

UNIVERSIDAD: Universidad Nacional de La Plata, Universidad de La República

NUCLEO DISCIPLINARIO: Química

Título del trabajo: **BÚSQUEDA COMPUTACIONAL DE NUEVOS FÁRMACOS PARA EL TRATAMIENTO DEL CÁNCER DE COLON**

AUTOR(es): Carolina L. Bellera, Ailín Leticia Buzzi, Alan Talevi, Gloria Serra, Luis E. Bruno-Blanch

CORREOS ELECTRÓNICOS DE LOS AUTORES: cbellera@biol.unlp.edu.ar, ailinleti@hotmail.com, atalevi@biol.unlp.edu.ar, gserra@fq.edu.uy, lbb@biol.unlp.edu.ar

PALABRAS CLAVES: cribado virtual – anticancerígenos – cáncer de colon

PALAVRAS CHAVES: seleção virtual – drogas anticâncer – câncer de cólon

INTRODUCCIÓN

El cáncer colorrectal (CCR) es el tercer cáncer de mayor incidencia a nivel mundial en hombres (después del cáncer de pulmón y el de próstata) y el segundo en mujeres (luego del cáncer de mama) [1,2]. La letalidad del CCR en países europeos ronda el 50% de los pacientes y la misma es mayor en países en vías de desarrollo [2]. Estudios epidemiológicos han demostrado que la alta incidencia del CCR en el mundo occidental se debe a factores ambientales (dieta rica en grasas y proteínas y pobre en fibras y vegetales) y al progresivo envejecimiento de la población [3]. Justamente, se supone que la dieta rica en carnes de Uruguay y Argentina es una de los factores fundamentales que explica que la incidencia del CCR en estos dos países (junto con Bolivia) sea muy superior a la del resto de América del Sur (ver Fig. 1) y similar a la de países desarrollados (Estados Unidos, Canadá, Australia y países Europeos).

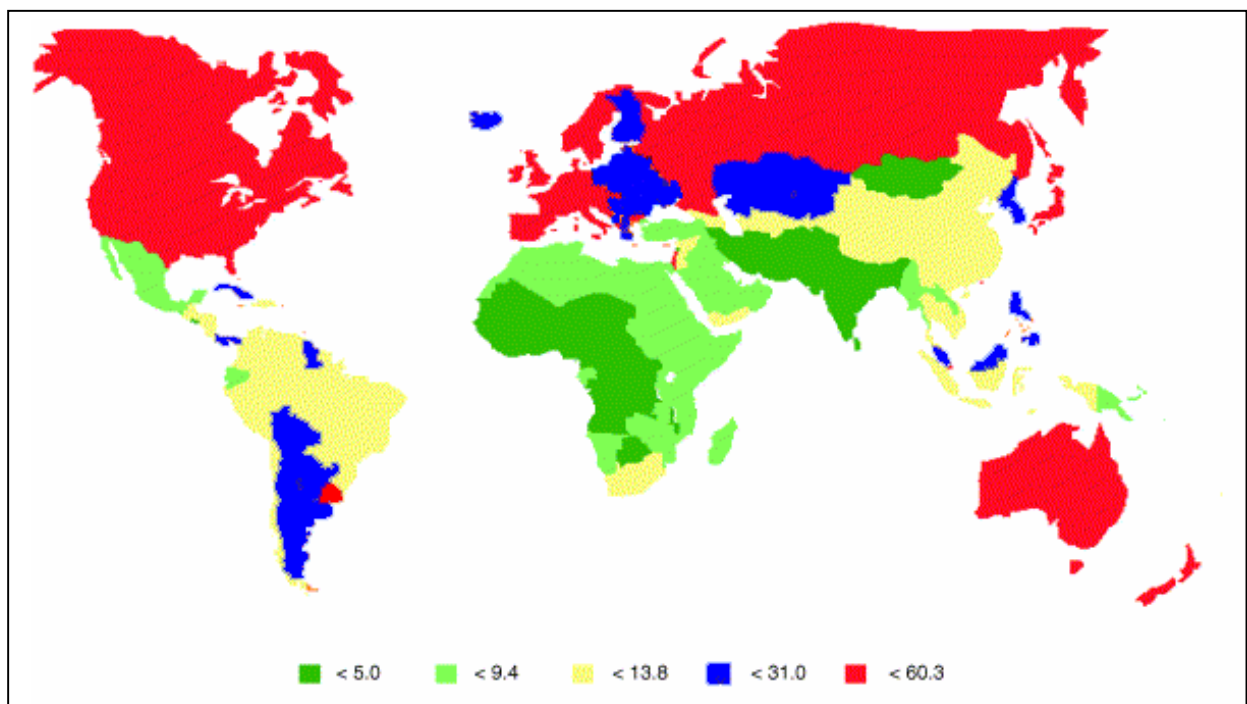


Figura 1. Incidencia del CCR a nivel mundial (hombres, todas las edades). Se observa que la incidencia en Argentina, Uruguay y Bolivia se asimila a la de países desarrollados (extraído de Parkin y col, 2001).

En el marco de la búsqueda racional de nuevos fármacos, el cribado virtual o screening virtual (SV) consiste en aplicar métodos computacionales para seleccionar, a partir de grandes bases de datos de compuestos químicos (típicamente, del orden de cientos de miles a millones de compuestos) aquellos con mayores probabilidades de presentar una actividad farmacológica de

interés. Por varios motivos (su bajo costo relativo, las ventajas bioéticas debido a la reducción/optimización de ensayos in vivo, la posibilidad de encontrar y patentar segundos usos terapéuticos para fármacos ya conocidos reduciendo los costos de ensayos clínicos y los tiempos necesarios para introducir un nuevo medicamento en el mercado) se trata de una metodología de gran potencial para su aplicación en nuestra región, donde la inversión en investigación y desarrollo de nuevos principios activos por parte de la industria farmacéutica es más bien escasa o nula, concentrándose los esfuerzos de dicho sector en el desarrollo galénico, es decir, el desarrollo de nuevas formas farmacéuticas para usos terapéuticos ya conocidos de fármacos ya conocidos [4].

En este trabajo hemos aunado esfuerzos con grupos de investigación de la Universidad de La República (UdeLaR, Uruguay) para combinar nuestra experiencia en la búsqueda racional de nuevos fármacos y nuevas aplicaciones terapéuticas de fármacos ya conocidos, mediante SV [5-11], con la experiencia en síntesis orgánica y evaluación farmacológica de nuevos fármacos para el tratamiento del CCR del grupo de la UdeLaR [12-16], a fin de abordar el desafío del desarrollo de nuevos agentes anticancerígenos desde un enfoque multidisciplinario que integre todas las herramientas utilizadas en la investigación y desarrollo de nuevas entidades terapéuticas. Para ello hemos desarrollado modelos computacionales capaces de identificar agentes promisorios para el tratamiento de CCR a fin de racionalizar la selección de aquellos compuestos que serán sintetizados y evaluados por el grupo de pares uruguayo. Estos modelos han sido aplicados en la selección de 478 compuestos químicos promisorios de entre 525.063 compuestos que conforman la base de datos de compuestos químicos de nuestro laboratorio.

OBJETIVOS

Identificar, mediante SV, fármacos estructuralmente novedosos potencialmente activos y selectivos sobre la línea HCT-15 de cáncer de colon, a fin de orientar la síntesis orgánica y evaluación farmacológica de nuestro grupo de colaboración.

MATERIALES Y MÉTODOS

A fin de identificar compuestos promisorios en la base de datos se generaron 2 funciones discriminantes (FD1 y FD2) mediante el uso del análisis lineal discriminante del módulo "Discriminant analysis" del programa Statistica 7.0 (Statsoft Inc., 2004).

Las funciones discriminantes son modelos matemáticos (o, para ser más precisos, modelos químico-matemáticos) de la forma:

Categoría = f (descriptor molecular 1, descriptor molecular 2, ..., descriptor molecular n)

donde "categoría" es una variable dependiente arbitraria, que indica si un compuesto químico determinado es "activo" o "inactivo"; los descriptores moleculares son variables independientes del modelo que describen o caracterizan algún aspecto determinado de la estructura molecular de los compuestos que forman el conjunto de entrenamiento del modelo químico-matemático. El descriptor "traduce" características de la estructura molecular en números que pueden utilizarse para cuantificar el aporte de algún aspecto de la estructura química a la propiedad que se está modelando (en nuestro caso, actividad anticancerígena en células HCT-15). El número de descriptores moleculares definidos actualmente es muy grande y crece constantemente [17]; los aspectos de la estructura molecular reflejados por un descriptor determinado pueden ser muy diversos, por ejemplo: volumen molecular, distribución de cargas en la molécula, topología molecular, frecuencia de aparición de un grupo funcional o un arreglo de átomos determinado en la molécula, lipofilidad, etc. La función discriminante puede ser una combinación lineal o no lineal de los descriptores incorporados al modelo; en el caso del análisis lineal discriminante la función toma la forma:

$$\text{Categoría} = b_0 + \sum_{i=1}^n b_i \text{descriptor}_i$$

donde b_0 es el término independiente y b_i es el coeficiente de regresión que acompaña al descriptor $_i$. En nuestro caso, se definió una variable dependiente binaria codificante de la categoría, que asume valor 1 para los compuestos activos del conjunto de entrenamiento, y valor -1 para los compuestos inactivos.

Como descriptores moleculares se utilizaron descriptores constitucionales, topológicos y frecuencias de grupos funcionales del software Dragon para cálculo de descriptores moleculares (Milano Chemometrics, 2003).

Para generar y validar la FD1 se utilizaron 128 compuestos químicos: 58 compuestos con IC_{50} (concentración inhibitoria 50) sobre células HCT-15 $< 5 \mu\text{molar}$ y 70 compuestos “similares a fármacos” o *druglike* sin antecedentes de actividad citotóxica sobre la línea HCT-15. El concepto de *druglike* implica que un compuesto químico tiene características estructurales consistentes con las de la mayoría de los fármacos conocidos (por ejemplo, determinado peso molecular, presencia de átomos de O o N en la estructura, valores de coeficiente de reparto octanol-agua en un rango determinado, etc.). La idea de FD1 es entonces identificar características estructurales que diferencien a un compuesto activo sobre HCT-15 de fármacos de otras categorías terapéuticas. La utilización de compuestos inactivos (señuelos) similares a fármaco implica lógicamente mayores exigencias sobre el modelado que si se consideraran compuestos inactivos no similares a fármaco (por ejemplo, compuestos inorgánicos), ya que en este último caso las diferencias entre la categoría activa e inactiva serían muy evidentes y por lo tanto, fáciles de modelar. Debe notarse que la idea del modelo es identificar características no evidentes, mejorar la capacidad de discriminación con respecto a la que tendría un ojo humano entrenado. Los 128 compuestos fueron divididos en un conjunto de entrenamiento de 80 compuestos (40 activos y 40 inactivos) y un conjunto de prueba independiente para la validación de la función, formado por los compuestos restantes (18 activos y 30 inactivos).

La segunda función discriminante, FD2, es a nuestro entender la que reviste mayor interés, ya que fue generada a partir de un conjunto de entrenamiento de 80 compuestos integrado por 40 compuestos con IC_{50} en HCT-15 $< 5 \mu\text{molar}$ y 40 *señuelos miméticos*, compuestos inactivos con características comunes a los compuestos activos (peso molecular similar, presencia de determinados grupos funcionales, pertenencia a las mismas familias químicas, etc.). Desde luego, la mayor similitud de los inactivos constituye un test de stress, un desafío, mucho mayor para el método de modelado. En un excelente artículo de publicación reciente [18], Nicholls ha comparado el proceso de Screening Virtual con una ronda de presos: si se incluyen en la misma, junto con el sospechoso, sujetos de características claramente diferenciales (por ejemplo, una anciana o un niño) la ronda pierde validez como prueba para inculpar al criminal; si por el contrario los sujetos inocentes que participan en la ronda de presos poseen caracteres comunes con el sospechoso, el testimonio del testigo gana validez como prueba ante la corte. En nuestro caso se trata de desafiar al modelo obligándolo a diferenciar compuestos activos e inactivos similares entre sí, para que el procedimiento de SV realmente cobre sentido. Se utilizaron 36 compuestos de características similares a los del conjunto de entrenamiento como conjunto de prueba para evaluar el desempeño esta segunda función.

Los modelos generados fueron validados mediante técnicas estándar de validación: validación interna (validación cruzada Leave Group Out – LOO; ensayo de aleatorización) y validación externa con los conjuntos de prueba independientes del conjunto de entrenamiento reservados para tal fin [19]. La validación LOO consiste en, una vez generado el modelo, remover aleatoriamente un grupo de n compuestos del conjunto de entrenamiento, generando un nuevo modelo con los mismos descriptores presentes en el modelo original con los compuestos remanentes. Este segundo modelo, lógicamente, tendrá coeficientes de regresión distintos a los del original, y se usará para predecir la categoría de los compuestos removidos. El procedimiento se repite por lo menos tantas veces como sea necesario para remover al menos una vez cada compuesto del conjunto de entrenamiento. La aleatorización consiste desordenar los valores de la variable dependiente entre los compuestos del conjunto de entrenamiento mediante un *scrambling* de la misma, generando nuevos modelos aleatorizados con los mismos descriptores del modelo original. Lógicamente, se espera que la capacidad discriminante de estos modelos sea muy baja, asimilándose a una clasificación al azar, ya que en los modelos aleatorizados no existe correlación real entre la variable dependiente y los descriptores. La validación externa consiste en aplicar el modelo original en la clasificación de los casos reservados para tal fin como conjunto de prueba independiente.

Una vez validados adecuadamente, se construyeron a partir de la distribución de los valores que cada función discriminante asumía para el conjunto de entrenamiento los Diagramas de Distribución Farmacológica (DDF) [20], en los que se grafica la expectativa de encontrar un compuesto activo o un compuesto inactivo para cada intervalo de valores de las funciones discriminantes. Las expectativas se calculan mediante los algoritmos:

$$\text{Expectativa activo} = \frac{\% \text{ de activos}}{\% \text{ de inactivos} + 100}$$

$$\text{Expectativa inactivo} = \frac{\% \text{ de inactivos}}{\% \text{ de activos} + 100}$$

Construidos los DDF para ambas funciones, los mismos se utilizaron para definir qué intervalos de valores de cada función discriminante se utilizarían para seleccionar compuestos promisorios en el SV. Se eligieron intervalos en los que hubiera mínima superposición de las categorías activa e inactiva, ya que la superposición de las categorías es una fuente de error que produce clasificaciones erróneas en el proceso de screening.

Se procedió entonces al SV de una base de datos de 525.063 obtenidos de la base de datos de acceso público ZINC database 5.0 [21].

Los compuestos seleccionados fueron rankeados según los valores de FD2 (criterio 1), valores de FD1 (criterio 2) y la distancia euclidiana de cada compuesto respecto al centroide del conjunto de entrenamiento de FD2 y FD1 en el espacio m-dimensional, donde $m = n+1$ y n es el número de descriptores incorporados en cada modelo, al que se le suma un término más para considerar la respuesta como una dimensión adicional [22]. La distancia euclidiana se considera un criterio para definir si los compuestos seleccionados pertenecen o no al dominio de aplicabilidad del modelo, es decir, si se encuentran dentro de la región del espacio químico definida por el conjunto de entrenamiento de cada modelo. En este sentido, debe considerarse al conjunto de entrenamiento como una muestra de la que se obtienen inferencias (el modelo); esa muestra define una población sobre la cual el modelo puede usarse confiablemente para predecir la actividad anticancerígena: el modelo no puede usarse para extrapolar las inferencias a compuestos químicos de características muy distintas a los del conjunto de entrenamiento. Se consideró que las predicciones son más fiables cuando la distancia euclidiana respecto al centroide, para un compuesto seleccionado de la base de datos, es menor a la distancia euclidiana promedio del conjunto de entrenamiento al centroide.

RESULTADOS

Presentamos los descriptores incluidos, parámetros estadísticos, desempeño en la clasificación del conjunto de entrenamiento, resultados del proceso de validación interna (Fig. 2) y desempeño en la clasificación de los respectivos conjuntos de prueba independientes de FD1 y FD2.

FD1 (señuelos similares a fármaco/drug-like)

Descriptores incluidos: nCO, nCaR, C-016, VRA1, GATS4e, JGI2, nC=N , JGI7 (se mantiene la nomenclatura del programa Dragon).

Wilks' $\lambda = 0.199$ $F = 35.79$ p valor < 0.0000

% buenas clasificaciones en conjunto de entrenamiento: 96.25%

% de buenas clasificaciones en el conjunto de prueba independiente: 91.67%

% medio de buenas clasificaciones en el conjunto de entrenamiento, validación LGO (10 rondas, remoción aleatoria de 8 compuestos p/ ronda): 96.40% (sd = 0.72)

% medio de buenas clasificaciones en el conjunto de prueba, validación LGO: 91.67% (sd = 2×10^{-6})

La relación N/d es adecuada (10), y de la similitud entre el desempeño sobre el conjunto de entrenamiento y sobre el conjunto de prueba se infiere que no hay sobre-ajuste apreciable (ajuste de los datos del conjunto de entrenamiento en detrimento de la generalizabilidad del modelo).

FD2 (señuelos miméticos)

Descriptores incluidos: nCO, C-035, HNar, C-016 (se mantiene la nomenclatura del programa Dragon)

Wilks' $\lambda = 0.355$ $F = 34.113$ $p \text{ valor} < 0.0000$

% buenas clasificaciones en conjunto de entrenamiento: 88.75%

% de buenas clasificaciones en el conjunto de prueba independiente: 77.77%

% medio conjunto de entrenamiento, validación LGO (10 rondas, remoción aleatoria de 8 compuestos p/ ronda): 88.47% (sd = 1.47)

% medio conjunto de prueba, validación LGO: 77.21% (sd = 1.17)

La relación N/d es muy buena (20).

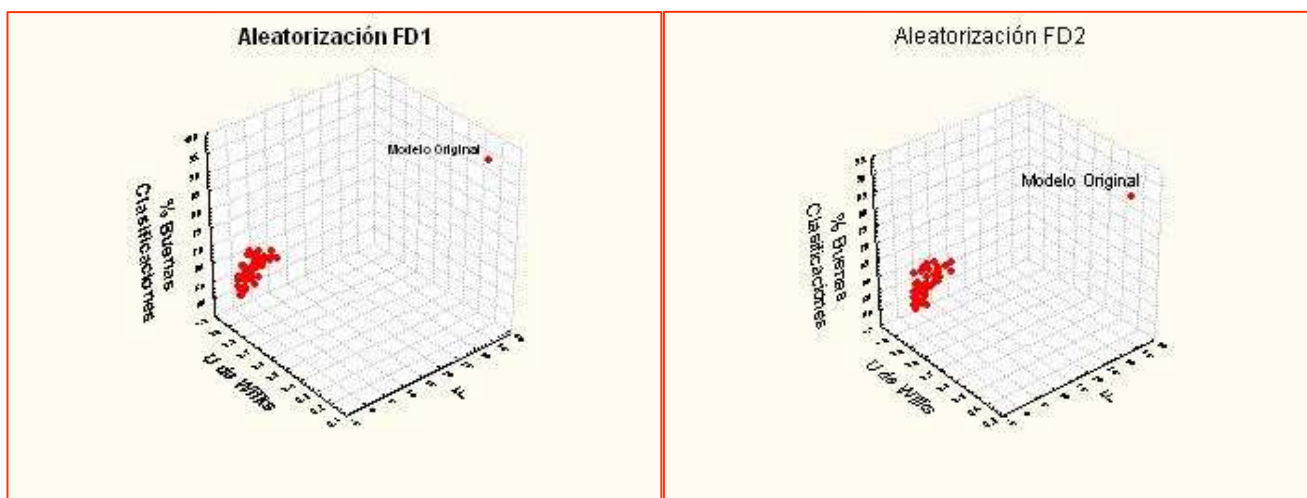


Figura 2. Resultados del ensayo de aleatorización; obsérvese que en ambos casos el modelo original es muy superior en términos estadísticos y en términos de desempeño en la clasificación a los modelos generados por scrambling de la variable dependiente entre los compuestos del conjunto de entrenamiento.

Los DDF obtenidos pueden apreciarse en la Fig. 3:

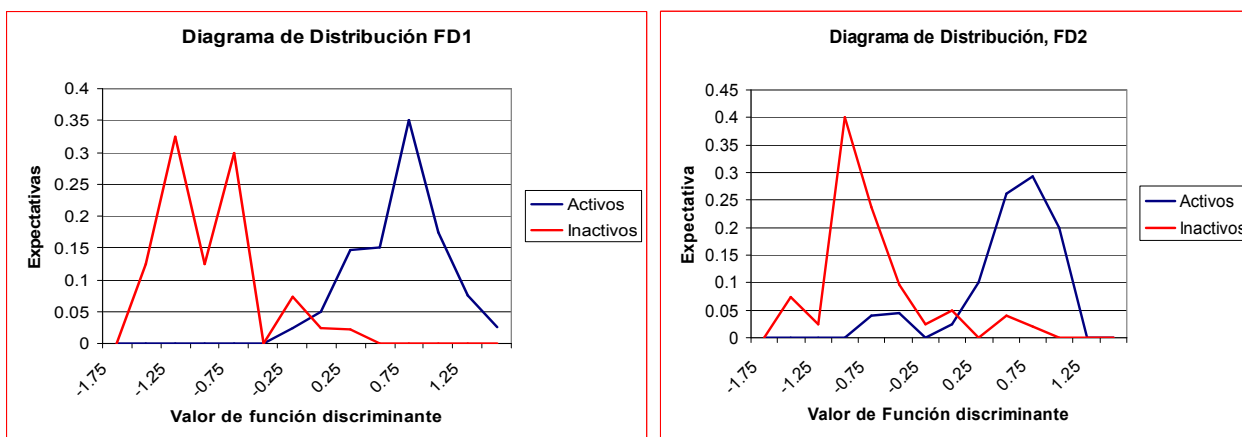


Fig. 3. DDF obtenidos para ambas funciones.

En base a los DDF se eligieron como valores de corte para FD1 valores de la función discriminante mayores a 0.5 y para FD2 valores de la función discriminante entre 1.0 y 1.25 (que, según los diagramas, corresponden a zonas de mínima superposición entre las categorías y minimizan los errores en la clasificación). La aplicación de las funciones discriminantes, con esos intervalos, en el SV de la base de datos dio lugar a la selección de 478 compuestos potencialmente activos. Efectuando una búsqueda bibliográfica de los antecedentes de actividad anticancerígena de los compuestos seleccionados, se observó que varios de ellos (12) poseen actividad sobre diversas líneas celulares de células cancerígenas, y en particular, algunos compuestos (4) han demostrado ser activos en líneas celulares de CCR en ensayos del National Cancer Institute de Estados Unidos (ver Fig. 4).

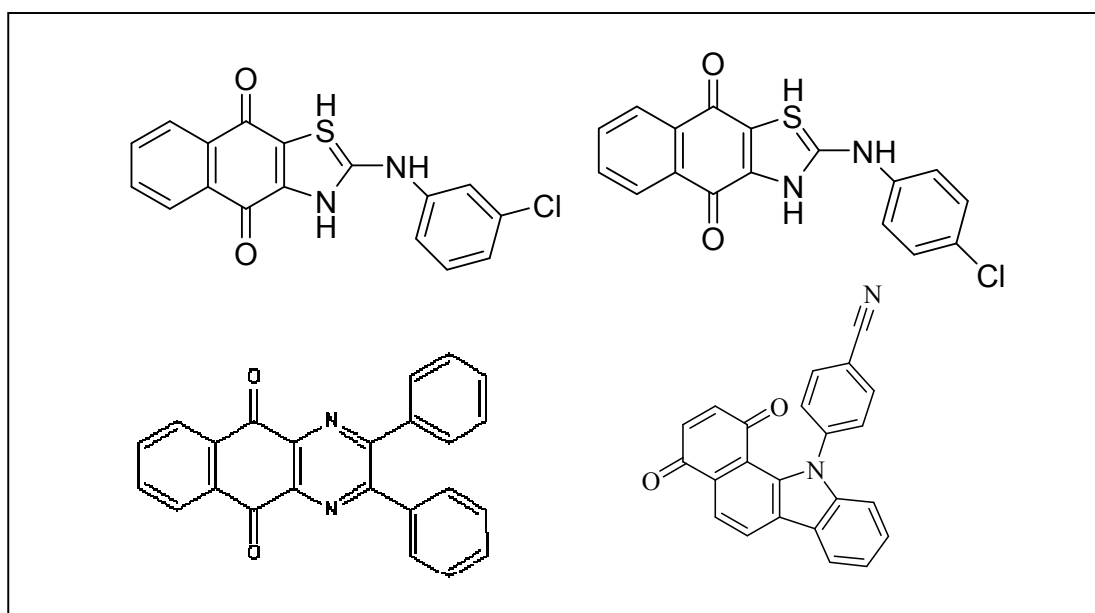


Fig. 4. Algunos de los compuestos seleccionados en el SV con antecedentes de actividad anticancerígena en alguna línea celular de cáncer colorrectal.

DISCUSIÓN

El desempeño de los modelos en los correspondientes conjuntos de entrenamiento y de prueba y en la validación interna sugiere que el modelado de la actividad anticancerígena sobre la línea celular HCT-15 de cáncer de colon ha sido exitoso. Aunque el desempeño de ambas funciones es bueno, se observa un mejor desempeño de FD1, consistente con la idea de que es más sencillo diferenciar compuestos anticancerígenos de señuelos similares a fármaco,

pertenecientes a otras categorías terapéuticas, que diferenciar compuestos activos de señuelos inactivos miméticos.

La búsqueda bibliográfica de antecedentes de actividad anticancerígena de los compuestos seleccionados demostró que varios de ellos tienen antecedentes de actividad anticancerígena, y en particular, 4 compuestos tienen antecedentes de actividad anticancerígena en líneas de CCR, lo que nos alienta a proceder a la segunda etapa del trabajo interdisciplinario: la síntesis y evaluación farmacológica de los compuestos seleccionados que no presentan antecedentes de la actividad buscada.

CONCLUSIÓN

El SV constituye una herramienta eficiente y poco costosa para iniciar la búsqueda racional de nuevos fármacos, siendo su uso de especial interés potencial en América Latina, donde el sector privado no realiza inversiones significativas, actualmente, en la Investigación y Desarrollo de nuevas entidades quimioterapéuticas, actividad que es desarrollada fundamentalmente en nuestra región a través de los centros públicos de investigación, generalmente con recursos limitados. De acuerdo a los resultados del presente trabajo, los datos de la validación de los modelos generados parecen indicar que hemos tenido éxito en el modelado de la actividad de interés. En breve se procederá a la validación definitiva de los modelos generados mediante la síntesis y evaluación in vitro de la actividad citotóxica sobre la línea celular de interés.

AGRADECIMIENTOS

A. Talevi agradece al CONICET por su beca postdoctoral. L. Bruno-Blanch agradece a la Universidad Nacional de La Plata (Incentivos UNLP) y a la ANPCyT (PICT 11985). G. Serra agradece a DINACyT.

BIBLIOGRAFÍA

1. Parkin DM, Bray FI, Devesa SS. Eur J Cancer. 2001, 37, S4-S66
2. Parkin DM, Whelan SL, Ferlay J, Teppo L, Thomas DB. Cancer incidence in five continents. 2002, World Health Organization Press
3. Ershler WB, Longo DL. J Natl Cancer Inst. 1997, 89, 1489-97
4. Talevi A, Bruno-Blanch LE. Lat. Am. J. Pharm. 2009, 28, 141-50.

5. Bruno-Blanch LE, Gálvez J, García Domenech R. *Bioorg. Med. Chem. Lett.* 2003, 13, 2749-54
6. Prieto JJ, Talevi A, Bruno-Blanch LE. *Mol. Divers.* 2006, 10, 361-75
7. Talevi A, Sella-Craverio M, Castro EA, Bruno-Blanch LE. *Bioorg. Med. Chem. Lett.* 2007, 17, 1684-90
8. Talevi A, Bellera CL, Castro EA, Bruno-Blanch LE. *J. Comput. Aid. Mol. Des.* 2007, 21, 527-38
9. Bellera CL, Talevi A, Bruno-Blanch LE. *Lat. Am. J. Pharm.* 2007, 26, 243-51
10. Gavernet L, Talevi A, Castro EA, Bruno-Blanch LE. *QSAR Comb. Sci.* 2008, 27, 1120-9
11. Talevi A, Gavernet L, Bruno-Blanch LE. *Current Comput-Aided Drug. Des.* 2009, 5, 23-37
12. Mahler G, Serra G, Manta E. *Synth. Comm.* 2005, 35, 1481-92
13. Sellanes D, Scarone L, Manta E, Wipf P, Serra G. *Lett. Org. Chem.* 2006, 3, 309-12
14. Sellanes D, Scarone L, Mahler G, Manta E, Baz A, Dematteis S, Saldaña J, Domínguez L, Wipf P, Serra G. *Lett. Drugs Des. Disc.* 2006, 3, 625-32
15. Mahler G, Serra G, Dematteis S, Saldaña J, Domínguez L, Manta E. *Bioorg. Med. Chem. Lett.* 2006, 3, 309-12
16. Sellanes D, Manta E, Serra G. *Tetrahedron Lett.* 2007, 48, 1827-30
17. Dudek AZ, Arodz T, Gálvez J. *Comput. Chem. High Throughput Screen* (2006), 9, 213-28
18. Nicholls A. *J. Comput. Aided Mol Des.* 2008, 22, 239-55
19. Yasri A, Hartsough D. *J. Chem. Inf. Comput. Sci.* (2001), 41, 1218-27
20. Gálvez J, García-Domenech R, Alapont de Gregorio C, de Julián Ortiz JV, Popa L. *J. Mol. Graph.* 1996, 14, 272-6
21. Irwin JJ, Shoichet BK. *J. Chem. Inf. Model.* 2005, 45, 177-82. <http://zinc.docking.org/>
22. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. *ATLA* (2005), 33, 445-59