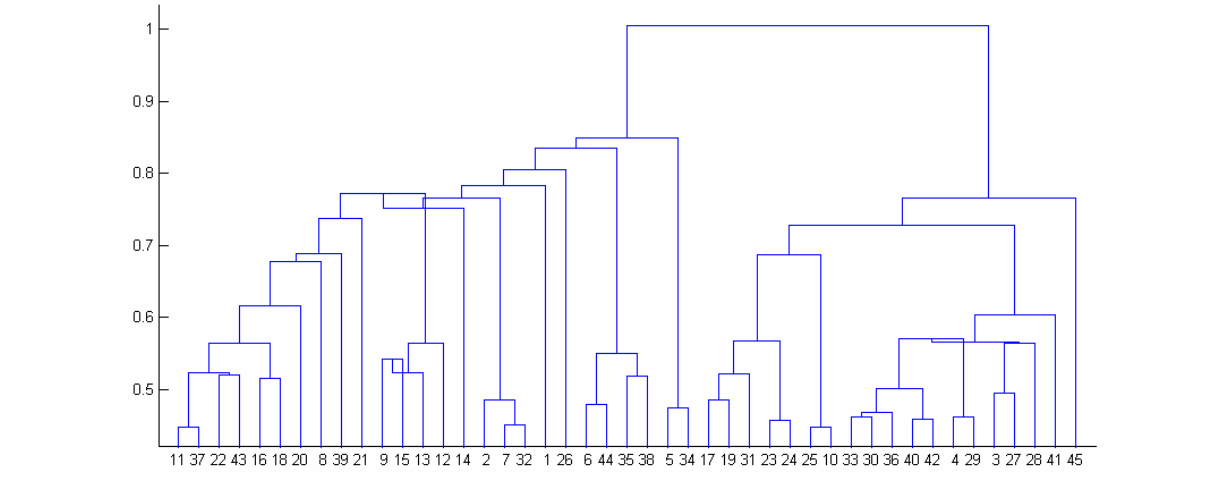


Figura 4: Ligamiento por centroide – problema del ligamiento no-monótono

Como conclusión final de los métodos jerárquicos, podemos agregar que estos métodos son sensibles a los elementos outliers (anómalos) o puntos con ruido. Consecuentemente, la configuración final de los grupos debe ser cuidadosamente analizada, para detectar tales situaciones. Una idea sensata es probar distintas medidas de distancia y distintos criterios de ligamiento de manera de obtener resultados consistentes y agrupamientos naturales.

2.5 Métodos de agrupamiento no jerárquicos

Las técnicas de agrupamiento no jerárquicas, fueron pensadas para agrupar elementos, más que variables, en un conjunto de K grupos. El número de grupos, K, puede ser especificado a priori o puede determinarse como parte del procedimiento. Dado que no es necesario calcular ni almacenar la matriz de distancia (asociación) antes de comenzar con el procedimiento, este tipo de métodos puede ser aplicado a mayor cantidad de datos.

En estos métodos se puede partir tanto de una división inicial en grupos de los elementos, o de un conjunto inicial de puntos, los cuales actúan como centros de los futuros grupos. Estas configuraciones iniciales deben ser dispuestas de manera de que se evite cualquier tipo de tendencias (biases). Un modo posible es comenzar con una división azarosa de los elementos o con un conjunto aleatorio de puntos-centro.

El método de K-medias, es uno de los más populares de este tipo, el cual va asignando cada elemento al centroide más cercano. El método puede describirse mediante el Algoritmo 2.

```
1- Particionar el conjunto de elementos en K grupos.
2- Calcular las distancias euclídeas (estandarizada o no) de cada elemento
a cada uno de los K centros y asignarlo al grupo cuyo centro esté más
próximo. Recalcular los nuevos centroides después de cada asignación de un
nuevo elemento para el grupo del cual se va y para el grupo al cual llega.
3- Definir un criterio de optimalidad y comprobar si una nueva reasignación
lo mejora. En ese caso se vuelve al paso 2-
```

Algoritmo 2: Método de las K-medias

El paso 1 puede reemplazarse por la variante de comenzar con los puntos-centro y a partir de ahí, agregar e intercambiar elementos en función de los centroides formados. El criterio de optimalidad del método de K-medias es minimizar la suma de cuadrados dentro de los grupos (ecuación 10).

$$SCDG = \sum_{g=1}^{K} \sum_{j=1}^{p} \sum_{i=1}^{n_g} (x_{ijg} - \bar{x}_{ig})^2 \tag{10}$$

Una desventaja importante de este método es su alta dependencia de la partición inicial o la selección inicial de los puntos-centro. Si dos o más puntos-centro caen dentro de lo que conformaría un mismo cluster, ambos grupos estarían pobremente diferenciados. Asimismo la existencia de outliers, produce al menos un grupo con elementos demasiado dispersos. Además, el no conocer el valor de K puede dar lugar a agrupamientos no naturales. Por estas razones, para mejorar la estabilidad del método, es deseable volver a correr el algoritmo con otras configuraciones iniciales. Más información sobre el uso de procedimientos no jerárquicos de agrupamiento puede encontrarse en (Anderberg, 1973; Everitt, 1974).

2.6 Visualización

En adición a un dendrograma, existen otras formas de mostrar la similitud o diferencias entre los elementos. En primer lugar, podemos considerar PCA (*Principal Component Analysis*) como un método que nos permite visualizar los datos utilizando un número menor de dimensiones, asegurando que en esas componentes se rescata el mayor porcentaje de la varianza, y por ende de la información. En el caso de variables en diferentes unidades de medidas, se requiere como preprocesamiento la estandarización por desvío de la matriz de covarianzas transformándola a una de correlación, de manera que no influyan los rangos de

valores de las distintas variables.

Similar al anterior es el método de escalado multidimensional o análisis por coordenadas principales. El mismo intenta reducir el número de dimensiones de manera que, en las proyecciones a las coordenadas elegidas, la distancia representada sea la más parecida a la real para un número fijo de coordenadas.

Existen otros métodos de visualización especialmente diseñados para el análisis de información multivariada, en el cual outliers, relaciones y grupos puedan ser identificados mediante una simple inspección visual. Entre ellos se encuentra los gráficos de cubrimiento bidimensionales, que consiste en poner en forma de matriz las distintas combinaciones de gráficos de dos dimensiones, i.e. el elemento $A_{2,3}$ de la matriz contiene la representación gráfica de la proyección sobre el plano formado por las variables 2 y 3. Otras herramientas de este estilo aunque muy distintas en la forma de comunicar su información, corresponden al modelo de Estrellas, matrices sombreadas (shadow matrices), Gráficos de Andrew y Caras de Chernoff. Información detallada de estás técnicas pueden encontrarse en (Johnson and Wichern, 1992; Grinstein *et al*, 2001).

3 EVALUACIÓN EXPERIMENTAL DE LAS TÉCNICAS DE CLUSTERING

En la presente sección se muestra la aplicación del análisis de agrupamiento a un conjunto de compuestos químicos de carácter multivariado. El desarrollo del experimento aquí mostrado es producto de muchas pruebas y de repetidos intentos con diferentes combinaciones de parámetros y métodos, en búsqueda de una mayor comprensión de la relación estructura-actividad de una molécula, disciplina conocida como QSAR (Quantitative Structure-Activity Relationship). En particular, lo que se busca descubrir es la relación existente entre la estructura molecular de un compuesto y su hidrofobia, propiedad importante para el desarrollo de nuevos fármacos.

3.1 Predicción de logP

Durante la última década ha quedado en evidencia la complejidad del mecanismo de acción de las drogas y se ha investigado especialmente paradigmas que toman en cuenta las llamadas propiedades ADME-Tox (Absorción, Distribución, Metabolismo, Excreción y Toxicidad) de compuestos candidatos a drogas. En épocas anteriores, el interés de la comunidad científica estaba fijado sólo en la potencia de la droga, y era poca la atención dada a su comportamiento en un sistema biológico in-vivo. Consecuentemente, en los últimos 10-15 años se han desarrollado numerosas investigaciones tendientes a evaluar, mediante técnicas in-vitro (*e.g*. high-throughput assays), distintas propiedades de los compuestos candidatos (Selick et al. 2002).

Nuevas investigaciones se han orientado a la predicción de propiedades ADME por computadora (Jónsdottir et al. 2005). Estos modelos in-silico, desarrollados en forma confiable, permiten la clasificación y selección de compuestos candidatos a drogas utilizando muchos menos recursos económicos que los requeridos por los estudios convencionales. Además, y a diferencia de los ensayos in-vitro, estos modelos permiten ser aplicados a compuestos antes de ser sintetizados. Otra ventaja de estas nuevas herramientas es que permiten un mejor conocimiento de las relaciones entre las propiedades ADME-Tox, la estructura y las características físico-químicas de los compuestos.

En particular, nuestras investigaciones apuntan al desarrollo de herramientas que asistan en la predicción de la propiedad fisicoquímica logP (coeficiente de partición octanol/agua), la cual brinda una medida de la hidrofobia (hidrofilia) de una sustancia. Dentro del contexto ADME, esta propiedad, se relaciona con la absorción, metabolismo y toxicidad. Según la reconocida “regla de los 5” de Lipinski (Lipinski *et al.*, 1997), un compuesto para ser considerado como

candidato a fármaco de absorción oral, debe cumplir que el valor de logP esté entre 0 y 5.

En la tarea de predicción de propiedades, resulta un factor clave la elección de los descriptores moleculares. Existen dos tipos de descriptores: teóricos y experimentales (Todeschini y Consonni, 2000). Un descriptor teórico se calcula a través de un procedimiento lógico y matemático que transforma información química de una molécula, codificada en una representación simbólica, en un número útil. En cambio, el valor de un descriptor experimental se obtiene mediante algún experimento estandarizado. Asimismo, los descriptores teóricos se dividen en distintas familias en función de su propia naturaleza.

De los muchos descriptores existentes, no todos son importantes para modelar la propiedad a predecir. También se puede dar que los descriptores elegidos brinden información redundante o altamente correlacionada. Una elección de descriptores inapropiada puede ocasionar correlaciones *by chance* (Topliss and Edwards, 1979) o un escaso poder predictivo.

En base a la revisión realizada en la bibliografía de modelos de cálculo de LogP, existen fuertes evidencias que muestran la potencialidad de las redes neuronales como método apropiado para la predicción de propiedades (Winkler, 2004; Duprat et al., 1998; Agatonovic-Kustrin and Beresford 2000; Taskinen and Yliruusi, 2003). Esto se debe básicamente a la capacidad de las redes de actuar como "aproximadores universales", capaces de modelar intrínsecamente cualquier función continua al nivel de precisión deseado, si es provista con suficiente cantidad de datos de entrenamiento. Son no-lineales e ideales para modelar sistemas complejos, dado que no requieren de conocimiento previo para la construcción de modelos.

Las redes neuronales corresponden a un método de regresión supervisada que también puede ser usado para predicción. Primero, se procede a entrenar un conjunto de datos de entrenamiento, y luego se aplica este mismo modelo a un conjunto de datos de validación, en el cual se intenta generalizar o predecir la propiedad antes entrenada.

Sin embargo, y al igual que otras técnicas de regresión, las redes neuronales pueden sufrir del sobreajuste (*overfitting*), proveniente del exceso de pesos ajustables en comparación al número de elementos a entrenar, y generar correlaciones *by chance*. Otro inconveniente es el sobreentrenamiento, el cual resulta de entrenar por más iteraciones que las suficientes los datos de entrenamiento, consiguiendo en consecuencia una memorización de los datos, en lugar de un aprendizaje de sus características. En definitiva, ambos problemas generan un detrimento en la capacidad de generalizar o inferir a partir de lo entrenado (Tetko et al., 1995). Esto hace necesario recurrir a tests de validación rigurosos que permitan cuantificar la calidad y precisión de las predicciones inferidas.

El objetivo de aplicar análisis de agrupamiento previo a la tarea de predicción, reside en confirmar la hipótesis de mejorar la validación en las redes neuronales, aplicando selectividad mediante una división de los compuestos con algún criterio de los vistos en la sección 2 (en modo Q). De esta manera, se pretende lograr que en cada grupo se concentren compuestos similares que compartan características propias. Luego, nuestra hipótesis es que entrenando las características de cada cluster con redes neuronales específicamente definidas para cada grupo, se puede mejorar la capacidad predictiva sobre las muestras usadas en la validación. La idea es aplicar para cada compuesto del conjunto de validación el modelo de red que pertenezca al grupo entrenado más cercano a dicho compuesto. De hecho, ya existen otros enfoques que si bien no utilizan análisis de agrupamiento, sí emplean medidas de similaridad para mejorar la predicción (Tetko 2002a, b). A continuación se detallan los experimentos realizados, los cuales, confirman, al menos para nuestro caso de estudio, la hipótesis recientemente planteada.

3.2 Diseño de experimentos

El trabajo experimental aplicado consiste en la predicción de la propiedad físico-química logP mediante redes neuronales. La hipótesis a comprobar es si se obtiene un mejor resultado aplicando análisis de agrupamiento al conjunto de datos a entrenar y generando, luego, un modelo de red neuronal para cada grupo resultante. A los datos separados para validación se les aplica la red neuronal correspondiente, según sea la distancia de ese compuesto al grupo más cercano. Esta medida de distancia tomada, así como el criterio de cercanía a un grupo, es la misma que la utilizada al momento de hacer la partición sobre el entrenamiento. En la Figura 5 se esquematiza este procedimiento de inferencia.

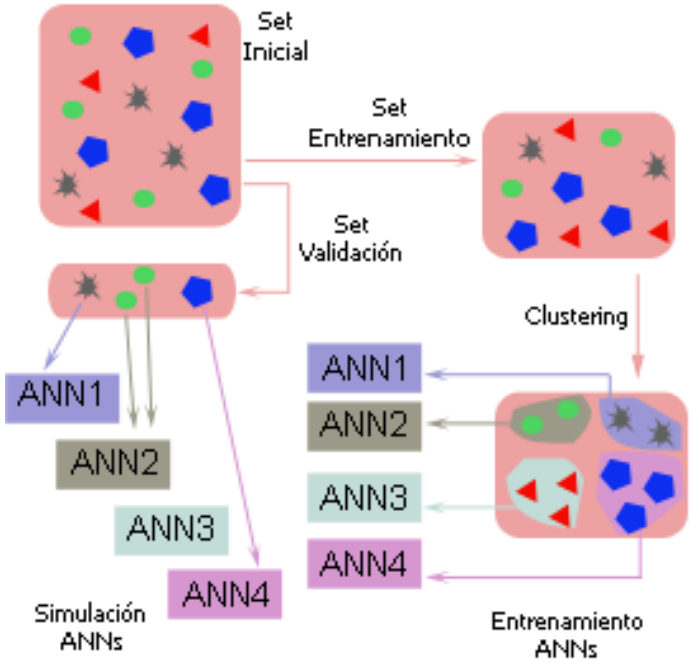


Figura 5: Esquema conceptual del procedimiento de inferencia.

Los datos utilizados para este trabajo corresponden a un subconjunto de los disponibles en la base de datos PHYSPROP (url:http://www.syrres.com/esc/), la cual contiene el código SMILES y valor de logP de alrededor de 13000 compuestos químicos. Para cada uno de los datos se calcularon cerca de 1500 descriptores moleculares. En base a la literatura consultada y a experimentaciones realizadas, se eligieron dos familias de descriptores para los cálculos en la predicción de logP: los constitucionales y los de conteo de grupos funcionales (Todeschini and Consonni, 2000). Sin embargo, para la tarea previa de análisis de agrupamiento, sólo los descriptores de la segunda familia fueron utilizados. Este grupo consiste de 121 variables de conteo relativos a la presencia de ciertos grupos funcionales químicos. Para la tarea de predicción, se utilizó la familia anterior de grupos funcionales en conjunción con los siguientes 16 descriptores constitucionales:

- peso molecular
- suma de los volúmenes atómicos de van der Waals
- suma de las electronegatividades atómicas de Sanderson
- suma de las polarizabilidades atómicas
- suma de los estados electrotopológicos de Kier-Hall
- suma de los órdenes de enlaces convencionales
- cantidad total de átomos

- cantidad de átomos de hidrógeno
- cantidad de átomos de fluor
- cantidad de átomos de carbono
- cantidad de átomos de nitrógeno
- cantidad de átomos de oxígeno
- cantidad de átomos de cloro
- cantidad de átomos de bromo
- cantidad de átomos de yodo
- cantidad de átomos de sulfuro

La cantidad total de compuestos químicos utilizada fue de 4575, de los cuales 4042 se utilizó para el entrenamiento y 458 para la validación (90%-10%). Tanto para el agrupamiento como para la predicción con redes neuronales, se redujo el número de descriptores, proyectándolos sobre sus componentes principales.

Dentro de cada cluster se utilizó la red neuronal que mejor comportamiento obtuviera, en base a las distintas corridas realizadas, y a los cambios sobre las numerosas variantes de arquitecturas, funciones de activación, reglas y modos de aprendizaje. La medida de distancia utilizada fue la distancia Coseno, mientras que el ligamiento completo fue elegido como criterio de comparación de distancia intergrupo. El algoritmo de aprendizaje de la red fue el de descenso por gradiente con optimización por momento y variación de paso adaptivo (*gradient with momentum and adaptive learning rate*).

3.3 Resultados de los experimentos

En la Figura 6 se observa que se tomó como umbral de corte del agrupamiento una distancia intergrupo de 1.85. De los 8 grupos así formados, uno de ellos fue descartado por su baja cantidad de elementos (119). Nótese que las hojas del árbol corresponden no a elementos sino a conjuntos de elementos.

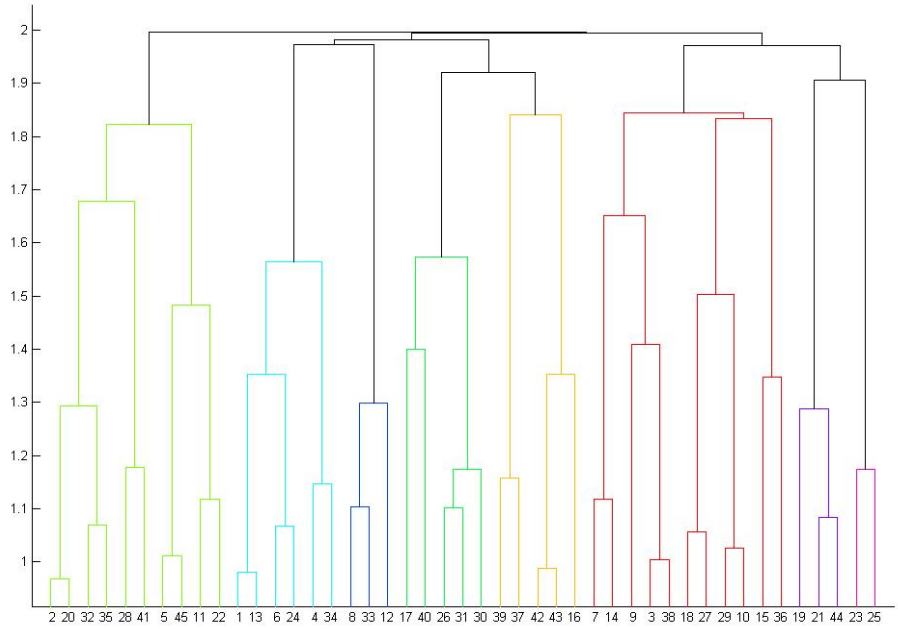
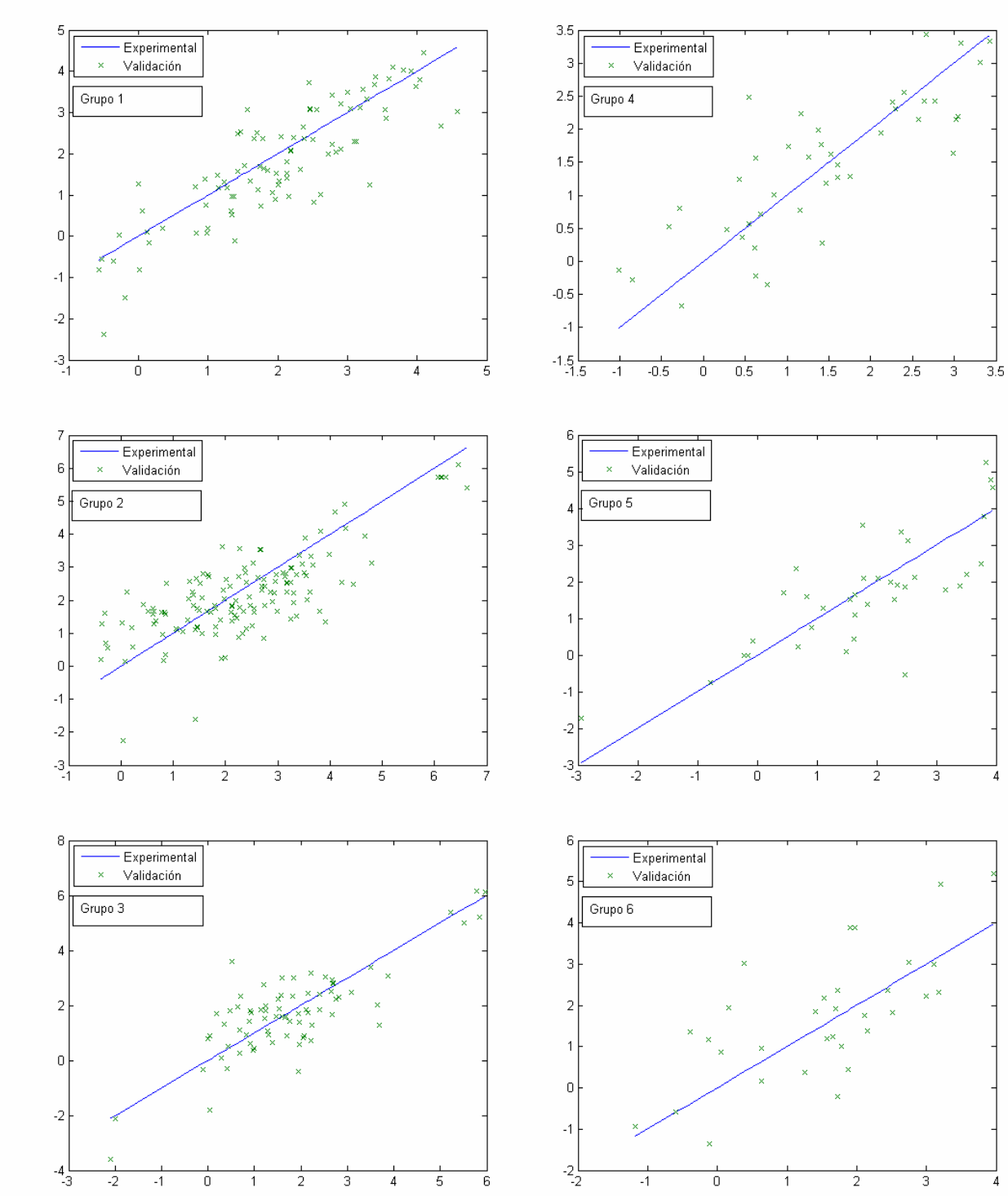


Figura 6: Dendrograma sobre el conjunto de entrenamiento

En las Figura 7 se muestran gráficos de los resultados de la validación que se obtuvo por cada grupo. Cada uno de ellos muestra el error entre el valor experimental y el valor calculado por la red, y se tomó el RMSE (root mean square error) como medida general de error. Los RMSE obtenidos en unidades de logP por cada grupo son: $RMSE_{G1}$ = 0.57, $RMSE_{G2}$ = 1.00, $RMSE_{G3}$ = 0.87, $RMSE_{G4}$ = 1.06, $RMSE_{G5}$ = 1.02, $RMSE_{G6}$ = 1.26 y $RMSE_{G7}$ = 0.59



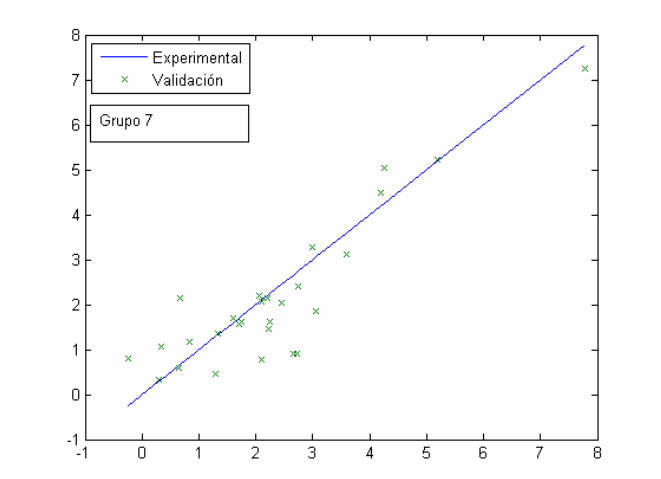
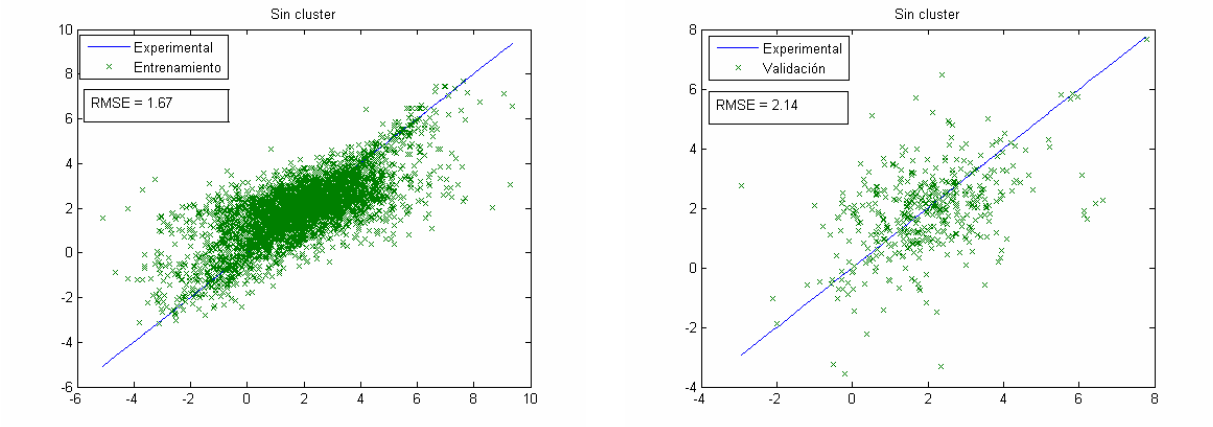


Figura 7: Resultados de las validaciones de las siete redes neuronales

3.4 Análisis de los resultados

A partir de los resultados experimentales pueden obtenerse conclusiones interesantes. La primera y más importante es que aplicar un preprocesamiento sobre los datos utilizando análisis de agrupamiento ayudó en la mejora de la capacidad predictiva. En la Figura 8, se visualiza la diferencia entre utilizar un solo modelo de red neuronal sobre todo el conjunto de datos (arriba), y aplicar distintas redes según el grupo al que pertenece (abajo). En particular, el gráfico del extremo inferior derecho de la Figura 8 corresponde a la "superposición" de los gráficos de la Figura 7. Como se puede apreciar, el empleo de análisis de agrupamiento mejora considerablemente el ajuste sobre los datos experimentales del entrenamiento (RMSE de 1.26 y 1.67 con y sin agrupamiento respectivamente). El mismo resultado se da en cuanto a la mejora en la predicción sobre el conjunto de validación (RMSE de 0.53 y 2.14 con y sin agrupamiento respectivamente).

La hipótesis de mejorar la predicción de las redes a partir de la subdivisión del problema en partes menores, fue corroborada por los experimentos realizados. El agrupamiento brinda la posibilidad de separar características dentro de cada grupo y, de esta manera, la red sólo se encarga de aprender características más específicas, lo que en definitiva mejora la capacidad de predicción global.



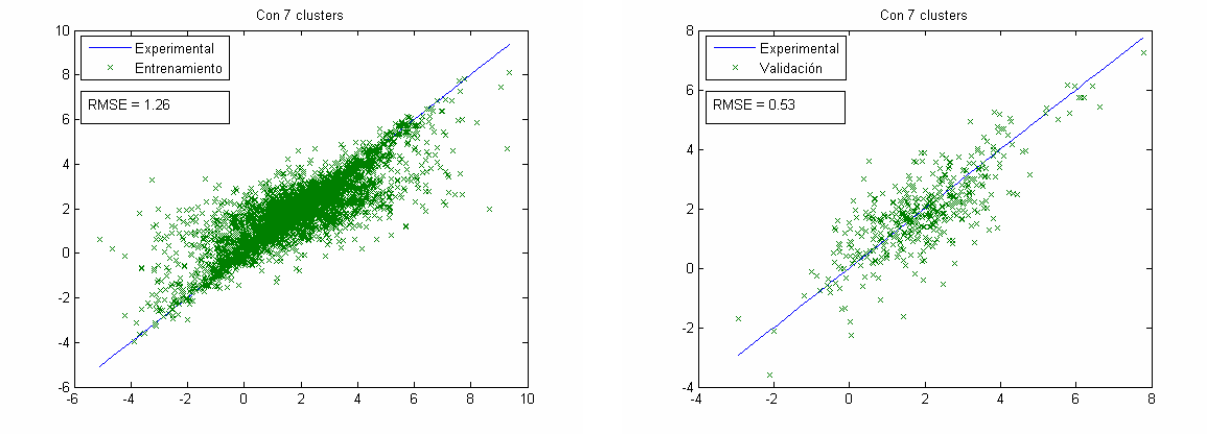
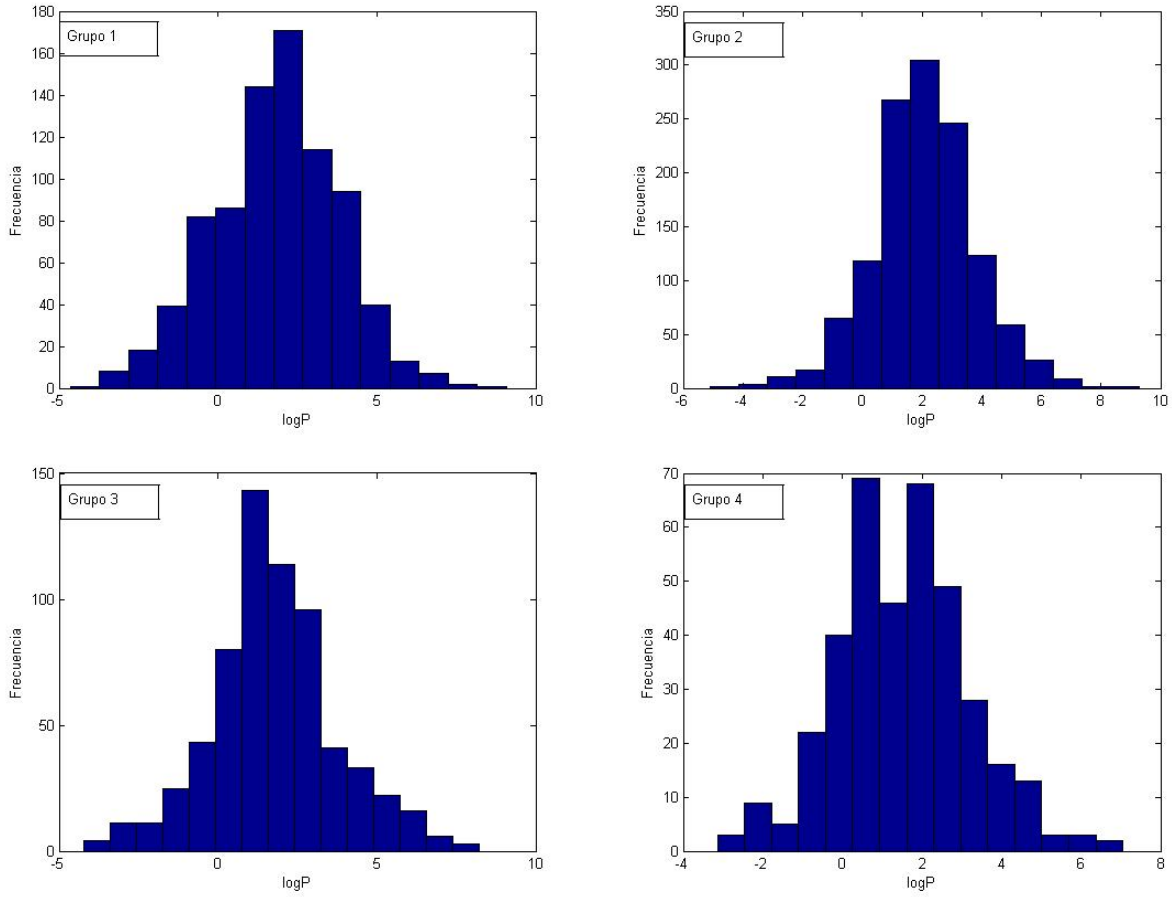


Figura 8: Entrenamiento (izquierda) y validación (derecha), sin aplicar análisis de agrupamiento (arriba) y con la aplicación de análisis de agrupamiento (abajo)

En la Figura 9, se visualizan los histogramas de cada uno de los grupos y el correspondiente a la unión de todos ellos (último histograma). En cada uno se dividió el rango de valores de logP en 15 subrangos. Analíticamente se obtuvo que el 14% de todos los los compuestos tienen valor de logP por debajo de 0 y el 11% es mayor a 4. En nuestro análisis, el grupo 1 contiene un mayor número de compuestos con valores extremos de logP, (19% menores a 0 y el 13% mayores a 4); los grupos 4 y 6 prácticamente no contiene valores de logP altos (6% y 7%, respectivamente); mientras que los grupos 5 y 6 tienen gran cantidad de compuestos muy bajos (18% y 24%, respectivamente); finalmente los grupos 3 y 7 presentan un comportamiento similar al promedio



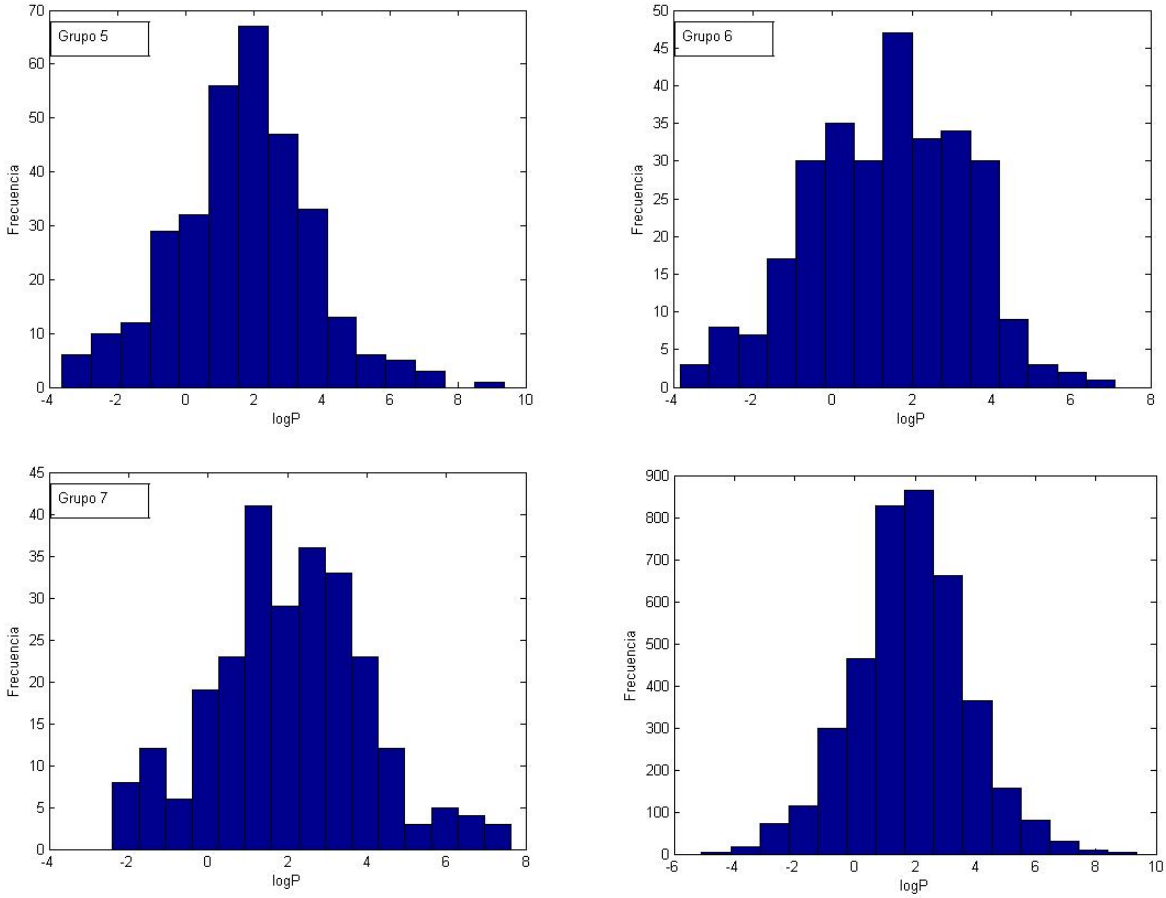


Figura 9: Histogramas de cada uno de los grupos y del total de compuestos.

4 CONCLUSIONES Y TRABAJO A FUTURO

Los resultados aquí presentados forman parte de una línea de investigación en avance sobre la compleja tarea de mejorar la predicción de propiedades ADME-Tox de compuestos químicos. A partir de la necesidad de realizar un análisis exploratorio sobre el conjunto de datos, se planteó utilizar análisis de agrupamiento como paso previo a la predicción de una fórmula de correlación para la propiedad logP. La propuesta consiste en entrenar una red neuronal especifica para cada grupo de datos obtenido a partir del análisis de agrupamiento, de forma tal de obtener una fórmula de correlación del logP adecuada a las características de cada grupo. Siguiendo esta estrategia se obtuvo resultados satisfactorios, dado que se mejoró la calidad y precisión de la predicción del logP respecto al uso de redes neuronales sin análisis de agrupamiento. Dado estos logros preliminares, consideramos que nuestra propuesta es muy promisoria, razón por la cual se proyecta seguir trabajando en este enfoque.

En primer término, se planifica bajar el umbral de distancia máxima intracluster. Si bien esto nos originará más grupos, y por ende más entrenamiento de redes, creemos que esta mayor segmentación incrementará el poder predictivo del método, dada la mayor semejanza entre los compuestos. Esto requerirá aumentar la cantidad de compuestos a utilizar tanto en el entrenamiento como en la validación de las predicciones. El análisis de otros descriptores en la predicción de propiedades también constituye una tarea de continuo estudio. Para esto se planea utilizar enfoques similares a los usados en (So and Karplus, 1996; Yasri and Hartsough, 2001) en donde la elección de los descriptores es asistida por algoritmos genéticos, así como también otros

enfoques estadísticos (Withley *et al*, 2000). Asimismo se planea, haciendo uso de la partición realizada sobre los datos, variar los descriptores según el grupo de pertenencia.

Finalmente, se pretende utilizar algún método de clasificación en relación a rangos de valores de logP. Para esto se piensa utilizar análisis de discriminantes y/o support vector machines. En cuanto a la tarea de predicción también se piensa evaluar otros métodos de inteligencia computacional para predicción, como árboles de decisión y support vector machines.

AGRADECIMIENTOS

Los autores desean expresar su agradecimiento a la Agencia Nacional de Promoción Científica y Tecnológica de la Argentina por la subvención otorgada en el marco del Programa de Modernización Tecnológica, Contrato de Préstamo BID 1728/OC-AR, al PICT Nº11-12778, denominado "Procesamiento paralelo distribuido aplicado a ingeniería de procesos", aprobado por Resolución ANPCYT Nº117/2003 y al PICTO-UNS Nº917, denominado "Re-Ingeniería de un sistema de soporte de decisión para localización estratégica de sensores en plantas industriales".

También queremos agradecer al CONICET por la subvención otorgada al Proyecto de Investigación Plurianual (PIP 5930): “Métodos Computacionales para Predicción de Propiedades, Simulación e Instrumentación de Procesos Industriales”, y a la Secretaría de Ciencia y Tecnología de la Universidad Nacional del Sur por la subvención otorgada al Proyecto de Grupos de Investigación (PGI 24/N019): "Aplicaciones de computación científica".

REFERENCIAS

M.R. Anderberg. *Cluster Analysis for Applications.* New York: Academic Press, 1973.

S. Agatonovic-Kustrin, R. Beresford. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, 22, 5:717-727, 2000.

A. Duprat, T. Huynh, and G. Dreyfus. Towards a principled methodology for neural network design and performance evaluation in qsar; application to the prediction of logp. *Journal of Chemical Information and Computer Science*, 38:586–594, 1998.

M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95:14863–14868, 1994.

B. Everitt. *Cluster Analysis*. London, Heinemann Educational Books, 1974.

G. Grinstein, M. Trutschl, U. Cvek. High-dimensional visualizations. *Proceedings of the Visual Data Mining Workshop, KDD*, 2001.

S. Guha, R. Rastogi and K. Shim. ROCK: A Robust Clustering Algorithm for Categorical Attributes. *Information Systems*, 25, 5:345-366, 2000.

W. J. Heeringa. Measuring Dialect Pronunciation Differences using Levenshtein Distance. PhD Thesis. Rijksuniversiteit Groningen, 2004.

R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis, capítulo 12*, 3rd ed., Prentice Hall, 1992.

S.Ó. Jónsdottir, F.S. Jørgensen and, S. Brunak. Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates. *Bioinformatics*, 21:2145-2160, 2005.

C.A. Lipinski, F. Lombardo, B.W. Dominy and P.J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Review*, 23:3–25, 1997.

H.E. Selick, A.P. Beresford and M.H. Tarbit. The emerging importance of predictive ADME simulation in drug discovery. *Drug Discovery Today*, 7, 2:109-116. 2002.

D.K. Slonim. From patterns to pathways: gene expression data analysis comes of age. *Nature genetics supplement*, 32, 2002.

S.S. So, M. Karplus. Evolutionary Optimization in Quantitative Structure-Activity Relationship: An Application of Genetic Neural Networks. *Journal of Medicinal Chemistry*, 39:1521-1530, 1996.

J. Taskinen and J. Yliruusi. Prediction of physcochemical properties based on neural networks modelling. Advanced Drug Delivery Reviews, 55:1163–1183, 2003.

I.V. Tetko, D.J. Livingstone and A.I. Luik neural Network Studies. 1. Comparison of Overfitting and Overtraining. *Journal of Chemical Information and Computer Science*, 35:826-833, 1995.

I.V. Tetko. Neural network studies. 4. Introduction to associative neural networks. *Journal of Chemical Information & Computer Sciences*, 42:717-728, 2002a.

I.V. Tetko. Associative Neural Network. *Neural Processing Letters*, 16, 2: 187-199, 2002b.

The Physical Properties Database (PHYSPROP) es sustentada por Syracuse Research Corporation (SRC), North Syracuse, USA. URL http://www.syrres.com/esc/.

R. Todeschini and V. Consonni. *Handbook of Molecular Descriptors*. Wiley-VCH, Weinheim (Germany), 2000.

J.G. Topliss and R.P. Edwards. Chance Factors in Studies of Quantitative Structure-Activity Relationships. *American Chemical Society*, 22,10:1238-1244, 1979.

J.H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236-244, 1963.

D.C. Whitley, M.G. Ford, D.J. Livingstone. Unsupervised Forward Selection: A Method for Eliminating Redundant Variables. *Journal of Chemical Information and Computer Science*, 40:1160-1168, 2000.

D.A. Winkler. Neural Networks in ADME and Toxicity Prediction, invited review. *Drugs of the Future*, 2004.

A. Yasri, D. Hartsough: Toward an Optimal Procedure for Variable Selection and QSAR Model Building. *Journal of Chemical Information and Computer Sciences*, 41: 1218-1227, 2001.