Bol. Soc. Esp. Mat. Apl.
nº 39(2007), 135–149

TOPOLOGíA MOLECULAR

J.M. AMIGÓ*, A. FALCÓ**, J. GÁLVEZ† Y V. VILLAR‡

*Centro de Investigación Operativa, Universidad Miguel Hernández de Elche
**Dpto. de Ciencias Físicas, Matemáticas y de la Computación, Universidad CEU Cardenal Herrera
†Dpto. de Química Física, Universitat de Valencia
‡Dpto. de Fisiología, Farmacología y Toxicología, Universidad CEU Cardenal Herrera

jm.amigo@umh.es, afalco@uch.ceu.es, jorge.galvez@uv.es, vicente.villar@uv.es

Resumen

Este artículo pretende dar una vista panorámica de la topología molecular, que es una aplicación de la teoría de grafos muy utilizada en la industria química y, sobre todo, en la farmacéutica, pero poco conocida por la comunidad matemática. El objetivo de la topología molecular es la caracterización estructural de moléculas mediante unos invariantes sencillos, llamados índices topológicos. Estos índices, una vez procesados estadísticamente, juegan un papel decisivo en el descubrimiento de nuevas aplicaciones de moléculas conocidas y en el diseño de moléculas con propiedades químicas y farmacológicas específicas. En el apartado de aplicaciones nos limitamos, por su importancia, a las farmacológicas.

Palabras clave: *Topología molecular, Teoría de grafos, Análisis discriminante*
Clasificación por materias AMS: *92E10*

1 Introducción

La "irrazonable eficacia de las matemáticas en las ciencias naturales" es una característica de la historia de la ciencia y de la técnica modernas que no ha hecho sino acentuarse aún más, si cabe, desde que el físico matemático Eugene Wigner (1902-1995), introductor de la teoría de grupos en la física atómica, titulara así su famoso artículo [22]. No hace falta decir que la sorpresa de Wigner (y de tantos otros científicos y pensadores) se debe al hecho de que las matemáticas, "la más pura de las ciencias", es una creación del intelecto que, en principio, no busca aplicación alguna. Nuestro objetivo en este artículo es, precisamente, ilustrar esta eficacia con una aplicación de la teoría de grafos a las ciencias de la salud y, más concretamente, al descubrimiento de nuevos fármacos y al diseño de moléculas para aplicaciones terapéuticas específicas.

Fecha de recepción: 18/05/2007

Esta disciplina, que lleva el nombre de *topología* (o *conectividad*) *molecular*, es todavía joven y a su desarrollo han contribuido significativamente la industria fisico-química y, sobre todo, la bio-química, ya que la posibilidad de predecir las propiedades de una molécula antes de sintetizarla permite ahorrar tiempo y dinero en la investigación farmacéutica, agrícola, ganadera, etc. Aunque el ámbito de aplicación de la topología es amplio, nosotros nos centraremos, por concreción e importancia, en las aplicaciones farmacológicas.

Recordemos, de paso, que la teoría de grafos es un magnífico exponente de matemática pura que ha encontrado con el tiempo un abanico de aplicaciones. Desarrollada en el siglo XIX por los matemáticos ingleses A. Cayley y J.J. Silvester (aunque fue Leonhard Euler, ¿quién si no?, el que había iniciado su estudio un siglo antes), la teoría de grafos se ha convertido en una herramienta imprescindible en las muchas áreas de la actividad científica y técnica en las que estructura y conectividad juegan un papel relevante. Por citar sólo unos pocos ejemplos: redes de comunicación y transporte, diseño de circuitos eléctricos (por ejemplo, en computadores), optimización de líneas de suministro, epidemiología, etc. Valga, pues, este artículo también como un pequeño tributo de reconocimiento y admiración a Euler (1707-1783) en el tercer centenario de su nacimiento.

Esencialmente, la topología molecular sirve para encontrar correlaciones entre una propiedad física, química o biológica y estructuras moleculares, basándose en la caracterización numérica de éstas mediante unos descriptores topológicos llamados *índices topológicos*. Supongamos, por ejemplo, que buscamos nuevos compuestos con determinada actividad farmacológica. Una vez calculados los índices topológicos de compuestos conocidos con dicha propiedad, se obtienen (por lo común, mediante análisis lineal discriminante) funciones de clasificación que permitan discriminar entre compuestos activos e inactivos. A continuación, las funciones de clasificación se aplican a bases de datos de estructuras químicas, para la selección de sustancias potencialmente activas. Finalmente, se realizan los ensayos experimentales *in vivo* o *in vitro* encaminados a confirmar la actividad predicha. Cuando se proponen nuevos índices topológicos, se estudia si los nuevos índices suponen una mejora en la funciones de clasificación, es decir, si las nuevas funciones demuestran mayor eficacia en la predicción de la actividad farmacológica objeto de estudio.

El origen de la topología molecular hay que buscarlo en los años 1970, cuando Kier y Hall y otros investigadores comenzaron a utilizar 'índices' derivados de las propiedades de conectividad de las moléculas para estudiar algunas propiedades fisico-químicas (como el calor de formación y la temperatura de ebullición) de compuestos orgánicos. Además de la topología molecular, hay actualmente tres metodologías en diseño molecular, caracterizadas por la técnica empleada: (i) QSAR (*quantitative structure activity relantionship*), basada en descriptores moleculares físico-químicos; (ii) mecánica molecular, basada en la mecánica clásica y en programas modeladores-constructores de representación gráfica tridimensional de moléculas; (iii) mecánica cuántica, que tiene en cuenta las posiciones y energías de los átomos y moléculas. Cuando se trata de diseño *ex novo*, es decir, de cabezas de serie para el diseño de nuevos fármacos, las

técnicas mecanocuánticas y moleculares requieren un conocimiento previo de las estructura del receptor biológico sobre el que la molécula va a interaccionar. Precisamente, la característica que distingue más claramente a la topología molecular de estas dos técnicas, es el hecho de no requerir un conocimiento previo de la estructura del receptor para generar nuevos cabezas de serie. A ello hay que añadir que la topología molecular aventaja también a los métodos QSAR por su mayor simplicidad.

Este artículo está divido en dos partes. En la primera (Sección 2), introducimos aquellos conceptos básicos de teoría de grafos que necesitamos para nuestra exposición y repasamos algunos de los índices topológicos más importantes, en particular, los de Hosoya, Kier y Hall, y Randić, que cuentan entre los pioneros. En la segunda parte (Sección 3), dedicada a las aplicaciones de la topología molecular en farmacología, nos detenemos primero, por completitud, a recordar cómo se construyen funciones de clasificación en análisis lineal discriminante, antes de presentar, con distinto grado de detalle, unos pocos ejemplos reales. Por supuesto, la lista de ejemplos se podría haber alargado mucho más, hasta incluir estudios en curso sobre el cáncer y el sida, dos de los grandes retos de la medicina moderna. Pero nuestro propósito no es abrumar al lector con datos, sino darle una descripción sencilla de la topología molecular y de alguna de sus aplicaciones, a la vez que mostrarle el alto grado de interdisciplinaridad que su práctica conlleva.

2 Las bases de la topología molecular: índices topológicos.

Como dijimos en la Introducción, la topología molecular se basa en la aplicación de teoría de grafos [4] a la descripción de las estructuras moleculares, siendo un *grafo* un conjunto de puntos (llamados *nodos* o *vértices*) con algunos pares de ellos conectados mediante uniones llamadas *aristas* o *ejes*. Los nodos del grafo G los numeraremos arbitrariamente y denotaremos por e_{ij} al eje que une los nodos i y j. Utilizaremos N para denotar el número de nodos de un grafo, mientras que $E(G)$ denotará el conjunto de ejes y $|E(G)|$ su cardinalidad (es decir, el número de ejes de G). Si dos nodos están conectados por un eje, se llaman *adyacentes*. El número de ejes que salen de un nodo dado se llama *grado* del nodo. Un *camino* o *trayectoria* p en G es el subgrafo obtenido al conectar consecutivamente varios nodos adyacentes; si, además, conectamos el primer punto y el último punto de p, obtenemos un *ciclo*. La *longitud* de un camino o ciclo es el número de ejes que lo componen. Cuando la teoría de grafos se aplica a moléculas, los nodos representan átomos y las aristas, enlaces químicos, normalmente enlaces covalentes puesto que es en la química orgánica donde la topología molecular ha encontrado su mayor campo de aplicación. El grafo resultante, que nos dice cómo están ligados los átomos y el camino (o caminos) que une(n) un átomo a otro en la misma molécula, se llama *grafo molecular*.

Supongamos, pues, que queremos caracterizar estructuralmente un compuesto orgánico. En el procedimiento general, se empieza eliminando los átomos de hidrógeno de la molécula; hay formulaciones en las que esto no es

así pero, por simplicidad, nosotros no las consideraremos aquí. En segundo lugar, los átomos restantes (los vértices del grafo molecular) se numeran de forma conveniente. Por último, la caracterización estructural contenida en el grafo molecular puede ser, a su vez, encapsulada de muy diversas maneras como, por ejemplo, mediante matrices, índices numéricos, polinomios, espectros, grupos u operadores. En esta sección daremos una idea general de las herramientas más sencillas utilizadas en topología molecular.

2.1 Matrices asociadas a grafos moleculares

La *matriz de adyacencia* es quizá el instrumento algebraico más sencillo en nuestro contexto ya que únicamente da información sobre qué pares de nodos están unidos mediante aristas. Si la molécula en cuestión consta, pues, de N átomos, la matriz de adyacencia del grafo molecular G, $\mathbf{A} = \mathbf{A}(G)$, es una matriz $N \times N$ simétrica cuyas componentes son

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{si los átomos } i, j \text{ están ligados,} \\ 0 & \text{en caso contrario.} \end{cases}$$

La Figura 1 muestra el grafo de la molécula 2,3-dimetil-butano y su matriz de adyacencia $\mathbf{A} = (\mathbf{A}_{ij})_{1 \le i,j,N}$.

Si, como ocurre con el 2,3-dimetil-butano, la molécula sólo tiene enlaces simples, entonces la suma de todos los elementos de la fila i de $\mathbf{A}$, $\sum_{j=1}^{N} \mathbf{A}_{ij}$, así como la suma de todos los elementos de la columna i, $\sum_{j=1}^{N} \mathbf{A}_{ji}$, nos dan indistintamente el número total de ejes que confluyen en el nodo o átomo i, es decir, el grado (o *valencia topológica*) del nodo i, que denotaremos por $\deg_i$. Si $\mathbf{DEG} = \mathbf{DEG}(G) = (\mathbf{DEG}_{ij})_{1 \le i,j \le N}$ es la matriz diagonal de componentes $\mathbf{DEG}_{ij} = \deg_i \delta_{ij}$ (donde δ_{ij} es la delta de Kronecker, es decir, $\delta_{ij} = 0$ si $i \neq j$ y $\delta_{ii} = 1$), entonces la *matriz Laplaciana* del grafo G, $\mathbf{L} = \mathbf{L}(G) = (\mathbf{L}_{ij})_{1 \le i,j,N}$, se define como [14]

$$\mathbf{L}(G) = \mathbf{DEG}(G) - \mathbf{A}(G).$$

A partir de la matriz de adyacencia $\mathbf{A}$ (y de otras matrices relevantes, como la Laplaciana y algunas que veremos enseguida), la teoría de grafos nos enseña cómo obtener una serie de herramientas algebraicas muy útiles para caracterizar G. Aquí consideraremos sólo dos de entre las más sencillas: (i) el polinomio característico,

$$p_{\mathbf{A}}(x) = \det(\mathbf{A} - x\mathbf{I}),$$

donde $\mathbf{I}$ denota la matriz unidad $N \times N$, y (ii) el espectro (es decir, el conjunto de valores propios λ_k de la matriz $\mathbf{A}$, $p_{\mathbf{A}}(\lambda_k) = 0$), si bien sólo suelen utilizarse los valores propios máximo $\lambda_{\text{máx}}$ (que caracteriza la ramificación o *branching* de G) y mínimo $\lambda_{\text{mín}}$. Puede probarse fácilmente que $\lambda_{\text{mín}} < 0 < \lambda_{\text{máx}}$. Digamos de paso que el polinomio característico de la matriz Laplaciana, $p_{\mathbf{L}}(x)$, se llama *polinomio Laplaciano*.

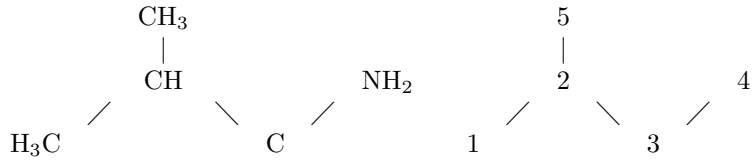
La *matriz de distancia*, $\mathbf{D} = \mathbf{D}(G)$, es una matriz $N \times N$ simétrica cuyas componentes son las 'distancias topológicas'

$$\mathbf{D}_{ij} = \begin{cases} \text{longitud mínima de los caminos que unen } i \text{ y } j, \text{ si } i \neq j, \\ 0, \text{ si } i = j. \end{cases}$$

Así, pues, $\mathbf{D}$ proporciona una imagen cualitativa de las relaciones de proximidad o lejanía entre los átomos de la molécula. La suma de las distancias topológicas entre el vértice i y todos los demás vértices del grafo molecular, se llama la *suma de distancias* del vértice i:

$$\mathbf{DS}_i = \sum_{j=1}^{N} \mathbf{D}_{ij} = \sum_{j=1}^{N} \mathbf{D}_{ji}. \tag{1}$$

Obsérvese, finalmente, que la matriz de distancia puede obtenerse a partir de la matriz de adyacencia. La Figura 1 da asimismo la matriz $\mathbf{D} = (\mathbf{D}_{ij})_{1 \leq i,j,N}$ para la molécula 2,3-dimetil-butano.



$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 0 & 1 & 2 & 3 & 2 \\ 1 & 0 & 1 & 2 & 1 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 3 \\ 2 & 1 & 2 & 3 & 0 \end{bmatrix}$$

Figura 1: El grafo de la molécula 2,3-dimetil-butano y sus matrices de adyacencia y de distancia.

Otras matrices utilizadas en topología molecular incluyen la matriz $\chi = \chi(G) = (\chi_{ij})_{1 \leq i,j \leq N}$ [18], donde

$$\chi_{ij} = \begin{cases} (\deg_i \cdot \deg_j)^{-1/2} & \text{si } e_{ij} \in E(G), \\ 0 & \text{en caso contrario,} \end{cases}$$

la matriz de distancias recíprocas [9], la matriz *detour* [20], la matriz de resistencia [12], etc.

2.2 Índices topológicos

La finalidad de los *índices topológicos* es codificar información topológica sobre las moléculas de forma puramente numérica. Este formato facilita enormemente

la búsqueda automatizada de moléculas con propiedades estructurales comunes y, por tanto, posibles candidatos a compartir propiedades químicas o farmacológicas deseadas. La relación entre grafos e índices topológicos no es unívoca, de manera que dado un índice topológico o un conjunto de ellos, en general no es posible identificar el grafo molecular correspondiente; esto es lo que se llama el *problema de degeneración*. Es esta degeneración, justamente, la que permite identificar grupos de moléculas con propiedades comunes mediante índices topológicos. El adjetivo 'topológico' hace referencia a que la información contenida en los índices es invariante a propiedades 'no esenciales', como pueden ser la numeración de los nodos, las distancias reales entre átomos o todas aquellas distorsiones de la molécula que no modifiquen el grafo subyacente. De la gran cantidad de índices topológicos que han sido propuestos en la literatura (véase, por ejemplo, [2]), aquí nos conformaremos con presentar una selección de aquéllos que han probado ser más útiles en la práctica, agrupados por el tipo de información que contienen. Por brevedad, nos limitaremos a índices basados en las matrices de adyacencia y de distancia.

2.2.1 Índices basados en la matriz de adyacencia

Randić introdujo en 1975 un primer número de conectividad χ_R, llamado *índice de Randić*, para caracterizar la ramificación del grafo molecular [16]. Si $E(G)$ denota el conjunto de aristas del grafo G, sea $f : E(G) \to \mathbb{R}$ la función

$$f(e_{ij}) = \left(\deg_i \cdot \deg_j\right)^{-1/2}. \tag{2}$$

Entonces, χ_R se define como

$$\chi_R(G) = \sum_{e_{ij} \in E(G)} f(e_{ij}).$$

Como consecuencia, a una mayor ramificación de la molécula, le corresponde un valor menor de χ_R.

Posteriormente, Randić introdujo un segundo número de conectividad, llamado *número de identificación ID* [17]. Dado el camino p de longitud m en G, definimos la función f^* del conjunto de caminos en G con valores en $\mathbb{R}$ como

$$f^*(p) = \prod f(e_{ij}), \tag{3}$$

donde el producto es sobre las aristas e_{ij} que componen el camino p. Entonces,

$$ID(G) = N + \sum_p f^*(p), \tag{4}$$

donde N es el número de vértices del grafo.

El índice de Randić fue generalizado por Kier y Hall [10, 11]. Para definir los *índices* ${}^m\chi$ *de Kier y Hall*, se descompone primero el grafo molecular en subgrafos con m aristas contiguas. Entonces,

$${}^m\chi(G) = \sum \left(\deg_{i_1} \cdot \deg_{i_2} \cdot \ldots \cdot \deg_{i_{m+1}}\right)^{-1/2}, \tag{5}$$

donde la suma es sobre todos los caminos en G formados por m aristas $e_{i_1 i_2}$, ..., $e_{i_m i_{m+1}}$. Obsérvese que ${}^1\chi = \chi_R$. En particular, se denotan mediante

$$ {}^m\chi_p,\ {}^m\chi_c,\ {}^m\chi_{pc},\ {}^m\chi_{ch}, \ldots \tag{6} $$

los índices de Kier y Hall obtenidos mediante subgrafos del tipo *path*, *cluster* (estrella), *path/cluster*, *chain* (polígonos), etc.

2.2.2 Índices basados en la matriz de distancia

Los índices basados en la matriz de distancia se llaman a veces *índices geométricos*, para distinguirlos de aquéllos derivados de la matriz de adyacencia.

El *índice de Wiener* W fue definido originalmente como el número de enlaces entre todos los pares de átomos de una molécula acíclica [21]. La definición de W basada en la matriz de distancia $\mathbf{D}$ fue propuesta por Hosoya [19] como la semisuma de los elementos no-diagonales de la matriz $\mathbf{D}$:

$$ W(G) = \frac{1}{2}\sum_{i,j=1}^{N} \mathbf{D}_{ij} = \sum_{i<j} \mathbf{D}_{ij} = \sum_{i>j} \mathbf{D}_{ij}, $$

(ya que $\mathbf{D}_{ii} = 0$). El índice de Wiener es uno de los índices topológicos más estudiado debido a su simplicidad y buena correlación con las propiedades estructurales del grafo molecular. Sin embargo, su degeneración es relativamente alta, lo que obliga a utilizarlo con otros índices (por ejemplo, con los índices de información que veremos más abajo).

Por su parte, Hosoya introdujo asimismo el índice que actualmente se conoce como *índice de Hosoya* [19]:

$$ Z(G) = \sum_{k} n(G,k), $$

donde $n(G,k)$ es el número de maneras en que podemos elegir k aristas no-adyacentes de G. Por definición, $n(G,k) \equiv 1$ y $n(G,1) = |E(G)|$.

El *índice de Balaban* [1] del grafo molecular G se define por la fórmula

$$ J(G) = \frac{|E(G)|}{\mu + 1} \sum_{e_{ij} \in E(G)} (\mathbf{DS}_i \cdot \mathbf{DS}_j), \tag{7} $$

donde $\mathbf{DS}_i$ y $\mathbf{DS}_j$ denotan las sumas de distancias (1) de los vértices i y j, respectivamente, extremos del eje $e_{ij} \in E(G)$ y μ denota el número de ciclos de G. Se ha probado analítica y computacionalmente que el índice de Balaban tiene muy poca degeneración.

Fue también Balaban quien sugirió sustituir en la definición del número de identificación de Randić 4, la función $f(e_{ij})$, ec. (2), por

$$ g(e_{ij}) = (\mathbf{DS}_i \cdot \mathbf{DS}_j)^{-1/2} $$

y la función $f^*(p)$, ec. (3), por

$$g^*(p) = \prod g(e_{ij}).$$

De forma análoga a (4), se define entonces el *número de identificación selectivo* SID como

$$SID(G) = N + \sum_p g^*(p).$$

Como indica su nombre, se ha encontrado que el índice SID es altamente selectivo [3].

Para concluir este breve listado, mencionaremos, además, el *índice molecular topológico* MTI que se define a partir de las matrices $\mathbf{A}$ y $\mathbf{D}$. En primer lugar, se introduce el vector $\mathbf{E} = \mathbf{E}(G)$ de descriptores estructurales para vértices,

$$\mathbf{E}(G) = \mathbf{Deg}(G)(\mathbf{A} + \mathbf{D}), \tag{8}$$

donde $\mathbf{Deg}(G) = (\deg_1, ..., \deg_N)$ es el *vector de grados* de G (que identificamos con una matriz $1 \times N$) que multiplica matricialmente en (8) a la matriz $\mathbf{A} + \mathbf{D}$, de dimensiones $N \times N$. Entonces,

$$MTI(G) = \sum_{i=1}^{N} \mathbf{E}_i.$$

Este índice topológico, como todos los anteriores, ha sido estudiado intensamente en la literatura y probado ser muy útil en topología molecular.

2.2.3 Índices basados en la teoría de la información

Por último, repasaremos algunos índices geométricos, inspirados en la teoría de la información.

Los *índices de información* I_D^E y $\bar{I}_D^E$, que representan la información total y promedio sobre las distancias en el grafo molecular G con N vértices, se definen como

$$\begin{aligned} I_D^E(G) &= \frac{N(N-1)}{2} \log_2 \frac{N(N-1)}{2} - \sum_{k=1}^{l} d(G,k) \log_2 d(G,k), \\ \bar{I}_D^E(G) &= \frac{2I_D^E(G)}{N(N-1)} = -\sum_{k=1}^{l} \frac{2d(G,k)}{N(N-1)} \log_2 \frac{2d(G,k)}{N(N-1)}, \end{aligned}$$

donde $d(G,k)$ denota el número de pares de vértices en G separados una distancia k, y l es el 'diámetro' de G (es decir, el mayor elemento de la matriz de distancia $\mathbf{D}$). Por otra parte, los índices de información I_D^W y $\bar{I}_D^W$, que representan la información total y promedio sobre la distribución de distancias

en G, se calculan mediante las fórmulas

$$\begin{aligned} I_D^W(G) &= W\log_2 W - \sum_{k=1}^{l} d(G,k)k\log_2 k, \\ \bar{I}_D^W(G) &= \frac{I_D^W(G)}{W} = -\sum_{k=1}^{l} d(G,k)\frac{k}{W}\log_2\frac{k}{W}, \end{aligned}$$

donde W es el índice de Wiener del grafo G.

La lista de índices basados en la teoría de la información incluye también, entre otros varios, los *índices* U, V, X e Y, cuya definición precisa los siguientes descriptores locales: (i) la información local media sobre la magnitud de las distancias,

$$u.inf_i = -\sum_{i\neq j}^{N} \frac{\mathbf{D}_{ij}}{\mathbf{DS}_i}\log_2\frac{\mathbf{D}_{ij}}{\mathbf{DS}_i},$$

donde $\mathbf{DS}_i$ es la suma de distancias (1) del vértice i; (ii) la información local sobre la magnitud de las distancias,

$$v.inf_i = \mathbf{DS}_i\log_2\mathbf{DS}_i - u.inf_i;$$

(iii) la información local extendida media sobre la magnitud de las distancias,

$$y.inf_i = \sum_{i\neq j}^{N} \mathbf{D}_{ij}\log_2\mathbf{D}_{ij};$$

y (iv) la información local extendida sobre la magnitud de las distancias,

$$x.inf_i = \mathbf{DS}_i\log_2\mathbf{DS}_i - y.inf_i.$$

Finalmente,

$$\begin{aligned} U(G) &= \frac{|E(G)|}{\mu+1}\sum_{e_{ij}\in E(G)}(u.inf_i\cdot u.inf_j)^{-1/2}, \\ V(G) &= \frac{|E(G)|}{\mu+1}\sum_{e_{ij}\in E(G)}(v.inf_i\cdot v.inf_j)^{-1/2}, \\ X(G) &= \frac{|E(G)|}{\mu+1}\sum_{e_{ij}\in E(G)}(x.inf_i\cdot x.inf_j)^{-1/2}, \\ Y(G) &= \frac{|E(G)|}{\mu+1}\sum_{e_{ij}\in E(G)}(y.inf_i\cdot y.inf_j)^{-1/2}, \end{aligned}$$

donde, al igual que en (7), μ es el número de ciclos de G.

3 Aplicaciones

Esta sección pretende mostrar los procedimientos y aplicaciones de la topología molecular.

3.1 Predicción de propiedades específicas y selección de nuevas moléculas

El objetivo último de la topología molecular es el establecimiento de ecuaciones de correlación entre propiedades moleculares e índices topológicos. Para ello suele utilizarse el *análisis lineal discriminante*, introducido por R.A. Fisher en 1936, si bien algunos autores prefieren técnicas no lineales como, por ejemplo, redes neuronales. Nosotros nos limitaremos al análisis lineal por ser el más extendido y, por tanto, el mejor contrastado.

El objetivo del análisis discriminante es encontrar una función capaz de distinguir (o discriminar, de ahí el nombre) entre dos o más categorías o grupos de objetos. La capacidad discriminante se mide determinando el porcentaje de objetos clasificados correctamente dentro de cada grupo.

En la práctica, el análisis discriminante se realiza mediante alguno de los paquetes estadísticos disponibles en el mercado (SPSS, BMDP, statistica, etc.). La selección de los descriptores se basa en el parámetro F-Snedecor, y el criterio de clasificación es la menor distancia de Mahalanobis, que es la distancia entre el valor individual y el valor medio global que aparece en la ecuación de regresión. El programa estadístico elige las variables usadas en la computación de las funciones de clasificación (normalmente lineal) paso a paso: en cada una de estas fases, la variable que aporta más a la separación de los grupos se introduce en la función de discriminación, mientras que la que aporta menos, se elimina. La calidad de la función de discriminación se evalúa con el parámetro lambda de Wilks (llamado también U-estadístico), empleando el test de igualdad de las medias de grupo para las variables de la función de discriminación.

En el caso que nos ocupa, los descriptores son los índices topológicos y el objetivo es clasificar una molécula dada como 'buena' (o activa) o 'mala' (o inactiva) respecto de cierta propiedad farmacológica, empleando para ello una función lineal discriminante. Supongamos, pues, que se ha elegido un conjunto determinado de p índices topológicos, que denotaremos por $x_1, x_2, \ldots, x_p$. A continuación, seleccionamos una muestra proveniente de una base de datos de moléculas, de las que sabemos cuáles poseen determinada propiedad y cuáles no. De las moléculas que tengan dicha propiedad, diremos que están en la clase $\mathcal{G}$ (*good*) y, en caso contrario, diremos que están en la clase $\mathcal{B}$ (*bad*). Supondremos, además, que la propiedad farmacológica en cuestión es medible y, en consecuencia, que dicha característica viene determinada por el valor de una variable y que toma valores en un intervalo cerrado I. Por último, establecemos la hipótesis de trabajo de que la variable y se puede escribir como combinación lineal de los índices topológicos considerados:

$$y = c_1 x_1 + \cdots + c_p x_p.$$

La propuesta de Fisher es la siguiente. Denotemos por la igualdad matricial

$$\mathbf{y} = \mathbf{X}\mathbf{c} \tag{9}$$

al sistema de igualdades formado por las n observaciones realizadas, siendo $\mathbf{y} = (y_1, y_2\ , \ldots, y_n)^T \in \mathbb{R}^n$ (el superíndice T denota, como de costumbre, transposición, de manera que $\mathbf{y}$ es un vector columna), $\mathbf{X}$ una matriz $n \times p$ y $\mathbf{c} = (c_1, c_2\ , \ldots, c_p)^T \in \mathbb{R}^p$. Supongamos, además, que tenemos $n_\mathcal{G}$ muestras de la población $\mathcal{G}$ y $n_\mathcal{B}$ muestras de la población $\mathcal{B}$ ($n_\mathcal{G} + n_\mathcal{G} = n$). Entonces podemos escribir el sistema lineal (9) (poniendo primero las muestras 'buenas' y, a continuación, las 'malas') como

$$\begin{pmatrix} \mathbf{y}_\mathcal{G} \\ \mathbf{y}_\mathcal{B} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_\mathcal{G} \\ \mathbf{X}_\mathcal{B} \end{pmatrix} \mathbf{c}.$$

donde $\mathbf{X}_\mathcal{G}$ es una submatriz $n_\mathcal{G} \times p$ y $\mathbf{X}_\mathcal{B}$ es una submatriz $n_\mathcal{B} \times p$.

Sea $\mathbf{m}_\mathcal{G} \in \mathbb{R}^p$ (respectivamente, $\mathbf{m}_\mathcal{B} \in \mathbb{R}^p$) el vector de medias calculadas por columnas en $\mathbf{X}_\mathcal{G}$ (respectivamente, $\mathbf{X}_\mathcal{G}$). Definamos del mismo modo la media $\mathbf{m}$ calculada por columnas empleando $\mathbf{X}$. Denotemos por $\overline{y}_\mathcal{G}, \overline{y}_\mathcal{B}, \overline{y}$ las medias calculadas por columna empleando los vectores $\mathbf{y}_\mathcal{G}, \mathbf{y}_\mathcal{B}$ e $\mathbf{y}$, respectivamente. Entonces podemos escribir la dispersión total de las observaciones y_i, $1 \leq i \leq n$, como

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \mathbf{y}^T\mathbf{H}\mathbf{y} = \mathbf{c}^T\mathbf{X}^T\mathbf{H}\mathbf{X}\mathbf{c} = \mathbf{c}^T\mathbf{T},$$

donde $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ es una matriz simétrica $n \times n$, $\mathbf{I}$ es la matriz unidad $n \times n$, $\mathbf{1}_n \in \mathbb{R}^n$ es el vector que tiene todas sus entradas iguales a 1 y $\mathbf{T} = \mathbf{X}^T\mathbf{H}\mathbf{X}$ es una matriz $p \times p$ denominada *matriz de dispersión total*. La dispersión dentro de grupos $\mathcal{G}$ y $\mathcal{B}$ la podemos describir mediante la expresión

$$\mathbf{Y}_\mathcal{B}^T\mathbf{H}_\mathcal{B}\mathbf{Y}_\mathcal{B} + \mathbf{Y}_\mathcal{G}^T\mathbf{H}_\mathcal{G}\mathbf{Y}_\mathcal{G} = \mathbf{c}^T \left(\mathbf{X}_\mathcal{B}^T\mathbf{H}_\mathcal{B}\mathbf{X}_\mathcal{B} + \mathbf{X}_\mathcal{G}^T\mathbf{H}_\mathcal{G}\mathbf{X}_\mathcal{G}\right) \mathbf{c},$$

La matriz $\mathbf{W} = \mathbf{X}_\mathcal{B}^T\mathbf{H}_\mathcal{B}\mathbf{X}_\mathcal{B} + \mathbf{X}_\mathcal{G}^T\mathbf{H}_\mathcal{G}\mathbf{X}_\mathcal{G}$ es simétrica y de dimensiones $p \times p$.

Por otra parte, la dispersión total entre estos mismos grupos se puede expresar como

$$\begin{aligned} n_\mathcal{G}(\overline{y}_\mathcal{G} - \overline{y})^2 + n_\mathcal{B}(\overline{y}_\mathcal{B} - \overline{y})^2 &= n_\mathcal{G}(\mathbf{c}^T(\mathbf{m}_\mathcal{G} - \mathbf{m}))^2 + n_\mathcal{B}(\mathbf{c}^T(\mathbf{m}_B - \mathbf{m}))^2 \\ &= \mathbf{c}^T\mathbf{B}\mathbf{c}, \end{aligned}$$

donde

$$\mathbf{B} = \frac{n_\mathcal{G} n_\mathcal{B}}{n} \left(\mathbf{m}_\mathcal{G} - \mathbf{m}_B\right) \left(\mathbf{m}_\mathcal{G} - \mathbf{m}_\mathcal{B}\right)^T.$$

Finalmente, se puede demostrar que

$$\mathbf{c}^T\mathbf{T}\mathbf{c} = \mathbf{c}^T\mathbf{W}\mathbf{c} + \mathbf{c}^T\mathbf{B}\mathbf{c}.$$

La idea de Fisher consiste en la elección un vector $\mathbf{c}$ que maximice el cociente

$$\frac{\mathbf{c}^T\mathbf{B}\mathbf{c}}{\mathbf{c}^T\mathbf{W}\mathbf{c}}$$

La solución general de este problema viene dada por el vector propio de la matriz $\mathbf{W}^{-1}\mathbf{B}$ que corresponde al valor propio de mayor magnitud. En nuestro caso, la matriz $\mathbf{W}^{-1}\mathbf{B}$ tiene un único valor propio que es igual a su traza (es decir, a la suma de sus elementos diagonales) y el correspondiente vector propio es $\mathbf{c} = \mathbf{W}^{-1}(\mathbf{m}_{\mathcal{G}} - \mathbf{m}_{\mathcal{B}})$. La regla para discriminar entre grupos se basa en la *distancia de Mahalanobis*. Dada una observación $\mathbf{x} \in \mathbf{R}^p$, diremos que se encuentra en la clase $\mathcal{G}$ si la distancia de Mahalanobis a $\mathbf{m}_{\mathcal{G}}$ es menor que la distancia de Mahalanobis a $\mathbf{m}_{\mathcal{B}}$, esto es, si

$$(\mathbf{x} - \mathbf{m}_{\mathcal{G}})^T \mathbf{W}^{-1} (\mathbf{x} - \mathbf{m}_{\mathcal{G}}) < (\mathbf{x} - \mathbf{m}_{\mathcal{B}})^T \mathbf{W}^{-1} (\mathbf{x} - \mathbf{m}_{\mathcal{B}}).$$

Se puede comprobar que esta expresión es equivalente a

$$\mathbf{c}^T \left(\mathbf{x} - \frac{1}{2} (\mathbf{m}_{\mathcal{G}} + \mathbf{m}_{\mathcal{B}}) \right) > 0.$$

En consecuencia, la función discriminante la podemos expresar del modo siguiente:

$$\mathbf{x} \in \mathcal{G} \quad \text{si} \quad \mathbf{c}^T \left(\mathbf{x} - \tfrac{1}{2} (\mathbf{m}_{\mathcal{G}} + \mathbf{m}_{\mathcal{B}}) \right) > 0,$$

$$\mathbf{x} \in \mathcal{B} \quad \text{si} \quad \mathbf{c}^T \left(\mathbf{x} - \tfrac{1}{2} (\mathbf{m}_{\mathcal{G}} + \mathbf{m}_{\mathcal{B}}) \right) \leq 0.$$

Podemos emplear entonces esta función para determinar si nuevas moléculas poseen o no la característica farmacológica sujeta a estudio (es decir, si son activas o inactivas) en base únicamente a sus índices topológicos.

3.2 Aplicaciones concretas

Como dijimos en la Introducción, la topología molecular nació del estudio de las propiedades físico-químicas de compuestos orgánicos. Como es de esperar, las aplicaciones biológicas y farmacológicas no suelen ser tan determinantes, debido a la gran variabilidad individual de los animales de laboratorio o de las personas sometidas a tratamiento. Además, muchas de las propiedades biológicas son menos específicas que las físicas o químicas, es decir están menos ligadas a la estructura molecular y más relacionadas con la presencia de grupos químicos funcionales, tamaño molecular, etc. Y aunque esto es una gran limitación, algunos autores han obtenido resultados muy interesantes, como por ejemplo Murakami [15] que estableció relaciones cuantitativas entre los índices de conectividad molecular y la unión a proteínas celulares de una serie de pesticidas. Por su parte, Kier y Hall llegaron a una correlación excelente para la acción anestésica local con el índice ${}^1\chi$ y, en otro estudio similar, se probó la correlación entre el logaritmo de la concentración de inhibición del M. Tuberculosis por un conjunto de alquil-bromofenoles y el índice ${}^1\chi^v$ [10] (${}^m\chi^v$ se obtiene formalmente como ${}^m\chi$, sustituyendo en (5) $\deg_i$ por $\deg_i^v = Z_i^v - H_i$, donde Z_i^v y H_i son, respectivamente, el número de electrones de valencia del átomo i y el correspondiente número de átomos de hidrógenos suprimidos).

Veamos un ejemplo concreto de aplicación del método topológico a la analgesia [7]. En este estudio, la función discriminante contenía 11 índices

topológicos (todos ellos índices de Kier y Hall), a saber:

$$^{0}\chi,\ ^{1}\chi,\ ^{1}\chi^{v},\ ^{2}\chi^{v},\ ^{3}\chi_{p},\ ^{3}\chi_{c},\ ^{3}\chi_{p}^{v},\ ^{3}\chi_{c}^{v},\ ^{4}\chi_{p},\ ^{4}\chi_{c}\ ,^{4}\chi_{pc},$$

(véase (6)). Como resultado del análisis discriminante, se seleccionó un conjunto de 17 moléculas con acción analgésica entre las que había derivados de estructuras químicas con reconocida acción analgésica, como el ácido acetilsalicílico (componente activo de la Aspirina) y las pirazolonas, mientras que otras eran inéditas y que, por tanto, se podrían utilizar como cabezas de serie para una nueva gama de analgésicos.

Se realizaron pruebas farmacológicas para confirmar la actividad analgésica, para determinar la dosis eficaz 50 (es decir, la que es óptima para el 50 % de los animales de laboratorio estudiados) y, asimismo, para determinar la dosis letal 50 (es decir, la que provoca la muerte en la mitad de los animales tratados con ese fármaco) para estudios de toxicidad. Estas pruebas se realizaron en ratones de entre 20 y 30 gramos, en número suficiente para obtener datos estadísticos fiables. Las pruebas de analgesia se realizaron siguiendo el protocolo de Witkin [23]. De las 17 moléculas estudiadas, 10 presentaron una clara acción analgésica [7]. El resultado más interesante fue el descubrimiento de que una molécula inédita, concretamente el 2-(1-propenil) fenol, tiene un porcentaje de analgesia que es casi el doble que el del ácido acetilsalicílico. Otra molécula inédita, la 2',4'-dimetilacetofenona, obtuvo resultados parecidos a la anterior. Ambas moléculas fueron patentadas (Patentes n^{o} P.9101034 y P.9101134).

Pero eso no es todo. El índice terapeútico (IT), que es el más empleado para determinar el grado de inocuidad de un medicamento, viene dado por la relación entre la dosis letal 50 y la dosis eficaz 50. Debe ser igual o mayor de 10 para que un medicamento pueda considerarse seguro. Para el 2-(1-propenil) fenol, el índice terapéutico es de 21 y para la 2',4'-dimetilacetofenona, de 16 (compárense estos valores con el IT = 5 del ácido acetilsalicílico). El valor de este índice para la sulfadiazina está en medio de los dos anteriores, a saber, 18. Esto pone de manifiesto el aceptable grado de inocuidad de estos nuevos analgésicos en cuanto a la toxicidad aguda.

Compuesto	**Anelgesia (%)**	**DE50 (mg/kg)**	**DL50 (mg/kg)**	**IT**
Ácido acetilsalicílico	49 ± 1	100 ± 8	500 ± 20	5
2-(1-propenil) fenol	85 ± 1	34 ± 5	720 ± 10	21
2',4'-dimetilacetofenona	80 ± 1	45 ± 5	700 ± 10	16
p-metil-propiofenona	56 ± 1	100 ± 3	590 ± 20	6
Sulfadiazina	43 ± 1	112 ± 10	2000	18

Otro ejemplo significativo es el hallazgo de nuevos antivíricos eficaces contra el Herpes Simplex-tipo I. Aquí la topología molecular hizo posible el descubrimiento del ácido 1,2,3 triazol-4,5 dicarboxílico, que nunca se había utilizado como antivírico [5].

Para terminar, citemos dos ejemplos más: búsqueda de antihistamínicos [6] y búsqueda de antimaláricos [8].

Referencias

[1] A.T. Balaban, Highly discriminating distance-based topological index, *Chem. Phys. Lett.* **89** (1982), 399-404.

[2] A.T. Balaban, I. Motoc, D. Bonchev y O. Mekenyan, Topological indices for structure-activity correlations, *Top. Curr. Chem.* **114** (1984), 21-71.

[3] A.T. Balaban, Numerical modelling of chemical structures: Local graph invariants and topological indices. In, *Graph Theory and Topology in Chemistry* (R.B. King and D.H. Rouvray, Eds.), Elsevier, Amsterdam, *Stud. Phys. Theor. Chem.* **51**, 159-176.

[4] B. Bollobás, *Modern Graph Theory*, Springer Verlag, New York, 1998.

[5] J.V. de Julián-Ortiz, J. Gálvez, C. Muñoz, R. García-Domenech y C. Gimeno, Virtual Combinatorial Syntheses and Computational Screening of New Potential Anti-Herpes Compounds, *J. Med. Chem.* **42** (1999), 3308-3314.

[6] M.J. Duart, R. García-Domenech, J. Gálvez, P. Alemán, R.V. Martín-Algarra y G.M. Antón-Fos, Application of a mathematical topological pattern of antihistaminic activity for the selection of new drug candidates and pharmacology assays, *J. Med. Chem.* **49** (2006), 3667-3673.

[7] J. Gálvez, R. García-Domenech, J.V. de Julián-Ortiz y R. Soler, Topological Approach to Analgesia, *J. Chem. Inf. Comput. Sci.* **34** (1994), 1198-1203.

[8] J. Gálvez, J.V. de Julián-Ortiz y R. García-Domenech, Diseño y desarrollo de nuevos fármacos contra la malaria, *Enf. Emerg.* **7** (2005), 44-51.

[9] O. Ivanciuc, T.S. Balaban y A.T. Balaban, Design of topological indices. Part IV: Reciprocal distance matrix, related local verex invariants and topological indices, *J. Math. Chem.* **12** (1993), 309-318.

[10] L.B. Kier y L.H. Hall, *Molecular Connectivity in Chemistry and Drugs Research*, Academic Press, London, 1976.

[11] L.B. Kier y L.H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, Research Studies Press, Letchworth, 1986.

[12] D.J. Klein y M. Randić, Resistance distance, *J. Math. Chem.* **12** (1993), 81-95.

[13] L.C. Miller y M.L. Tainter, Calculation of ED50 and LD50, *Proc. Soc. Exp. Biol. Med.* **57** (1944), 261-264.

[14] B. Mohar, Laplacian matrices of graphs, *Stud. Phys. Theor. Chem.* **63** (1989), 1-8.

[15] M. Murakami y J. Fukami, J. Specific molecular connectivity analysis of pesticides and related compounds: a preliminary study. *Bull Environ Contam Toxicol.* ***34*** *(*1985), 775-9.

[16] M. Randić, On characterization of molecular branching, *J. Am. Chem. Soc.* **97** (1975), 6609-6615.

[17] M. Randić, On molecular identification numbers, *J. Chem. Inf. Comput. Sci.* **24** (1984), 164-175.

[18] M. Randić, Similarity based on extended basis descriptors, *J. Chem. Inf. Comput. Sci.* **32** (1992), 686-692.

[19] H. Hosoya, A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons, *Bull. Chem. Japan* **44** (1971), 2332-2339.

[20] N. Trinajstić, S. Nikolić, B. Lučić, D. Amić y Z. Mihalić, The detour matrix in chemistry, *J. Chem. Inf. Comput. Sci.* **37** (1997), 631-638.

[21] H. Wiener, Structural determination of paraffin boiling points, *J. Am. Chem. Soc.* **69** (1947), 17-20.

[22] E. Wigner, The unreasonable effectiveness of mathematics in the natural sciences, *Comm. Pure Appl. Math* **13** (1960).

[23] L.B. Witkin, C.F. Heubner, F. Galdi, E. O'Keefe, P. Spitaletta y A.J. Plummer, Pharmacology of 2-amino-indane hydrochloride (Su-8629): a potent non-narcotic analgesic, *J. Pharmacol. Exp. Ther.* **133** *(*1961), 400-408.