Reconocimiento de Patrones

JOSE ELGUERO
Instituto de Química Médica, CSIC
Juan de la Cierva, 3.- 28006 Madrid

1.0 INTRODUCCION

Este capítulo está basado en dos revisiones bibliográficas: una de 1979, de Kirschner y Kowalski (Kirschner, 1979), y otra de Dunn y Wold, de 1990 (Dunn, 1990). Dado que estos nombres corresponden a los autores más representativos en aplicaciones de reconocimiento de patrones al diseño de fármacos, sólo se añadirán algunas referencias a trabajos personales que, por su carácter menor, no han sido recogidos en las citadas revisiones. Excepcionalmente, algún trabajo muy significativo o muy reciente, no citado en ellas, también será comentado .

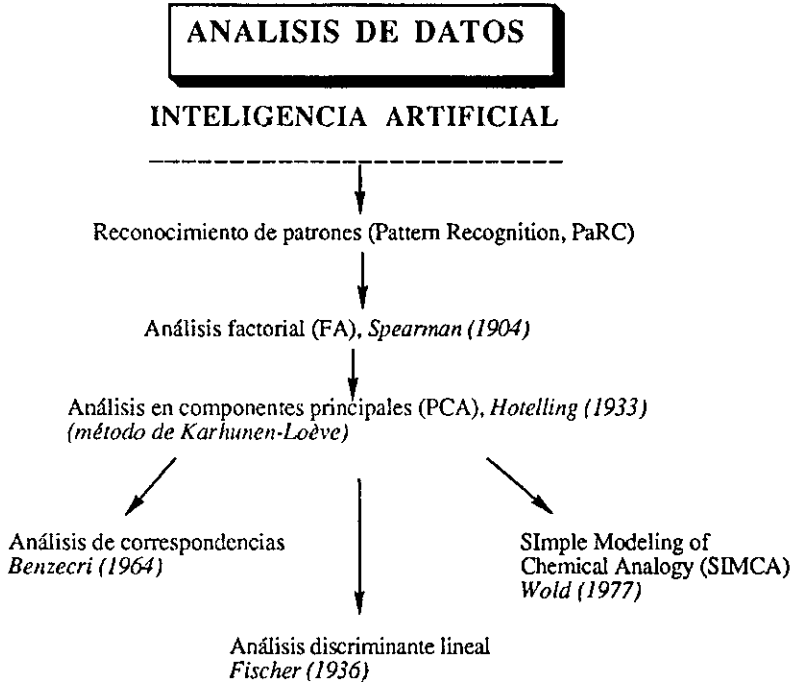
En el esquema 1 figura un ensayo de clasificación de la rama de la ciencia conocida como análisis de datos.

Como puede verse, se trata de métodos clásicos, bien conocidos en estadística, incluso su aplicación al diseño de fármacos tiene ya decenas de años de antigüedad (Martin, 1974). El desarrollo de la microinformática y la disponibilidad de numerosos programas, hacen de estos métodos un útil sencillo de manejar y, por lo tanto, algo que debe formar parte del arsenal de todo investigador en este campo.

Dado que el ser humano es el mejor "**reconocedor de patrones**" en un espacio de dos o tres dimensiones, todos los problemas reales se referirán a espacio de mayores dimensiones.

2.0. FUNDAMENTOS DE RECONOCIMIENTO DE PATRONES EN CUANTO A APLICACIONES EN DISEÑO DE FARMACOS Y METODOS QSAR

El objetivo principal de los métodos QSAR es encontrar modelos que predigan la actividad biológica de compuestos desconocidos o no ensayados. Además, si las predicciones son altamente significativas desde el punto de vista estadístico, deben proporcionar algún tipo de entendimiento acerca del efecto que produce la modificación de la estrucutra (dentro de una serie) sobre la modificación de la actividad biológica. Para que el objetivo de predicción se cumpla, dos condiciones que afectan al modelo y a los datos



Esquema 1.

deben cumplirse. El primero es que el modelo debe reproducir la actividad biológica tan bien como sea posible para los compuestos que han servido para establecerlo. En segundo lugar, las predicciones se deben hacer con el menor número de parámetros posibles comparados con el número de grados de libertad de los datos. Esta segunda condiciones es la más importante y la que más frecuentemente se olvida.

La aplicación del **reconocimiento de patrones** en diseño de fármacos está basada en la misma hipótesis sobre las relaciones entre estructura química y actividad biológica que los otros métodos QSAR: el principio de analogía de Hammett. Este principio dice: "desde sus orígenes la ciencia de la química orgánica ha dependido de la regla empírica y cualitativa según la cual las sustancias reaccionan de una manera similar y que cambios

similares de la estructura producen cambios similares de la reactividad". Según S. Wold, "la filosofía de la química está basada en gran parte en los conceptos de **similitud** y **analogía**. En este aspecto, la química difiere de la física y está mucho más cerca de la biología y de la medicina. Ello hace a la química un terreno de ensayos ideal para métodos de análisis en términos de **similitud**".

En **reconocimiento de patrones** se habla de objetos. Los objetos pueden pertenecer a dos conjuntos: el conjunto de aprendizaje (training set, learning set o reference set) y el conjunto de prueba (test set). La colección de objetos que poseen alguna propiedad conocida que interesa al investigador forman el conjunto de aprendizaje. Dichos objetos representan el conocimiento previo y, en consecuencia, deben ser seleccionados con el máximo cuidado. Los objetos sin clasificar u objetos para los cuales se desconoce la propiedad que interesa formarán parte del conjunto de prueba. Cada uno de los pasos del **reconocimiento de patrones** utilizará uno, otro o los dos conjuntos y se conocen como **fase de aprendizaje** y **fase de predicción**. He aquí una matriz de las usadas en métodos PaRc.

3.0. INFORMACION OBTENIDA A PARTIR DEL RECONOCIMIENTO DE PATRONES

En todos los métodos de clasificación es posible distinguir cuatro niveles:

Nivel 1.- Clasificar un objeto desconocido como perteneciente a una u otra clase.

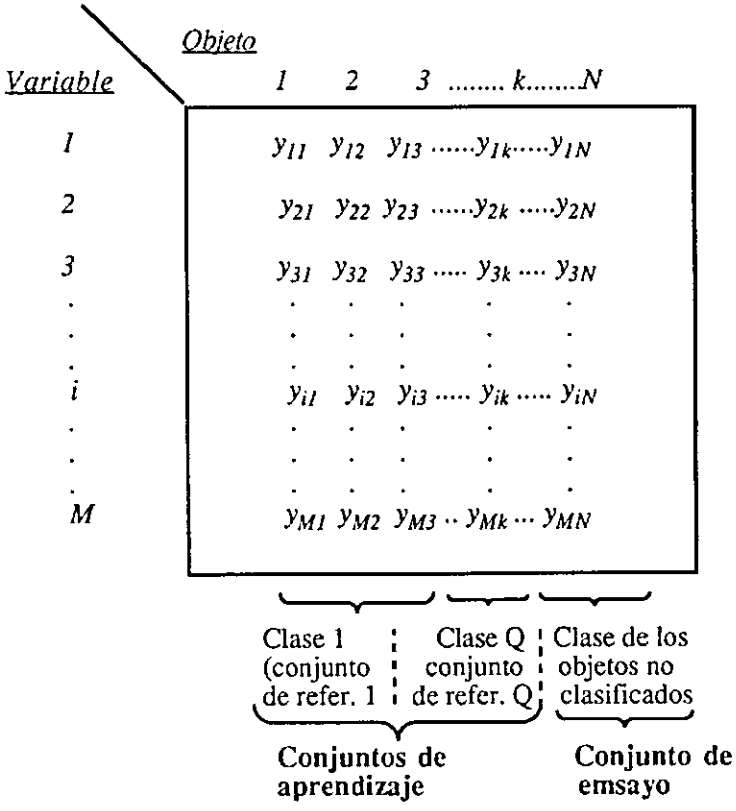
Nivel 2.- Lo anterior y además que tenga la posibilidad de que no sea miembro de ninguna de las clases definidad (es decir, que sea un "outlier" en la terminología de Wold).

Nivel 3.- Que ofrezca además la posibilidad de establecer relaciones cuantitativas entre la actividad (biológica) y la posición del objeto en el espacio de los factores **dentro de una clase.**

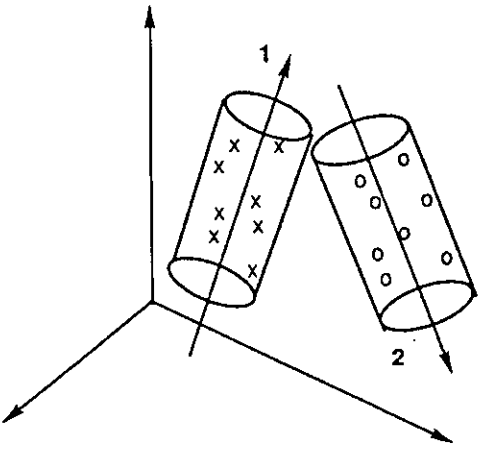
Nivel 4.- Posibilidad de utilizar varias respuestas simultáneamente. Este aspecto es muy importante en investigación farmacéutica, ya que lo que se desea optimizar es, a menudo, un compromiso entre varias respuestas (toxicidad-actividad; varias cepas; diferentes modelos animales, etc.).

4.0 LAS BASES GEOMETRICAS DEL RECONOCIMIENTO DE PATRONES

Para clasificar compuestos como miembros de diferentes clases es necesario establecer regiones del **espacio de patrones** ocupadas por los



compuestos de entrenamiento. Los métodos de **reconocimiento de patrones** difieren en cómo llevar a cabo la partición del espacio. Unos (**análisis discriminante, máquina de aprender lineal**) usan hiperplanos y buscan una línea, plano o hiperplano que separe una clase de otra. Los segundos métodos de **reconocimiento de patrones** son aquellos basados en distancias, entre ellos el método del **k-vecino más próximo.** finalmente, el tercer grupo está constituido por los métodos llamados de **modelización de clases** o métodos proyectivos. Tales métodos están basados en la proyección hacia estrucutras latentes (projections to latent structures, PLS) el más conocido delos cuales es el SIMCA de Wold. El esquema 2 representa el resultado de aplicar el método SIMCA a la clasificación de dos conjuntos: uno (cruces), con una determinada propiedad, y otro (círculos), con otra propiedad.Se ha elegido un espacio de tres dimensiones por conveniencia, pero hay que recordar que estos métodos se aplican a espacios multidimensionales con



Esquema 2.

dimensiones muy superiores a tres.

El cálculo de las desviaciones estándar (SD) de los residuales de cada clase permite definir un intervalo de confianza alrededor de cada modelo. Este intervalo (el cilindro) rodea cada clase.

5.0. ETAPAS EN UN ANALISIS POR RECONOCIMIENTO DE PATRONES.

Los estudios de PaRC se llavan a cabo por pasos. En la primera fase del estudio se determinan los objetivos del análisis y el nivel de información requerido. A continuación, se establecen los conjuntos de aprendizaje y se eligen las variables químicas necesarias para describir los compuestos. Se suelen tratar las variables de manera a transformarlas en variables homogéneas por técnicas de tipo autoescalado (una variable escalada tiene una media de cero y una varianza unidad). Se prueban modelos de clasificación, lo cual lleva a eliminar las variables cuya contribución no es importante. Usando el conjunto reducido de variables (es decir, todas aquellos significativas) se establecen nuevos modelos y se procede a su validación usando parte de los datos del conjunto de entrenamiento como si su actividad biológica fuera desconocida. Cuando el modelo se considera satisfactorio, se puede usar para clasificar nuevas sustancias. Eventualmente los nuevos resultados biológicos (que sean aciertos o fracasos de la predicción) se pueden incorporar al conjunto de entrenamiento, robusteciendo así el carácter predictivo del modelo.

6.0 LA NATURALEZA DE LAS VARIABLES QUE SE USAN EN LOS ESTUDIOS DE RECONOCIMIENTO DE PATRONES

La etapa esencial en los procedimientos de **reconocimiento de patrones** es la de selección de las variables que se usan para caracterizar un "objeto". El **reconocimiento de patrones** puede encontrar únicamente tanta información como se encuentra en las características seleccionadas; es decir, los únicos **patrones** que se pueden hallar son aquellos que están presentes en las variables. En primer lugar, las variables deben ser relevantes para el problema biológico considerado. Tales variables son (Manaut, 1993):

- El logaritmo del coeficiente de reparto (log P) (Taylor, 1990).
- La refractividad molar (Silipo, 1990; Taylor, 1990).
- La solubilidad (Taylor, 1990).
- El pK_a o pK_b (Bowden, 1990).
- Las constantes de sustituyente, como las σ de Hammett (Bowden, 1990).
- El parámetro π de Hansch y Fujita (Taylor, 1990).
- Los efectos estéricos, como es E_s de Taft o los B y L de Verloop (Silipo, 1990).
- Los fragmentos subestructurales (Silipo, 1990).
- Los índices de conectividad (Silipo, 1990).
- Los desplazamientos químicos del carbono-13 (Alcalde, 1991).

Como muchos de estos parámetros se pueden calcular fácilmente con un ordenador (ver, por ejemplo (Camilleri, 1993) para el cálculo de 254 propiedades moleculares mediante el programa GENPROP) es fácil tener grandes colecciones de variables. Su análisis con un método inapropiado, por ejemplo con el de los hiperplanos, suele conducir a modelos de escaso poder predictivo. Es bien conocido que tales variables suelen ser redundantes (colinearidad). Por ello es muy importante reducir tales variables a las llamadas **propiedades principales**, mediante un análisis en componentes principales.

Es práctica común en muchas aplicaciones PaRC el uso de variables de tipo "cero-uno" para describir la presencia-ausencia de fragmentos estructurales (conocidas en QSAR kcomo matrices *de novo*, de Free-Wilson o de Fujita-Ban (Elguero, 1982; Baumes, 1983; Cativiela, 1983)). La hipótesis de **continuidad** subyacente en la mayoría de los métodos PaRC, incluido el SIMCA, no se cumple bien con variables "cero-uno". Además, predicciones fuera de rango de substituyentes estudiados, son imposibles. Por ejemplo, no se puede utilizar el conocimiento químico de que un grupo nitro, en ciertos casos, se parece a un grupo ciano.

Los descriptores más cuantitativos, como algunos de los enumerados en la lista anterior, son preferibles. De todos modos, para evitar falsas espectativas, hay que considerar que los parámetros de sustituyente derivan todos de la influencia de dichos sustituyentes en sencillas reacciones de sencillos compuestos orgánicos. Teniendo eso en cuenta, su aplicabilidad a estudios de actividad biológica es sorprendentemente buena. El desarrollo sistemático de sistemas biológicos modelo y el estudio sistemático de la influencia de la estructura de los compuestos orgánicos en esos sistemas biológicos podría llevar a escalas biológicas de sustituyentes. Esto tendría un efecto mucho más trascendental sobre los métodos QSAR que la manipulación matemática de datos "históricos" es decir ya publicados.

Es frecuente que descriptores de los dos tipos sean utilizados simultáneamente. En un estudio de clasificación de sulfoniltioureas y tioureas como hiper o hipoglicemiantes (Dunn, 1980) se utilizaron π, MR (refracción molar) y σ_p para los sustituyentes en el anillo aromático, π y MR para los sustituyentes en el grupo urea, una variable indicadora (-1,0,1) para describir las propiedades electrónicas del anillo aromático y otra variable indicadora (0,1) para diferenciar ureas de tioureas. Finalmente, una variable de tipo R_M se usó para describir la lipofilia de la molécula entera.

Un breve comentario acerca de la respuesta biológica. Aunque ésta es frecuentemente cuantitativa, es bien conocido que está afectada de una gran incertidumbre. Algunos autores (Kirschner, 1979) proponen transformarla en una respuesta cualitativa, lo cual, a primera vista, no parece razonable. Consideremos un ejemplo imaginario. Supongamos una colección de moléculas que han sido ensayadas por su actividad y que los valores encontrados se hallan entre 0.0 (inactivas) y 100.0 (la más activa). Imaginemos que la replicación de los ensayos muestra una incertidumbre en las medidas de actividad del 20% (nada excepcional en ensayos biológicos in vivo). Si hay un número suficiente de moléculas se pueden construir tres categorías: muy activas (de 70 a 100), medianamente activas (de 30 a 70) e inactivas (de 0 a 30). El problema ahora es encontrar un patrón de características que permita clasificar las moléculas en esas tres clases. El químico le pide menos a los datos, pero los datos son más fiables. El procedimiento que hemos descrito está relacionado con los tests no-paramétricos. Tales tests son aplicables en más casos que los tests paramétricos, pero son menos exigentes con los datos.

7.0 ESTRUCTURA DE LOS DATOS EN PROBLEMAS DE RECONOCIMIENTO DE PATRONES

Una de las hipótesis básicas en reconocimiento de patrones es que cada conjunto de aprendizaje forma una agrupación ("cluster") bien definida

en el espacio de patrones. Eso es lo que se hace habitualmente cuando sólo se aplica el análisis QSAR a los compuestos activos, dejando de lado los inactivos (hipótesis subyacente, los inactivos pertenecen a otro "cluster"). Existen dos tipos de estructura de datos claramente diferenciados:

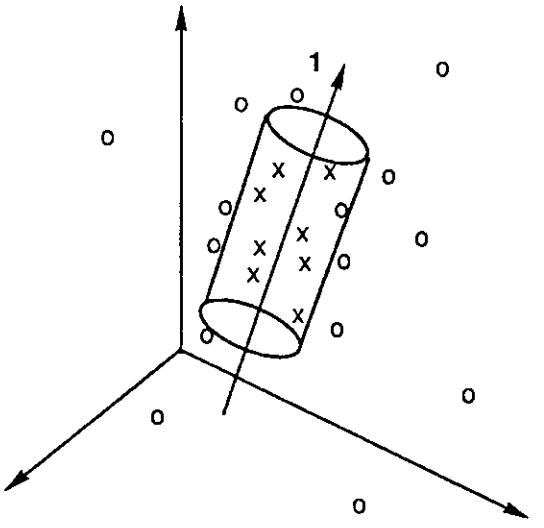
7.1.- Clases bien definidas: estructura de datos simétrica

El hecho de que compuestos que son química y farmacológicamente similares se agrupen en el espacio de los patrones es una consecuencia del principio de analogía de Hammett que hemos discutido previamente. Ejemplos de tal comportamiento se encuentran en compuestos que actúan sobre el mismo receptor, tales como agonistas y antagonistas, inhibidores y activadores enzimáticos, etc.

7.2.- El problema de los compuestos activos frente a los inactivos: estructura de datos asimétrica.

En numerosos casos de estudios sobre relaciones estructura-actividad biológica el problema de clasificación implica compuestos activos frente a compuestos inactivos, por ejemplo, tóxicos/no tóxicos, carcinogénicos/no carcinogénicos, etc. Estos casos han sido llamados con el nombre de **problema de la clasificación asimétrica.** Ello es debido a que una de las clases (la de los no activos) suele ser una clase mal definida e inhomogénea. Ello es fácil de entender, los hidrocarburos aromáticos policondensados forman una clase homogénea y bien definidad de sustancias cancerígenas, pero las sustancias no cancerígenas son un conjunto heteróclito de compuestos, desde el vidrio hasta el agua. Tales conjuntos tienen la apariencia del esquema 3.

En cierto sentido, todos los compuestos no activos (círculos) deben ser considerados "outliers" de la clase de compuestos activos (cruces). Esto lleva necesariamente a la conclusión que un compuesto que es clasificado como no-activo es no-activo **por el mismo mecanismos** que la clase de los activos (si es activo por otro mecanismo, entonces representa una nueva clase de compuestos activos).



Esquema 3.

8.0 APLICACIONES DE LOS METODOS DE RECONOCIMIENTO DE PATRONES EN QSAR Y EN DISEÑO DE FARMACOS

8.1.- Métodos generales

8.1.1. Análisis espectral

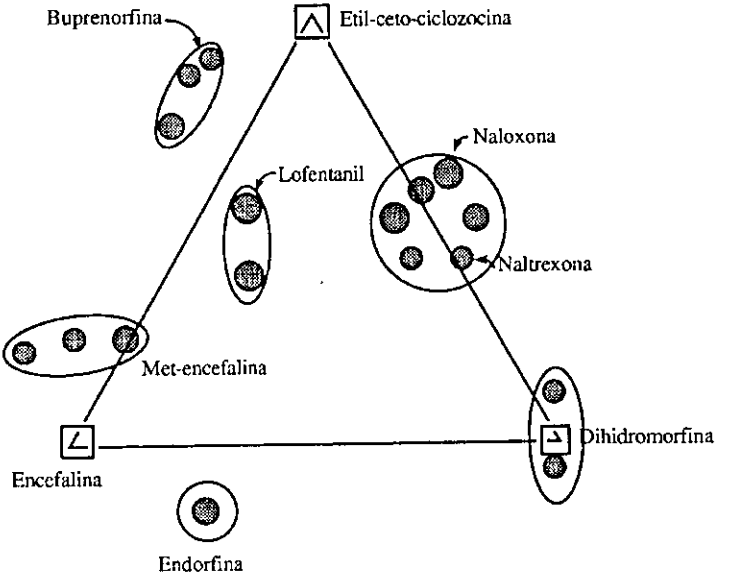
Paul Lewi, de los Laboratorios Janssen (Lewi, 1976; Calomme, 1984) ha desarrollado un método llamado "spectral mapping" (mapeo espectral) que ha aplicado con éxito considerable al análisis de resultados biológicos. El objeto del análisis multidimensional de datos en farmacología es encontrar relaciones entre la estructura de los compuestos ensayados y su actividad biológica. En un segundo paso, las hipótesis resultantes del análisis serán verificadas con nuevos ensayos. Tales datos se presentan generalmente en forma de tablas cuyas filas corresponden a los compuestos o fármacos y cuyas columnas corresponden a organismos vivos, órganos aislados, tejidos homogeneizados, etc.

Por ejemplo, si se analizan 26 narcóticos opioides en cuatro ensayos de afinidad (binding) frente a etilcetociclazocina, dihidromorfina, D-Ala,D-Leu encefalina y naloxona triaciadas (valores en CI_{50}). Tal tabla puede ser vista como las coordenadas de 26 puntos en un espacio de 4 dimensiones: cada compuesto ensayado corresponde a un punto y cada compuesto de

referencia está asociado con un eje. Los valores de CI_{50} serán las coordenadas en los 4 ejes. Como el ser humano tiene dificultades para visualizar los 26 puntos en el espacio de 4 dimensiones, Lewi usa su método (programa DATASCOPE) para reducir esa información a un espacio de 2 dimensiones, con la mínima pérdida de información (Esquema 4). La representación triangular, como en los clásicos problemas de mezclas o en los diagramas de fase, es una inteligente manera de obtener tres dimensiones en un plano, con la restricción de que la suma de las tres coordenadas tiene que ser uno.

El mapa espectral del esquema 4 contiene el 99% de la información. El análisis del mapa es función de los receptores κ, δ y μ, revela que hay muy pocos fármacos potentes frente al receptor κ, pero que algunos analgésicos muy potentes no se unen a tal receptor.

En el centro del triángulo aparecen los analgésicos "neutros" (clase $\mu\delta\kappa$) y las subclases $\mu\kappa$ y $\mu\delta$. Tales conclusiones eran imposibles de deducir del examen de la tabla original.



Esquema 4.

8.1.2. Análisis de correspondencias

Nosotros hemos usado el método de **análisis de correspondencia de Benzecri** para clasificar disolventes (Elguero, 1983). Se trata de un ejemplo

sencillo de cómo se aplican estos métodos, en este caso para clasificar disolventes. A partir del modelo tetraparamétrico de Koppel y Palm (Koppel, 1972) se encuentra un modelo de tres factores, de los cuales los dos primeros dan cuenta del 94% de la varianza de los datos. El análisis muestra, además, que los parámetros de acidez, basicidad, polaridad y polarizabilidad no son "puros", sino que cada uno contiene parte de la información de los otros.

8.1.3. Análisis discriminante lineal

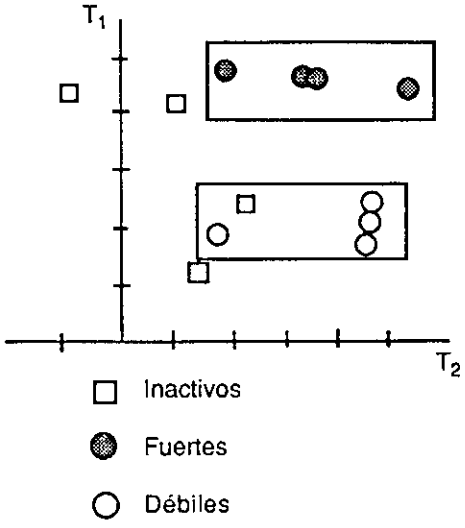
El **Análisis discriminante lineal** de Fischer ha sido aplicado a un conjunto de 66 compuestos (objetos), 30 sedantes y 36 tranquilizantes (Kirschner, 1979) utilizando como descriptores sus espectros de masas (hipótesis subyacente: tanto el espectro de masas como la actividad farmacológica están relacionadas con la estructura). Se eligieron 30 valores de m/z con sus respectivas intensidades, es decir cada uno de los 66 compuestos es un punto en un espacio de 30 dimensiones, cada relación m/z es una coordenada. Para verificar el buen funcionamiento del método de Fischer se recurre a un procedimiento muy frecuente en reconocimiento de patrones que se conoce con el nombre de **validación**: se quitan aleatoriamente, por ejemplo, dos compuestos; se usan los otros 64 como conjunto de aprendizaje y luego el método clasifica los dos restantes como clase 1 (sedantes) o clase 2 (tranquilizantes): ambos compuestos fueron clasificados correctamente. Se puede repetir el procedimiento tantas veces como se quiera. Si los resultados son satisfactorios, el carácter predictivo del modelo es elevado y cabe esperar que el espectro de masas de un producto, antes de ser evaluado farmacológicamente, permita decir si será sedante o tranquilizante (**nota:** sólo funcionará si se trata de una clase homogénea, ver más arriba, esquema 2).

8.1.4. Análisis factorial

El **análisis factorial** ha sido aplicado por Cammarata y Menon (Cammarata, 1976) para clasificar los agentes vasopresores en fuertes y débiles. El método está basado en la obtención de los vectores propios (eigenvectors) que den cuenta de la varianza de los datos. En el ejemplo citado Cammarata y Menon usaron un código topológico sencillo para describir una serie de aminas alifáticas y aromáticas. Dos vectores propios T_1 y T_2 (esquema 5) explican el 80% de la varianza y permiten una separación clara entre los dos niveles de actividad.

8.1.5. Análisis de autocorrelación

Guilles Moreau (Roussel-Ucalf) (Moreau, 1980; Moreau, 1988) ha introducido un método denominado **análisis de autocorrelación**. Este



Esquema 5.

método, conocido por sus iniciales inglesas ATS (Autocorrelation of the Topological molecular Structure) usa como descriptores propiedades de los átomos tales como conectividad, radio de van der Waals, densidad electrónica, etc., todos ellos calculables teóricamente. A partir de esos descriptores se contruye una matriz cuyo vector de autocorrelación se calcula. En su segundo trabajo (Moreau, 1988) este autor presenta un ejemplo de gran relevancia práctica. El objeto es detectar automáticamente nuevas moléculas activas (en el ejemplo citado, ansiolíticos) a partir de grandes colecciones de moléculas no ensayadas, caso frencuente en las empresas que, o bien adquieren moléculas de los laboratorios al azar o bien tienen conocimiento de tales moléculas por las publicaciones y sólo desean adquirir aquellas potencialmente activas en un campo farmacológico dado. En la fase de aprendizaje se logró unos porcentajes de acierto tanto en la clasificación de moléculas activas como en la de inactivas del orden del 85%. En la fase de predicción se calcularon los descriptores de 30.000 moléculas del catálogo de Roussel-Uclaf; la aplicación del método ATS reconoció como activas alrededor del 5% de ellas (unas 1.500), de las cuales 750 estaban disponibles y se ensayaron frente al receptor de las benzodiazepinas: 25 resultaron acticas (3%) porcentaje débil pero hay que tener en cuenta que corresponden a estructuras químicas muy diferentes de las de los ansiolíticos, razón por la cual nunca habían sido ensayadas (la probabilidad de encontrar una

actividad ansiolítica al azar es difícil de estimar, Moreau cita entre un 0,1 y un 1%).

8.1.6. Otros métodos

Diversos grupos han utilizado variantes de los métodos anteriores en sus estudios sobre clasificación de compuestos con actividad biológica. entre ellos podemos citar los trabajos de Horváth sobre análisis multivariado y su aplicación a quinazolinas inhibidoras de dihidrofolato reductasa (Chen, 1979); el método de mínimos cuadrados adaptable y su aplicación a guanidinas hipotensoras (Moriguchi, 1980); la aplicación del método de las nubes dinámicas de Diday por Grassy y colaboradores (Grassy, 1982) al caso de respuestas biológicas múltiples; finalmente el uso de los agrupamientos no-jerárquicos ("Nonhierarchic Cluster Analysis") para seleccionar compuestos destinados a ensayos biológicos (Willett, 1986).

8.2. Métodos de Wold (SIMCA) y PLS)

Una excelente presentación de estos métodos para químicos se encuentra en el libro de uno de los alumnos de Svante Wold (Carlson, 1992) (nosotros hemos utilizado estos métodos para la clasificación de cristales líquidos (Serrano, 1985) y en un estudio quimiométricos de heterociclos pentagonales nitrogenados (azoles) (Ebert, 1990)).

En la mayoría de los métodos PaRC (LDA, linear discriminant analysis; LLM, linear learning machine) el número de variables debe ser limitado, por ejemplo, si hay N objetos en el conjunto de aprendizaje, aproximadamente N/3 es el número máximo de variables permitido (en multiregresión, si el número de variables excede N/4, la regresión se vuelve altamente inestable). Esta limitación no se aplica al SIMCA. La principal diferencia entre el SIMCA y los otros métodos de PaRC es que en lugar de buscar diferencias entre clases, busca parecido dentro de cada clase.

El método SIMCA no opera al nivel 1, ya que la detección de elementos extraños (outliers) es automática. Está basado en el principio de analogía de Hammett y define una **región de analogía** en el espacio de patrones. Si un compuesto no está dentro de esta región (una por clase) es un elemento extraño. Al nivel 2, los datos del problema de **reconocimiento de patrones** eran los elementos de la matriz antes representada. Los niveles 3 y 4 de la clasificación de un compuestos se pueden obtener correlacionando la posición de un compuesto en el espacio fisicoquímico de su clase (dos

clases diferentes tienen variables significativas diferentes) por medio del método PLS (partial least squares).

Recientes aplicaciones de los métodos de Wold (que tienden a llegar a ser los métodos de referencia) se encuentran en un trabajo sobre caracterización de aminoácidos (Cocchi, 1993). Clementi y colaboradores (Baroni, 1993) han efectuado un excelente estudio sobre el uso del método PLS (GOLPE: Generating Optimal Linear PLS Estimations) para manipular problemas QSAR en 3 dimensiones. El método PLS también ha sido empleado con éxito para modelizar las relaciones estructura-actividad de inhibidores de catecol-O-metiltransferasa (Lotta, 1992).

8.3. Redes neuronales

En los últimos años hay una tendencia creciente a la aplicación de las redes neuronales (Artificial Neural Networks, ANN) en problemas de clasifición relacionados con métodos QSAR (Aoyama, 1990a; Aoyama, 1990b; Andrea, 1991; Tetko, 1993).

9.0. CONCLUSIONES

En lo que concierne al conjunto de entrenamiento, es necesario hacer una observación que proviene de otra área, la del diseño de experimentos (Box, 1978; Box, 1987; Atkinson, 1992; Carlson, 1992): si los objetos que forman el conjunto de entrenamiento están mal elegidos con respecto al modelo postulado, todos los esfuerzos serán inútiles. cuanto más se acerquen los puntos a un **diseño óptimo**, más fiables serán las consecuencias que saquemos de su estudio (Elguero, 1981). Si el conjunto nos es dado y nos apercibimos que está mal construido, será necesario añadir algunos objetos más elegidos con un criterio de optimalidad (Pahn Tan Luu, 1992). Para dos aplicaciones de los métodos modernos de construcción de matrices de experimentos óptimas en diseño de fármacos, ver las siguientes poblicaciones: Cativiela, 1983; Claramunt, 1984.

En resumen, el conjunto de métodos estadísticos recogid son de gran utilidad en investigación farmacéutica. En efecto, estos métodos permiten clasificar objetos pertenecientes a un espacio multidimensional (n >3 pero, en general, mucho mayor) de tal manera que las diferentes clases en que se agrupan son aparentes por simple examen visual (espacio de dos o tres dimensiones).

Esta operación de **clasificación** es esencial y debe necesariamente preceder todo intento más cuantitativo (QSAR), que sólo tiene sentido dentro

de una **clase homogénea** (química y biológicamente). Dada la sencillez de su empleo, los investigadores en este campo no deben dudar en aplicarlos, pero deben tener en cuenta:

1) No se puede partir de cero. Es necesario reunir una colección de datos biológicos suficientes para que el conjunto de entrenamiento permita encontrar un criterio de clasificación satisfactorio (en el trabajo de ansiolíticos citado en la sección 7.1.5 (Moreau, 1988) los datos de 3.000 compuestos fueron utilizados en la fase de aprendizaje).

2) Si tienen éxito, el mérito final recaerá sobre el investigador, ya que estos métodos sólo funcionan si el conjunto de descriptores elegidos para caracterizar un compuesto químico son **relevantes para su actividad biológica.**

Agradecimientos: Agradezco a D. Francisco Caballero la ayuda prestada en la preparación de este manuscrito.

10.0 BIBLIOGRAFIA

(1) ALCALDE, E.; DINARÉS, I.; FRIGOLA, J. (1991) *Eur. J. Med. Chem.* 26: 633.
(2) ANDREA, T.A.; KALAYEH, H. (1991) *J. Med. Chem.* 34: 2824.
(3) AOYAMA, T.; SUZUKI, Y.; ICHIKAWA, H. (1990a) *J. Med. Chem.* 33: 905.
(4) AOYAMA, T.; SUZUKI, Y.; ICHIKAWA, H. (1990b) *J. Med. Chem.* 33: 2583.
(5) ATKINSON, A.C.; DONEV, A.N. (1992) *"Optimun Experimental Designs"*, Oxfords Science Publications, Clarendon Press, Oxford.
(6) BARONI, M.; COSTANTINO, G.; CRUCIANI, G.; RIGANELLI, D.; VALIGI, R.; CLEMENTI, S. (1993) *Quant. Struct.-Act. Relat* 12:9.
(7) BAUMES, R.; TIEN DUC, H.N.C.; ELGUERO, J.; FRUCHIER, A. (1983) *An. Quím.* 79C, 128.
(8) BOWDEN, K. (1990) *"Electronic Effects in Drugs"* en Comprehensive Medicinal Chemistry (Hansch, C.; Sammes, P.G.; Taylor, J.B. Eds.), Pergamon Press, Oxford, 4: 205.
(9) BOX, G.E.P.; HUNTER, W.G.; HUNTER, J.S. (1978) *"Statistics for Experimenters"*, Wiley, New York.
(10) BOX, G.E.P.; DRAPER, N.R. (1987) *"Empirical Model-Building and Response Surfaces"*, Wiley, New York.
(11) CALOMME, G.; LEWI, P. (1984) *"Multivariate Analysis of Sctructure-Activity Data. Spectral Map of Opiod Narcotics in Receptor Binding"*, Actualités de Chimie Thérapeutique, 11ème Série, 121.
(12) CAMILLERI, P.; LIVINGSTONE, D.J.; MURPHY, J.A.; MANALLACK, D.T. (1993) *J. Comput Aided Mol. Design* 7: 61.
(13) CAMMARATA, A.; MENON, G.K. (1976) *J. Med. Chem.* 19: 739.
(14) CARLSON, R. (1992) *"Design and optimization in organic synthesis"*, Elsevier, Amsterdam.
(15) CATIVIELA,C.; ELGUERO, J.; MATHIEU, D.; MELÉNDEZ, E.; PHAN TAN LUU, R. (1983) *Eur. J. Med. Chem* 18: 359.
(16) CHEN, B.K.; HORVÁTH, C.; BERTINO, J.R. (1979) *J. Med. Chem.* 22: 483.
(17) CLARAMUNT, R.M.; ELGUERO, J.; MATHIEU, D.; PHAN TAN LUU, R. (1984) *An. Quím.* 80C: 30.
(18) COCCHI, M.; JOHANSSON, E. (1993) *Quant.Struc.-Act. Relat.* 12: 1.

(19) DUNN, W.J.; WOLD, S. (1990) *"Pattern Recognition Techniques in Drug Design"* en Comprehensive Medicinal Chemistry (Hansch,C.; Sammes, P.G.; Taylor, J.B. Eds.), Pergamon Press, Oxford, 4: 691.
(20) EBERT, C.; ELGUERO, J.; MUSUMARRA, G. (1990) *J. Phys. Org. Chem.* 3: 651.
(21) ELGUERO, J. (1981) *An. Real Acad. Farm.* 47: 137.
(22) ELGUERO, J.; FRUCHIER, A. (1982) *Afinidad* 39: 548.
(23) ELGUERO, J.; FRUCHIER, A. (1983) *An. Quím.* 79: 72.
(24) GRASSY, G.; TEULADE, J.C.; CHAPAT, J.P.; SIMEON DE BUOCHBERG, M.; ATTISO, M. (1982) *Eur. J. Med. Chem.* 17: 109.
(25) KIRSCHNER, G.L.; KOWALSKI, B.R. (1979) *"The Applications of Pattern Recognition to Drug Design" Drug. Design* 8: 73.
(26) KOPPEL, I.A.; PALM, V.A. (1972) *"The Influence of the Solvent on Organic Reactivity"* en Advances in Free Energy Relationships (Chapman, N.B.; Shorter, J. Eds.), Plenum Press, London.
(27) LEWI, P.J. (1976) *Drug Design* 7: 209.
(28) LOTTA, T.; TASKINEN, J.; BÄCKSTRÖM, R.; NISSINEN, E. (1992) *J. Comput.-Aided Mol. Design* 6: 253.
(29) MANAUT, F.; CLARAMUNT, R. (1993) "Parámetros o factores descriptores de las propiedades fisicoquímicas de los compuestos orgánicos y relaciones cuantitativas estructura-actividad (QSAR) en el diseño de fármacos" en Introducción a la Química Farmacéutica (Avendaño, C. Ed.), Interamerican-McGraw-Hill, p.73.
(30) MARTIN, Y.C.; HOLLAND, J.B.; JARBOE, C.H.; PLOTNIKOFF, N. (1974) *J. Med. Chem.* 17: 409.
(31) MOREAU, G.; BROTO, P. (1980) *New. J. Chem.* 4: 757.
(32) MOREAU, G.; BROTO, P.; FORTIN, M.; TURPIN, C. (1988) *Eur. J. Med. Chem.* 23: 275.
(33) MORIGUCHI, I.; KOMATSU, K.; MATSUSHITA, Y. (1980) *J. Med. Chem.* 23: 20.
(34) PHARN TAN LUU, R.; MATHIEU, D. (1993). Conjunto de programas NEMROD, Universidad de Aix-Marsella III, Francia.
(35) SERRANO, J.L.; MARCOS, M.; MELÉNDEZ, E.; ALBANO, C.; WOLD, S.; ELGUERO, J. (1985) *Acta Chem. Scand.* B39: 329.
(36) SILIPO, C.; VITTORIA, A. (1990) *"Three-Dimensional Structure of Drugs"* en Comprehensive Medicinal Chemistry (Hansch, C.; Sammes, P.G.; Taylor, J.B. Eds.), Pergamon Press, Oxford 4: 153.
(37) TAYLOR, P.J. (1990) *"Hydrophobic Properties of Drugs"* en Comprehensive Medicinal Chemistry (Hansch, C.;Sammes, P.G.; Taylos, J.B. Eds.) Pergamon Press, Oxford 4:241.
(38) TETKO, I.V.; LUIK, A.I.; PODA, G.I. (19993) *J. Med. Chem.* 36: 811.
(39) WILLETT, P.; WINTERMAN, V.; BAWDEN, D. (1986) *J. Chem. Inf. Computer Sci.* 26: 109.